
Functional analysis of position effects of inversion *2j*
in *Drosophila buzzatii*: gene *CG13617* silencing
and its adaptive significance

Análisis funcional de efectos de posición de la inversión *2j*
en *Drosophila buzzatii*: el silenciamiento del
gen *CG13617* y su significado adaptativo

Anàlisi funcional d'efectes de posició de la inversió *2j*
a *Drosophila buzzatii*: el silenciament del
gen *CG13617* i el seu significat adaptatiu

DOCTORAL THESIS

Marta Puig Font



Universitat Autònoma de Barcelona
Facultat de Biociències
Departament de Genètica i Microbiologia
Bellaterra, 2010

Memòria presentada per la Llicenciada en Biologia
Marta Puig Font per a optar al grau de Doctora.

Marta Puig Font

Bellaterra, 10 de novembre de 2010

El Doctor Alfredo Ruiz Panadero, Catedràtic del Departament de Genètica i Microbiologia de la Facultat de Biociències de la Universitat Autònoma de Barcelona, i el Doctor Mario Cáceres Aguilar, Professor d'Investigació ICREA de l'Institut de Biotecnologia i Biomedicina de la Universitat Autònoma de Barcelona,

CERTIFIQUEN que la Marta Puig Font ha dut a terme sota la seva direcció el treball de recerca realitzat al Departament de Genètica i Microbiologia de la Facultat de Biociències de la Universitat Autònoma de Barcelona que ha portat a l'elaboració d'aquesta Tesi Doctoral titulada “Functional analysis of position effects of inversion *2j* in *Drosophila buzzatii*: gene *CG13617* silencing and its adaptive significance”.

I perquè consti als efectes oportuns, signen el present certificat a Bellaterra, a 10 de novembre de 2010.

Dr. Alfredo Ruiz Panadero

Dr. Mario Cáceres Aguilar

Table of contents

ABSTRACT RESUMEN RESUM	13
1 INTRODUCTION	19
1.1 Chromosomal inversions	21
1.1.1 The generation of inversions	22
1.1.2 Adaptive value of inversions	23
1.1.3 Spread of inversions in populations	26
1.1.4 Traits and genes	29
1.2 Position effects of inversion breakpoints	31
1.3 Effects of transposable elements on gene expression	36
1.3.1 Regulatory changes	44
1.3.2 Epigenetic effects	48
1.4 The inversion <i>2j</i> of <i>D. buzzatii</i>	51
1.4.1 The origin of inversion <i>2j</i>	53
1.4.2 Genes flanking the inversion <i>2j</i> breakpoints	56
1.5 Objectives	61
2 MATERIALS AND METHODS	65
2.1 <i>Drosophila</i> lines	67
2.2 Nucleic acid isolation	68
2.3 RT-PCR and PCR	69
2.4 RACE	72
2.5 Real-time RT-PCR	72
2.6 DNA sequencing	73

2.7 Sequence analysis _____	75
2.7.1 Sequence annotation and comparative DNA sequence analysis	75
2.7.2 Protein analysis	76
2.8 Northern blot analysis _____	78
2.9 Whole-mount <i>in situ</i> hybridization in <i>Drosophila</i> embryos _____	78
2.10 dsRNA detection _____	79
2.11 RNA interference _____	79
2.11.1 Synthesis of a dsRNA molecule complementary to <i>D. melanogaster</i> gene <i>CG13617</i>	79
2.11.2 Microinjection	81
2.12 Microarrays _____	85
3 RESULTS _____	91
3.1 Position effect of inversion <i>2j</i> on <i>CG13617</i> gene expression in <i>D. buzzatii</i> _____	93
PUIG, M., CÁCERES, M. and RUIZ, A. (2004) Silencing of a gene adjacent to the breakpoint of a widespread <i>Drosophila</i> inversion by a transposon-induced antisense RNA. <i>Proc. Natl. Acad. Sci. USA</i> 101 : 9013-9018.	94
3.2 Functional consequences of <i>CG13617</i> silencing _____	106
3.2.1 Silencing of <i>D. melanogaster CG13617</i> gene expression by RNAi	107
3.2.2 Detection of gene-expression changes induced by <i>CG13617</i> silencing using microarrays	110
3.2.3 Molecular consequences of <i>CG13617</i> silencing in <i>D. buzzatii 2j</i> lines	121
3.3 Evolution and function of gene <i>CG13617</i>: comparative sequence analysis _____	127
3.3.1 <i>CG13617</i> genomic structure in <i>Drosophila</i> species	127
3.3.2 Sequence analysis of the <i>CG13617</i> protein in <i>Drosophila</i> species and other organisms	130
3.3.3 Identification of regulatory sequences in gene <i>CG13617</i>	141

4 | DISCUSSION --- 147

4.1 Position effect of inversion *2j* on *CG13617* gene expression --- 149

4.1.1 Possible causes of *CG13617* silencing 151

4.1.2 *CG13617* antisense transcript 161

4.1.3 Transposable elements can induce transcription of adjacent sequences 168

4.2 Consequences of *CG13617* silencing --- 173

4.2.1 Expression changes associated to *CG13617* silencing 173

4.2.1.1 Magnitude of the expression changes induced by *CG13617*
silencing 177

4.2.1.2 Specificity of *CG13617* silencing effect 181

4.2.2 Gene *CG13617* structure and function 183

4.2.2.1 Changes in gene structure within the *Drosophila* genus 184

4.2.2.2 Protein sequence analysis 185

4.2.2.3 *CG13617* homologous proteins in other species 187

4.2.2.4 Is *CG13617* a component of the Hedgehog signaling pathway? 189

4.2.3 *CG13617* and inversion *2j* evolutionary history 197

5 | CONCLUSIONS --- 205

APPENDIX I --- 209


CÁCERES, M., PUIG, M. and RUIZ, A. (2001) Molecular characterization of two natural hotspots in the *Drosophila buzzatii* genome induced by transposon insertions. *Genome Res* **11**: 1353-1364.

APPENDIX II --- 223

CLARK, A.G., EISEN, M.B., SMITH, D.R., BERGMAN, C.M., OLIVER, B., MARKOW, T.A., KAUFMAN, T.C., KELLIS, M., GELBART, W., IYER, V.N., *et al.* (2007) Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* **450**: 203-218.

APPENDIX III --- 241

MARZO, M., PUIG, M. and RUIZ, A. (2008) The *Foldback*-like element *Galileo* belongs to the P superfamily of DNA transposons and is widespread within the *Drosophila* genus. *Proc Natl Acad Sci U S A* **105**: 2957-2962.

APPENDIX IV _____	249
DELPRAT, A., NEGRE, B., PUIG, M. and RUIZ, A. (2009) The transposon <i>Galileo</i> generates natural chromosomal inversions in <i>Drosophila</i> by ectopic recombination. <i>PLoS One</i> 4 : e7883.	
BIBLIOGRAPHY _____	265
WEB REFERENCES  _____	281
ABBREVIATIONS _____	285
INDEX OF FIGURES _____	290
INDEX OF TABLES _____	293
INDEX OF BOXES _____	295
ACKNOWLEDGEMENTS AGRADECIMIENTOS AGRAÏMENTS _____	296

A la Sara i el Joan

Chromosomal inversions have been known for a long time to be maintained by natural selection in *Drosophila* populations. However, the molecular mechanisms underlying their adaptive value remain uncertain. In *D. buzzatii* natural populations, inversion *2j* forms a balanced polymorphism with the *2st* arrangement, in which *2j* individuals have a larger size and a longer developmental time compared to *2st* carriers. In this work we tested the hypothesis that a position effect of one of the inversion breakpoints could be the cause of these phenotypic changes by analyzing the expression of a gene adjacent to the proximal breakpoint, *CG13617*, in *D. buzzatii* lines with and without inversion *2j*. We have found that in *2j* embryos an antisense RNA originated in a copy of a *Galileo* family transposon inserted at the breakpoint causes a 5-fold decrease of the expression level of *CG13617*. In order to investigate the functional consequences of the reduction in *CG13617* expression, we have used RNA interference to reproduce this silencing in *D. melanogaster*. Microarray and real-time RT-PCR experiments comparing first instar larvae with and without *CG13617* expression revealed that 41 genes show reduced expression levels when *CG13617* is silenced, while none is up-regulated. Interestingly, genes involved in DNA replication and cell cycle are significantly enriched among those affected by *CG13617* silencing. Nine out of ten of these genes analyzed in *D. buzzatii* also show a reduced expression level in *2j* embryos, but not in first instar larvae, a stage where the *CG13617* expression difference between chromosomal arrangements is lower and the antisense RNA is no longer transcribed. To gain insight into the potential function of this gene we have carried out a comprehensive nucleotide and protein sequence analysis in the 12 available *Drosophila* genomes and also in other organisms. *CG13617* protein contains a conserved C2H2 zinc finger, three coiled coil regions, two PEST sequences, and putative nuclear localization and export signals, and shows similarity to human DZIP1 and zebrafish Iguana (a component of the Hedgehog signaling pathway) proteins, which indicates that its cellular role could be related to the transport of transcription factors in and out of the nucleus. These results suggest that gene *CG13617* could be involved in the regulation of DNA replication and that the position effect in *2j* carriers might contribute to explain the

phenotypic differences observed between $2st$ and $2j$ individuals as well as the adaptive value of the inversion.

Es sabido desde hace mucho tiempo que las inversiones cromosómicas son mantenidas por selección natural en muchas poblaciones de *Drosophila*. Sin embargo, los mecanismos que causan este valor adaptativo aún no se conocen. En las poblaciones naturales de *D. buzzatii*, la inversión *2j* forma un polimorfismo equilibrado con la ordenación *2st*, en que los individuos *2j* tienen un mayor tamaño y un tiempo de desarrollo más largo en comparación con los *2st*. En este trabajo hemos puesto a prueba la hipótesis de que un efecto de posición de uno de los puntos de rotura de la inversión podría ser la causa de estos cambios fenotípicos. Para ello, hemos analizado la expresión de un gen adyacente al punto de rotura proximal, *CG13617*, en líneas de *D. buzzatii* con y sin la inversión *2j*. Hemos encontrado que los embriones *2j* presentan un nivel de expresión de *CG13617* cinco veces menor causado por un RNA *antisense* originado en una copia de un transposón de la familia *Galileo* insertado en el punto de rotura. Las consecuencias funcionales de esta reducción en la expresión de *CG13617* se han investigado utilizando la técnica de RNA interferencia para reproducir este silenciamiento en *D. melanogaster*. Los experimentos con *microarrays* y RT-PCR en tiempo real comparando larvas de primer estadio con y sin expresión de *CG13617* han revelado que 41 genes muestran niveles de expresión reducidos cuando *CG13617* es silenciado, mientras que ningún gen presenta un incremento. Además, hay un exceso significativo de genes implicados en la replicación del DNA y en el ciclo celular entre los afectados por el silenciamiento de *CG13617*. Nueve de diez de estos genes fueron analizados en *D. buzzatii* y también tienen un nivel de expresión reducido en embriones *2j*, aunque no en larvas de primer estadio, una fase en la que la diferencia en la expresión de *CG13617* entre ordenaciones es menor y el RNA *antisense* ya no se transcribe. Para averiguar la posible función de este gen hemos llevado a cabo un exhaustivo análisis de secuencias nucleotídicas y proteicas en los 12 genomas de *Drosophila* disponibles y también en otros organismos. La proteína *CG13617* contiene un dedo de zinc tipo C2H2 muy conservado, tres regiones capaces de formar *coiled coils*, dos secuencias PEST y posibles señales de localización y exportación nuclear. También presenta similitud con la proteína humana DZIP1 y con Iguana, un componente de la vía de señalización Hedgehog en

el pez cebra, hecho que indica que su papel en la célula podría estar relacionado con el transporte de factores de transcripción hacia dentro y hacia fuera del núcleo. Estos resultados sugieren que el gen *CG13617* podría estar implicado en la regulación de la replicación del DNA y que el efecto de posición en los portadores de la inversión *2j* podría contribuir a explicar las diferencias observadas entre los individuos *2st* y *2j*, así como el valor adaptativo de esta inversión.

És sabut de fa molt temps que les inversions cromosòmiques son mantenides per selecció natural a les poblacions de *Drosophila*. No obstant, els mecanismes moleculars que generen aquest valor adaptatiu encara no es coneixen. A les poblacions naturals de *D. buzzatii*, la inversió *2j* forma un polimorfisme equilibrat amb l'ordenació *2st*, en què els individus *2j* tenen una mida més gran y un temps de desenvolupament més llarg en comparació amb els *2st*. En aquest treball hem posat a provat la hipòtesi de que un efecte de posició d'un dels punts de trencament podria ser la causa d'aquests canvis fenotípics. Per a fer-ho hem analitzat l'expressió d'un gen adjacent al punt de trencament proximal, *CG13617*, en línies de *D. buzzatii* amb i sense la inversió *2j*. Hem trobat que els embrions *2j* presenten un nivell d'expressió de *CG13617* cinc vegades menor causat per un RNA *antisense* originat en una còpia d'un transposó de la família *Galileo* inserit al punt de trencament. Les conseqüències funcionals de la reducció de l'expressió de *CG13617* s'han investigat utilitzant la tècnica de RNA interferència per reproduir aquest silenciament a *D. melanogaster*. Els experiments de *microarrays* i RT-PCR en temps real comparant larves de primer estadi amb i sense expressió de *CG13617* han revelat que 41 gens mostren nivells d'expressió reduïts quan *CG13617* és silenciament, mentre que cap gen presenta un increment. A més, hi ha un excés significatiu de gens implicats en la replicació del DNA i el cicle celular entre els afectats pel silenciament de *CG13617*. Nou de deu d'aquests gens van ser analitzats a *D. buzzatii* i també tenen un nivell d'expressió reduït en embrions *2j*, però no en larves de primer estadi, una fase en què la diferència d'expressió de *CG13617* entre ordenacions cromosòmiques és menor i el RNA *antisense* ja no es transcriu. Per a esbrinar la possible funció d'aquest gen hem dut a terme un exhaustiu anàlisi de seqüències nucleotídiques i proteiques en els 12 genomes de *Drosophila* disponibles i també en altres organismes. La proteïna *CG13617* conté un dit de zinc tipus C2H2 molt conservat, tres regions que poden formar *coiled coils*, dues seqüències PEST i senyals de localització i exportació nuclear. També presenta similitud amb la proteïna humana DZIP1 i amb Iguana, un component de la via de senyalització Hedgehog al peix zebra, fet que indica que el seu paper dins la cèl·lula podria estar relacionat amb el transport de factors de transcripció cap a

dins i cap a fora del nucli. Aquests resultats suggereixen que el gen *CG13617* podria estar implicat en la regulació de la replicació del DNA i que l'efecte de posició en els portadors de la inversió *2j* podria contribuir a explicar les diferències fenotípiques observades entre els individus *2st* i *2j*, així com el valor adaptatiu d'aquesta inversió.

INTRODUCTION

Tots els ulls miren, pocs observen, molt pocs hi veuen.

– ALBERT SÁNCHEZ PIÑOL, *La pell freda* (2002)

1.1 Chromosomal inversions

One of the main challenges faced by biology today is trying to decipher genome function and understand how genome sequences translate into individuals and organisms with different phenotypic characteristics. Solving this problem requires the identification and study of genomic variants within and between species to investigate their effect on phenotype. The sequencing of the genomes of many species has led to the discovery of a great degree of genome variation (ranging from single nucleotide polymorphisms to chromosomal rearrangements) that could be associated to complex traits. Probably most of these variants are neutral and do not have any effect on the individuals that carry them, but the next step is trying to identify those variants that do have functional consequences and to determine their role in evolution and phenotypic variation.

After a few years during which interindividual variation has been largely attributed to single nucleotide polymorphisms (SNPs), recent genomic techniques have uncovered an unexpectedly large extent of structural variation in many genomes and the interest for the mechanisms underlying the origin, evolution and especially the contribution to phenotypical diversity of this type of variants has increased substantially (2007, STRANGER *et al.* 2007). Deletions, duplications and insertions (now more commonly known as copy-number variants), as well as inversions and translocations, which can comprise millions of nucleotides of heterogeneity within every genome and can contribute to gene expression variation (STRANGER *et al.* 2007), have been detected in the human genome (KORBEL *et al.* 2007, LEVY *et al.* 2007, KIDD *et al.* 2008) but also in other species like mice (QUINLAN *et al.* 2010) or *Drosophila* (AULARD *et al.* 2004, DOPMAN and HARTL 2007).

In this work, we focused on one specific type of structural variation that has been known for a long time to be able to affect phenotype and be under selection in certain species: chromosomal inversions. Inversions are produced when a segment of a chromosome is excised and reinserted in the opposite orientation, which results in the inversion of the intervening sequences and the consequent alteration of gene order with respect to the original chromosomal arrangement. Inversions were the first type of structural variant to be studied and, since their discovery in *Drosophila* polytene chromosomes more than 80 years ago, an extraordinarily rich inversion polymorphism has been found in multiple *Drosophila* species. These studies have resulted in a very fertile line of experimental and theoretical research on different aspects of inversion biology, and for a long time inversions have been a paradigm of population genetics and evolutionary biology (DOBZHANSKY 1970, KRIMBAS and POWELL 1992).

In particular, in *Drosophila*, inversions were first discovered due to their effects on recombination during the construction of genetic maps (STURTEVANT 1917). Afterwards, the presence of polytene chromosomes in salivary glands made possible to identify the inverted chromosomal segments (PAINTER 1933) and check their distribution in different species and populations (STONE *et al.* 1960, SPERLICH and PFRIEM 1986). Since then, thousands of inversions have been identified in *Drosophila*, either as fixed differences between species or as polymorphisms within species (KRIMBAS and POWELL 1992) becoming the most frequent polymorphic and fixed chromosomal change in the *Drosophila* genus.

1.1.1 The generation of inversions

A first step towards investigating the potential biological consequences of inversions is to characterize their breakpoints at the molecular level. The analysis of inversion breakpoints has revealed several mechanisms able to cause their formation. Ectopic recombination between oppositely oriented copies of transposable elements (TEs) is the causing mechanism for three polymorphic inversions in *D. buzzatii* (CÁCERES *et al.* 1999, CASALS *et al.* 2003, DELPRAT *et al.* 2009). In other cases, even though some TEs were found at the breakpoints, it is not clear if they were involved in the generation of the inversion (MATHIOPOULOS *et al.*

1998, ANDOLFATTO *et al.* 1999). Inversions that appear to have been formed by simple cut-and-paste mechanisms (NHEJ) have also been detected, like inversion *In(3L)Payne* in *D. melanogaster* (WESLEY and EANES 1994). Also, a recent analysis of the breakpoints of the 29 inversions differentiating *D. melanogaster* and *D. yakuba* genomes has revealed duplicated sequences at the breakpoints that were not found anywhere else in the genome and that most likely result from staggered breaks (RANZ *et al.* 2007). In humans, most inversions occur by non-allelic recombination between homologous sequences (NAHR) situated in opposite orientation. Typically these sequences are large segmental duplications or smaller common repetitive sequences, like SINEs or LINEs (KEHRER-SAWATZKI and COOPER 2008). The relative importance of each mechanism has yet to be established but the prevalent mechanisms could be different for each lineage, just like the occurrence of inversions seems to be, at least within the *Drosophila* genus (CLARK *et al.* 2007) (see APPENDIX II).

1.1.2 Adaptive value of inversions

It has been known for a long time that polymorphic inversions (i.e. those segregating in natural populations) have an adaptive value in some species. This means that inversions are somehow able to affect the phenotype and be detected by natural selection, which can cause their spread and maintenance in populations. Polymorphic inversions are very common in a great number of *Drosophila* species, and many of them were soon found to be under selection, since latitudinal and altitudinal clines, as well as seasonal variations in inversion frequencies have been reported for several species (KRIMBAS and POWELL 1992). This is the case of *D. subobscura*, which in its colonization of America during the 1970s in less than five years reproduced both in North and South America the same latitudinal clines for the different arrangement frequencies that had been previously observed in European populations (PREVOSTI *et al.* 1988). In addition, inversion polymorphisms in *D. robusta* exhibit similar altitudinal clines in different mountains (LEVITAN 2001), and seasonal variations have been described in *D. pseudoobscura* (KRIMBAS and POWELL 1992) and *D. subobscura* (RODRÍGUEZ-TRELLES *et al.* 1996), even though patterns of seasonal changes in inversion frequencies can be complex. The reason underlying these variable geographical or temporal distributions of the distinct chromosomal arrangements within a species could be that each arrangement is better

adapted to distinct environmental conditions (those found at each end of a cline or in different seasons) and that natural selection favors a higher frequency of one or the other arrangement depending on the local or temporal conditions where each population is found. This climatic selection has prompted studies linking changes in the frequencies of *Drosophila* polymorphic inversion to recent climate change. For example, in *D. subobscura* the comparison of data on over 21 inversion polymorphisms gathered over 24 years has revealed that the

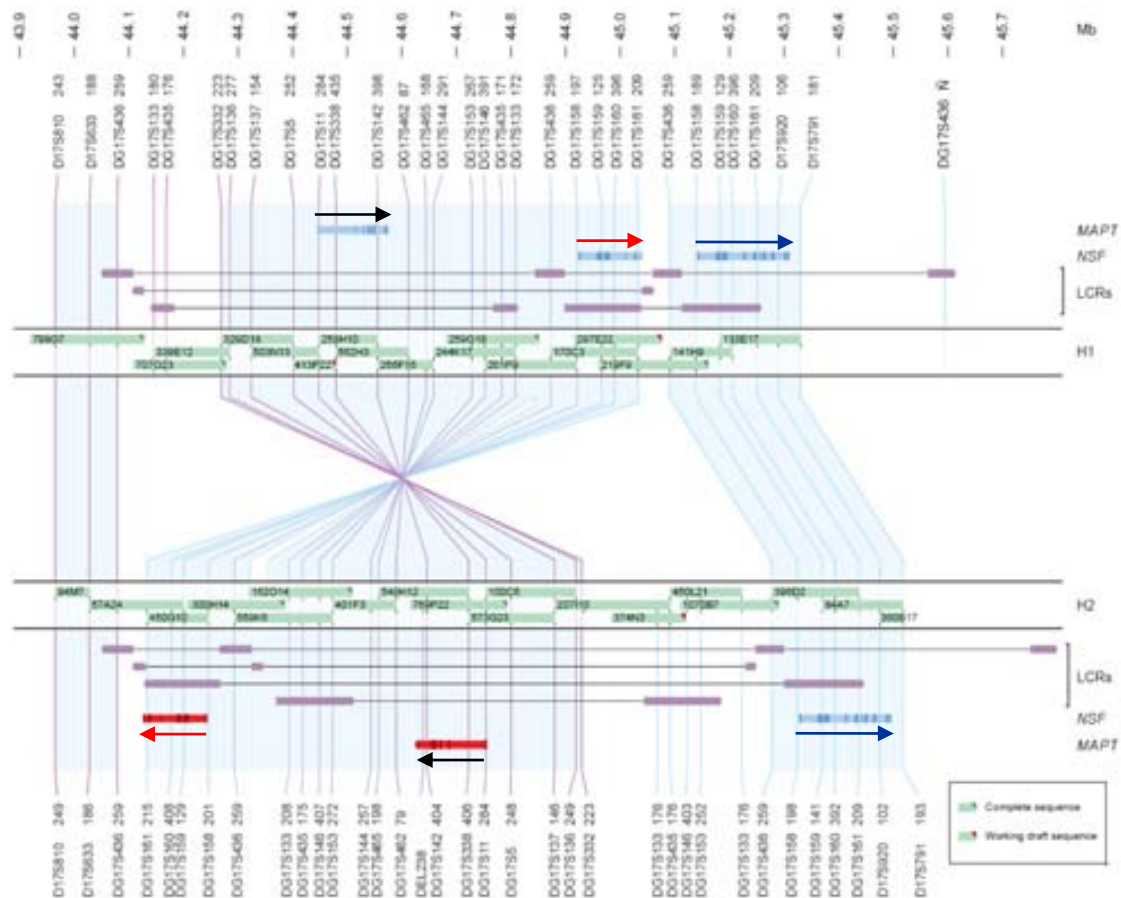


FIGURE 1 | 900-kb inversion polymorphism detected in human chromosome 17. Inversion breakpoints reside in 100-kb palindromic low-copy repeats (LCRs) shown in purple. Microsatellite markers are shown above and below the black lines representing the chromosomes. Vertical lines connect the equivalent marker in both arrangements. BAC clones are depicted as green bars. Features illustrated in each line are indicated at the left of the image. The upper chromosome corresponds to the major H1 haplotype, and the lower chromosome represents the inverted H2 variant. Full-length gene *NSF* (blue arrow) and a truncated copy of *NSF* including exons 1-13 (red arrow) are present in both chromosomes, but the distance between them and their respective orientation are different in both arrangements because the truncated copy is located inside the inverted segment and therefore, it has been moved to a more distant position. Gene *MAPT* (black arrow) is also located within the inverted fragment and is found in the opposite orientation in the two chromosomes. Figure modified from STEFANSSON *et al.* (2005).

frequencies of the different chromosomal arrangements have shifted toward a more low-latitude pattern in 21 out of 22 locations where temperature has increased during this period (BALANYÀ *et al.* 2006), indicating that inversion frequency is able to change in response to varying environmental conditions. Also, in *D. melanogaster*, the clinal pattern observed for inversion *In(3R)Payne* in the eastern coast of Australia has experienced a latitudinal shift equivalent to 7.3° in latitude in the last 20 years as a consequence of the increasingly warmer and drier conditions (ANDERSON *et al.* 2005, UMINA *et al.* 2005).

Inversions with adaptive effects have been described also in species other than *Drosophila*. In humans, a 900-kb polymorphic inversion in chromosome 17 (FIGURE 1) generated by NAHR between complex blocks of segmental duplications of 200-500 kb, is associated to a region of extended linkage disequilibrium that presents two main haplotypes that diverged as much as 3 million years ago. This inversion is found at a 20% frequency in European populations, while it is absent from African populations, a distribution consistent with the action of positive selection. In fact, in the Icelandic population, carrier females of the inverted arrangement present an increased fertility, with around 3% more children, and have higher recombination rates than non-carriers (STEFANSSON *et al.* 2005). Also, in white-throated sparrows (*Zonotrichia albicollis*), the presence of a white stripe together with a more aggressive and male-like behavior are associated to ZAL2^m arrangement in chromosome 2, which was created by two nested pericentric inversions that comprise 86% of the chromosome. This arrangement is maintained in the sparrow populations by sexual selection via a strong disassortative mating preference such that 96% of observed breeding pairs are composed of birds from both morphs (THOMAS *et al.* 2008). In plants, a widespread polymorphic inversion in yellow monkeyflower (*Mimulus guttatus*) causes the differences between the two ecotypes of this plant in North America: an annual form adapted to the dry habitats found inland, and a perennial form adapted to the cool and moist areas of the coast. By affecting flowering time so that the two forms are not available at the same time for pollination, this inversion also contributes to reproductive isolation (LOWRY and WILLIS 2010). Therefore, several studies on polymorphic inversions have provided ample evidence that inversions can have a positive effect on fitness and that they are adaptive. However, it remains unclear how these adaptive values are produced.

1.1.3. Spread of inversions in populations

For a long time the existence of so many inversions in some genus like *Drosophila* has generated an intense debate over the forces governing their fate (KIRKPATRICK 2010). After an inversion arises in a population, it can have different fates: it may be lost, spread to fixation or remain in a polymorphic state. In case that the inversion has deleterious effects, it will be eliminated by purifying selection. On the contrary, if the inversion turns out to be beneficial for the individuals carrying it, its frequency will increase thanks to positive selection and it will become fixed in the population, contributing to divergence between species. It is also possible that each arrangement presents an advantage and that both are actively maintained by natural selection. In this case, inverted and non-inverted chromosomes will remain in intermediate frequencies forming a balanced polymorphism. Therefore, natural selection could not only be involved in the increase or decrease of the frequency of inversions, but it can also be the force behind the stable inversion polymorphisms found in some natural populations.

There are four main explanations for the spread of inversions in natural populations (HOFFMANN and RIESEBERG 2008). The simplest case is the hypothesis that inversions are neutral, and that their probability of fixation or loss depends exclusively on genetic drift (and therefore on population size) and migration. Many small inversions segregating in populations could be neutral but, as mentioned above, in *Drosophila*, frequencies of large inversions usually present latitudinal or altitudinal clines or seasonal variations (KRIMBAS and POWELL 1992), which strongly suggest that inversions are influenced by selection for adaptation to different environmental conditions, and probably are maintained in populations through different mechanisms.

The next two hypothesis focus on the impact that inversions have on maintaining linkage disequilibrium between *loci* located within the inverted segment. Inversions maintain linkage disequilibrium because they suppress recombination in the inverted region in heterokaryotypes, both through the reduced pairing and crossing over between inverted regions, as well as selection against the unbalanced gametes that result from unique crossovers taking place within the inverted segment and that can not originate viable offspring. When an inversion is formed, all the *loci* contained in the inverted segment are in strong linkage

disequilibrium with the inversion. With time, due to the fact that viable gametes can be generated by double crossovers within the inverted segment, to gene conversion events, or to the creation of new mutations, we will expect to find linkage disequilibrium only in regions around the breakpoints (that can not pair in heterokaryotypes) and in those regions where selection maintains associations between alleles (HOFFMANN and RIESEBERG 2008). So, the reduction of recombination promotes the divergence of inverted and non-inverted chromosomes, even though genetic exchange is still possible between the different arrangements.

The coadaptation hypothesis (DOBZHANSKY 1970) suggests that inversions protect favorable (coadapted) combinations of alleles that work well together. In this hypothesis, the alleles at several *loci* within the inversion present epistatic interactions, or in other words, a specific combination of alleles has a higher fitness than predicted from the sum of the independent effects of each *locus*. Then, the reduced recombination maintains together the coadapted alleles and facilitates their spread through the population, and with them, the inversion also increases its frequency. It has to be taken into account that the allelic content of the inversion can also evolve after its generation. That is, once epistatic interactions between alleles have been established, fitness can still be incremented by recruiting additional beneficial alleles. However, there is not yet any particular example where the coadapted genes on which natural selection is acting to maintain the inversion have been identified. A case that may be an example of epistatic fitness effects is that of inversion *In(het-6)* in *Neurospora crassa* (MICALI and SMITH 2006). In this fungus species, non-self recognition processes that control cell fusion depend on the allelic variants found at several *loci*. One of these *loci* is a complex of two genes, *un-24* and *het-6*, that present strong linkage disequilibrium and are contained within a 18.6-kb inversion. It is remarkable that the allele combinations held together by the inversion have a higher fitness than the other combinations, which produce fewer viable progeny when generated artificially. This suggests that alleles of both genes combine to influence fitness (MICALI and SMITH 2006). Also, SCHAEFFER *et al.* (2003) examined DNA variation within eight genes and patterns of linkage disequilibrium among these genes on *D. pseudoobscura* chromosome 3 and found that, unlike genes outside the arrangements, those genes within the inverted regions showed fixed differences among chromosomal arrangements, and that there

were tight associations among some genes even when they were not adjacent, presumably because selection favors certain allele combinations.

An alternative hypothesis is that formulated by KIRKPATRICK and BARTON (2006), who suggest that locally favorable alleles kept together by an inversion that prevents recombination will present an advantage with respect to the same alleles in normally recombining chromosomes, if the population receives immigrants carrying disadvantageous alleles in those same *loci*. This occurs because favorable alleles in inverted chromosomes will never suffer the disadvantage of being found on a chromosome carrying deleterious immigrant alleles at the other *loci*. The favorable alleles on the inversion have then a higher fitness and, as their frequency increases in the population, the inversion spreads with them. In this case no epistasis is required between the adapted alleles to explain the advantage of the inverted chromosomes. The local adaptation mechanism operates because the inversion binds together alleles that are individually favored in a local population, with the result that each allele is always associated with the superior genetic background provided by the (potentially independent) fitness advantage of certain alleles at the other *loci*. It is important to mention that when we refer to different alleles in the two hypothesis above, we mean not only alleles with changes in the coding region, but also alleles that result in a different temporal or spatial expression pattern due to a change in their regulatory regions, since differential expression is also a mechanism to generate variation that can be selected (CARROLL 2005, WRAY 2007).

Finally, the position effect hypothesis postulates that the functional consequences of inversions stem from mutations created by their breakpoints. In this case, the inversion itself is the target of selection. The breakage and repair of DNA, together with the switching of the relative position of a segment of DNA sequence, are also genetic consequences of the generation of inversions that can have an impact on genes located close to the breakpoints. Inversion breakpoints can have multiple effects on adjacent genes that range from the disruption of coding sequences to the separation of a gene from its regulatory elements or the contribution of new regulatory sequences (with the corresponding effects on gene expression). However, there is actually little evidence that phenotypes associated with inversions segregating in natural populations are caused by position effects at their breakpoints.

In addition, it has to be taken into account that more than one mechanism could contribute to the increase of frequency of an inversion in the population. For example, it would be possible that a position effect provides the initial advantage for an inversion to be selected, but that later on coadapted gene complexes are established through time within the inverted segment as a consequence of the suppression of recombination in heterokaryotypes. Therefore, molecular studies of inversions are needed to assess which fraction of inversions has functional consequences and the relative importance of the different mechanisms involved.

1.1.4 Traits and genes

A large number of traits have been associated with different inversion polymorphisms in various organisms. In *Drosophila* these traits include viability, developmental time, longevity, body size, reproductive success, fecundity or resistance to extreme temperatures (TABLE 1). Individuals carrying different chromosomal arrangements show differences in these characters, so they are likely to be involved in the adaptation to different environments and therefore related to selection on inversion polymorphisms.

Associations between body size and inversion arrangements are particularly common and have been reported in several species. For example, inversion *2j* in *D. buzzatii* (RUIZ *et al.* 1991) or inversion *In(3R)Payne* in *D. melanogaster* (RAKO *et al.* 2006) alter body size in adult flies. However, it is interesting that these two inversions, both located in the same chromosome arm, have an opposite effect. While inversion *2j* causes an increase of body size in *D. buzzatii*, inversion *In(3R)Payne* decreases it in *D. melanogaster* carriers with respect to their respective standard arrangements, which indicates that size changes in both directions can be beneficial under certain conditions and selective pressures. Trait-inversion associations can be complex and depend on the environment or be sex specific (HOFFMANN *et al.* 2004). Also, inversion polymorphisms might be involved with antagonistic effects on different traits, and therefore in the establishment of trade-offs. This is the case of the arrangements in *D. buzzatii* chromosome 2, which have been associated with a trade-off between fast developmental time and decreased size (RUIZ *et al.* 1991, NORRY *et al.* 1995, BETRÁN *et al.* 1998, FERNÁNDEZ

TABLE 1 | Traits associated with chromosomal polymorphism in several species. Table extracted from HOFFMANN and RIESEBERG (2008).

Taxon	Traits associated with inversion polymorphism
<i>Drosophila</i> (various species)	Body size (including wing and thorax size)
	Wing shape and wing loading
	Resistance to heat and cold
	Longevity
	Developmental time
	Larval to adult viability
	Male mating success
	Fecundity
	Competitive ability
	Male and female fertility
	Starvation resistance
Seaweed fly (<i>Coelopa frigida</i>)	Body size
	Developmental time
	Viability
	Sperm displacement
Blackfly (<i>Wilhelmia paraequina puri</i>)	Male development
Grasshopper (<i>Trimerotropis pallidipennis</i>)	Body size
Grasshopper (<i>Sinipta dalmani</i>)	Body size and male survival
Fruitfly (<i>Rhagoletis pomonella</i>)	Eclosion time and diapause
Mosquito (<i>Anopheles gambiae</i>)	Aridity responses
	Insecticide resistance
Midge (<i>Chironomus ramosus</i> and others)	Nematode resistance
	Body size
Fungus (<i>Neurospora crassa</i>)	Self recognition

IRIARTE and HASSON 2000). Evidences that these two traits are related in a trade-off in *D. buzzatii* have been provided by CORTESE *et al.* (2002), who found that rapid developmental time responded to selection when selected alone or together with small wing length, but not when selected in conjunction with large size. The association between large size and slow development caused by the rearrangements affecting both traits at the same time is thought to be responsible for this lack of selective response.

However, there is currently little information about the genes that influence these traits and that are responsible for the fitness effects of inversions. Attempts to identify the genes

underlying variation in these traits linked to inversions focus on finding genes located within the inverted segment that present linkage disequilibrium with the inversion or among them, a fact that would suggest that they are undergoing selection (either with epistasis or without it). For example, by associating size with microsatellite variation KENNINGTON *et al.* (2007) mapped two regions located inside *D. melanogaster* inversion *In(3R)Payne* (but away from the breakpoints) that are good candidates to contain genes influencing size. These regions present strong linkage disequilibrium with the inversion and show clinal patterns just like the inversion. Also, the human polymorphic inversion mentioned above (STEFANSSON *et al.* 2005) is associated with expression changes in *MAPT*, a gene located within the inverted segment that shows a lower expression level in inverted H2 haplotypes (MYERS *et al.* 2007). In spite of studies like these, little progress has been done in the identification of the precise genes underlying these traits, and there is not yet a single inversion for which the genetic basis of its effect on fitness is known. As we will see in the next section, the alternative hypothesis of breakpoint mutational effects is not usually explored.

1.2 Position effects of inversion breakpoints

Apart from the effects on recombination derived from the inversion of a substantial portion of a chromosome, it is well-established that the presence of breakpoints (not only from inversions, but also deletions, translocations, etc.) in the vicinity of genes can account for alterations in their correct spatial, temporal or quantitative expression (SPERLICH 1966). Specifically, inversions can modify the expression of genes adjacent to their breakpoints either by directly disrupting the coding region, by separating regulatory elements (promoters, enhancers, transcription factor binding sites, etc.) from the corresponding coding regions, or by placing a gene under the control of new regulatory sequences or in a different transcriptional environment (for example, in *Drosophila*, genes relocated close to heterochromatin can be silenced by position effect variegation). A simplified representation of the different basic mechanisms by which inversion breakpoints could alter gene expression of adjacent genes can be found in FIGURE 2.

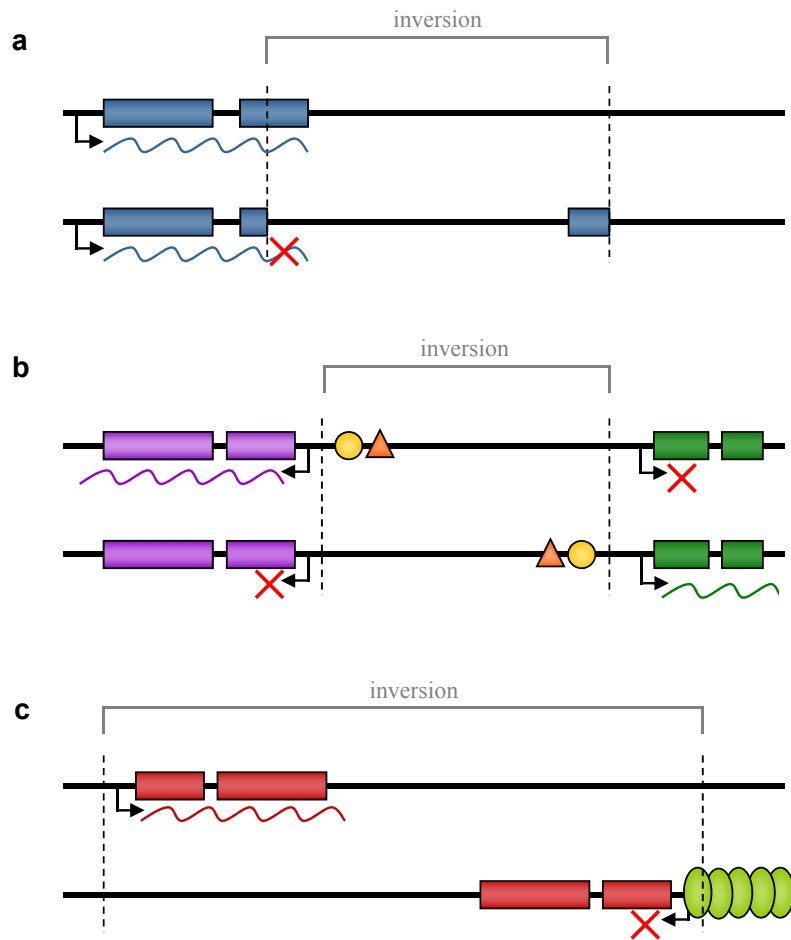


FIGURE 2 | Examples of possible mechanisms for position effects altering the expression of genes flanking inversion breakpoints. For each part of the figure, the top image corresponds to the standard non-inverted arrangement and the bottom one to the inverted chromosome. The inverted segment is indicated above each representation. Dashed lines mark the inversion breakpoints and colored boxes represent exons. Arrows indicate the direction of transcription. **(a)** Disruption of a coding region. The breakpoint can be positioned either within an exon or an intron. In this case the gene could be transcribed but the mRNA would be translated into a truncated protein. **(b)** Removal or addition of regulatory elements. The yellow and orange figures represent the regulatory sequences (promoters, enhancers, transcription factor binding sites, insulators, etc.) required for the proper spatial, temporal and quantitative expression of a gene. The removal of regulatory elements can prevent transcription of a gene in a particular tissue or time. Also, a gene can be placed next to new regulatory sequences that can cause its expression in the inverted chromosome in times or tissues where it is not normally expressed. Regulatory regions and coding sequence can also be separated by the breakpoint if the gene is located inside the inverted segment and the regulatory elements are found on the outside. **(c)** Relocation of genes. Those genes situated inside the inverted segment, change their relative position inside the chromosome and an inversion can place them close to heterochromatic regions, which can cause gene silencing through phenomena like position effect variegation (PEV). The heterochromatic DNA organization (represented with green ovals) can spread into the juxtaposed euchromatic DNA, silencing the nearby gene in a stochastic manner.

Many examples of genes affected by inversion breakpoints have been described in the literature, although most cases correspond to mutations with deleterious effects for those individuals carrying them. In some cases, inversions disrupt the coding region of a gene, causing the loss of its function. For example, a pericentric inversion truncates the *creA* gene in the fungus *Aspergillus nidulans*, which results in carbon catabolite derepression (ARST *et al.* 1990). Also, the proximal breakpoint of mouse inversion *rump white* disrupts gene *Dpp6*, a gene with unknown function that is expressed in embryos and that could be the underlying cause of the embryonic lethality observed in inversion homozygotes. The distal breakpoint of this same inversion is associated with the ectopic expression of the gene *Kit*, which may be responsible for the dominant pigmentation defect presented by heterozygous mice (HOUGH *et al.* 1998). An unusual expression pattern can be caused by the removal or exchange of the regulatory regions of a gene, another of the possible effects of inversion breakpoints. Such a situation occurs in the *Antp^{73b}* mutation in *D. melanogaster*, which is characterized by an inversion that exchanges the first exon and regulatory regions between the *Antennapedia* gene and another gene of unknown function. This results in the ectopic expression of *Antennapedia* in the head during development and causes the replacement of antennae by legs (FRISCHER *et al.* 1986). In mice, an inversion breakpoint located 13.3 kb upstream of gene *Sbb* separates its coding sequence (contained within the inverted segment) from several putative regulatory elements identified by interspecies comparison, causing brachydactyly in the individuals carrying the inversion through the down-regulation and ectopic expression of this gene during development (NIEDERMAIER *et al.* 2005).

In humans, genetic disorders may be caused by breakpoints affecting gene expression through position effects (KLEINJAN and VAN HEYNINGEN 1998), both by the direct disruption of coding regions or by affecting regulatory elements. For example, inversions disrupting the factor VIII human gene are a common cause of severe haemophilia A (LAKICH *et al.* 1993), and, in certain T-cell acute lymphoblastic leukemias, a recurrent inversion in human chromosome 7 that juxtaposes the distal part of the *HOXA* gene cluster to the T-cell receptor β locus leads to the transcriptional activation of genes *HOXA10* and *HOXA11*, whose deregulation had previously been suggested to be a key factor in leukaemic transformation (SPELEMAN *et al.* 2005). In fact, the detection of chromosomal rearrangements in patients with a genetic disease, followed by the characterization of the corresponding

breakpoints, can be used as a strategy to identify candidate genes that may be altered and that could be the underlying cause of the disorder. For example, the distal breakpoint of a paracentric inversion in chromosome 10 detected in individuals with autism, at the same time disrupts gene *TRIP8*, a transcriptional regulator, and abolishes expression of gene *REEP3*, 43 kb away from the breakpoint, putting forward both genes as candidates for autism (CASTERMANS *et al.* 2007). Also, using this method it has been established that long-range position effects caused by inversion breakpoints that down-regulate *TRPS1* expression are the probable cause of hypertrichosis in Ambras syndrome patients in humans and of the *Koala* phenotype in mice (FANTAUZZO *et al.* 2008).

However, the effects of inversion breakpoints on adjacent genes do not always have to be deleterious for the individuals carrying them. They can also be neutral or even beneficial. In an *Antirrhinum* inversion, the excision of one of the *Tam3* TEs that generated the rearrangement, placed the gene *nivea* under the control of the regulatory sequences of the gene *cycloidea^{radialis}*, which resulted in a novel distribution of anthocyanin pigment in the flower tube (LISTER *et al.* 1993). Extreme examples are those provided by the bacterial systems that control the expression of pili in *Escherichia coli* (ABRAHAM *et al.* 1985) or flagellar phase variation in *Salmonella typhimurium* (SILVERMAN and SIMON 1980), where transition between alternative phenotypes is accomplished by the recurrent inversion of certain segments of DNA that cause the activation or repression of adjacent genes. An unusual number of DNA inversion events that potentially control expression of many different components (including surface and secreted components, regulatory molecules, and restriction-modification proteins) has also been identified in *Bacteroides fragilis*, an opportunistic pathogen and inhabitant of the normal human colon microbiota (CERDEÑO-TÁRRAGA *et al.* 2005), revealing that this peculiar regulatory mechanism involving inversion of DNA segments is quite frequent in bacteria.

It is important to note that in most of the examples of position effects mentioned above there is a very clear distinction between the wild-type and mutant phenotypes, the latter being often deleterious and caused by the inverted arrangement. In these cases, the changes on gene activity derived from the inversion breakpoints have major consequences, which manifest in dramatically different phenotypes or diseases, and the fitness of those individuals carrying these rearrangements can be seriously impaired. Although all these mutations are

spontaneous, many of them have been isolated in conditions that facilitate the survival of carrier individuals, such as laboratory populations, or in human patients suffering an important disease. However, these circumstances are not representative of what these individuals may have encountered if they lived in a natural environment where their probability of producing offspring could be significantly reduced.

On the contrary, inversions found segregating in *Drosophila* natural populations have passed the filter of natural selection, and we do not expect them to cause major phenotypical effects, since this kind of changes are more likely to be eliminated by natural selection. This fact implies that the effects of natural inversions on the expression of genes adjacent to their breakpoints will probably be more subtle, and therefore, more difficult to detect. Indeed, traits associated to natural inversions are often quantitative (like body size, fertility, or resistance to heat and cold; see TABLE 1 for a more detailed list) and not always easy to measure. Also, inversions can affect traits important for survival in developmental stages other than the adult phase (e.g. developmental time or larval to adult viability), which further complicates the detection of differences between chromosomal arrangements. Small differences in any of these traits can result crucial for the fitness of the individuals carrying each arrangement, but they are definitely not as visible or apparent as the defects observed for those inversions generated in the laboratory or that result in human diseases.

The study of position effects in natural inversions is also limited by the fact that the specific location of the breakpoints (needed to identify the closest genes both inside and outside the inverted segment) is not known for most inversions. To start with, the sequencing of inversion breakpoints is a complicated task. In addition, so far the goal of those researchers undertaking it has been to determine the mechanism that generated the inversion, and expression analyses to detect possible position effects on nearby genes have not been usually performed (MATHIOPOULOS *et al.* 1998, ANDOLFATTO *et al.* 1999). As a consequence, few examples have been reported where the expression of genes adjacent to the breakpoints has been examined after their precise characterization. KEHRER-SAWATZKI *et al.* (2005) identified the breakpoints of the pericentric inversion differentiating human chromosome 4 from chimpanzee homologous chromosome 3 and found that *C4orf12*, a novel gene originated during mammalian evolution whose 3' end was located within the inverted segment 30 kb

away from one of the breakpoints, showed a 33-fold decrease in its expression level in chimpanzee with respect to humans in fibroblast cell lines. Also, a 3-fold higher expression level was detected in humans for gene *WDFY3*, located ~50 kb away from the same breakpoint and transcribed from a bidirectional promoter shared with *C4orf12*. Interestingly, gene *WDFY3* had been previously reported to present a 4-fold difference in expression between human and chimpanzee cerebral cortex (CÁCERES *et al.* 2003). However, further analyses are required to determine whether or not these expression differences are related to the adjacent inversion. Another putative position effect was detected for polymorphic inversion *In(3L)Payne* in *D. melanogaster*, where a Northern blot suggested that the distal breakpoint disrupted three transcripts normally expressed in the standard arrangement (WESLEY and EANES 1994). Nevertheless, the function of these transcripts and the consequences of their disruption remain unknown.

More recently, new methods have been developed that will facilitate to a large extent the detection of novel inversions in multiple species, as well as the detailed study of the ones that are already known. For example, paired-end mapping techniques have allowed the identification of many new human inversions (KORBEL *et al.* 2007, KIDD *et al.* 2008). Unusual patterns of linkage disequilibrium among alleles from high-density markers (such as SNPs) provide another method to locate inverted segments (BANSAL *et al.* 2007). And of course, the direct comparison of sequenced genomes could represent the most accurate approach (FEUK *et al.* 2005, CLARK *et al.* 2007, LEVY *et al.* 2007, RANZ *et al.* 2007). These techniques will increase greatly the number of available molecularly characterized inversions where inversion breakpoints have been precisely located and the adjacent genes (which are candidates to present expression changes caused by the breakpoints) can be identified. The availability of the data required to undertake the study of position effects will thus likely facilitate and enhance this kind of analyses in the next years.

1.3 Effects of transposable elements on gene expression

Transposable elements (TEs) have been found in the genomes of all kinds of organisms and constitute a large fraction of most eukaryotic genomes. They are DNA

sequences that are able to move and replicate within the genome using different replicative strategies that involve either RNA (class 1 or retrotransposons) or DNA intermediates (class 2 or DNA transposons) (Box 1). TEs have been commonly considered as selfish genomic parasites because their success (that is, their increase in copy number) is usually negatively correlated with the fitness of the host. However, it is undeniable that their presence in

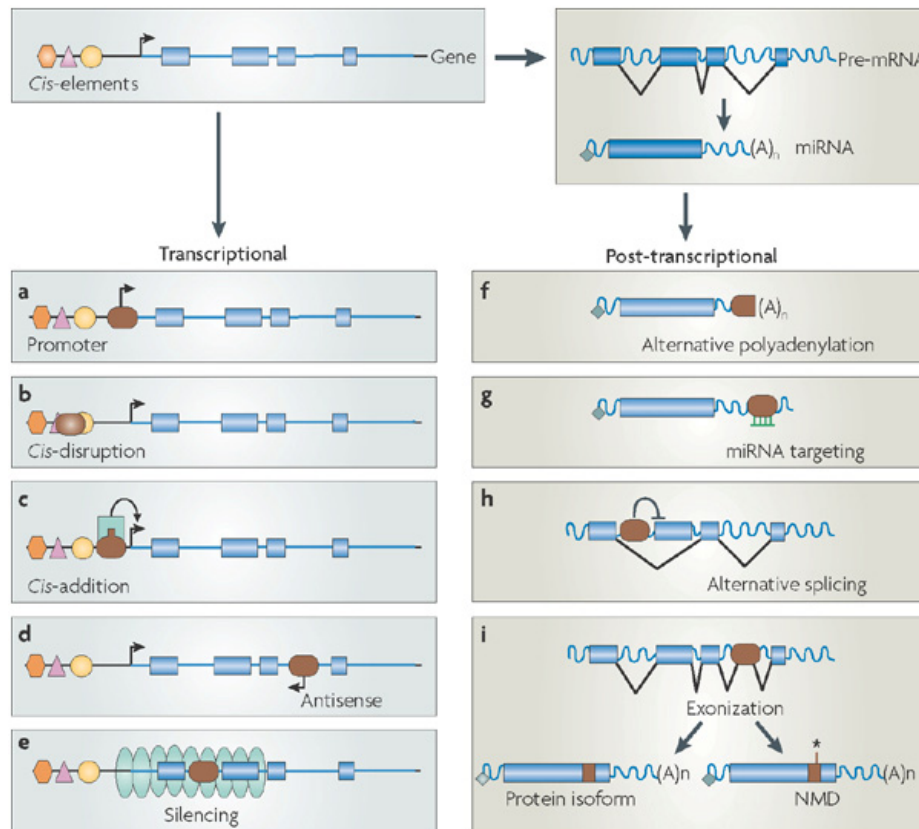


FIGURE 3 | TEs can influence gene expression through different mechanisms. At the transcriptional level (left column), a TE (shown in brown) that has inserted upstream of a gene can insert promoter sequences or introduce an alternative transcription start site (a), disrupt existing *cis*-regulatory elements (b), or introduce a new *cis* element such as a transcription factor binding site (c). In addition, a TE that has inserted within an intron (or in the 3' UTR) can drive antisense transcription and potentially interfere with sense transcription (d), or serve as a nucleation center for the formation of heterochromatin (green ovals), potentially silencing the transcription of adjacent genes (e). At the post-transcriptional level (right column), a TE that has inserted in the 3' UTR of a gene can introduce an alternative polyadenylation site (f) or a binding site for a microRNA (g) or for an RNA-binding protein (not shown). A TE that has inserted within an intron can interfere with the normal splicing pattern of a pre-mRNA (h), causing various forms of alternative splicing like intron retention or exon skipping, or if it contains cryptic splice sites the TE can be incorporated into the coding region (exonized) as an alternative exon (i). This can result in the translation of a new protein isoform or in the destabilization or degradation of the mRNA by the nonsense-mediated decay (NMD) pathway, especially if the exonized TE introduces a premature stop codon (represented by an asterisk). Figure extracted from FESCHOTTE (2008).

genomes creates variability that can contribute to host evolution in multiple ways (McDONALD 1995, KIDWELL and LISCH 2001) ranging from generating large-scale chromosomal rearrangements to altering gene expression or providing functional elements to genomes (new genes, exons, etc.). In particular, multiple mechanisms have been described through which TEs can affect gene expression of adjacent genes (FIGURE 3). Next we will review some examples to illustrate the enormous potential of TEs as modifiers of gene expression.

As mentioned above, TEs can originate inversions (among other chromosomal rearrangements) through different mechanisms like ectopic recombination between oppositely oriented copies or abnormal transposition events (DELPRAT *et al.* 2009) (see APPENDIX IV). In those inversions where TEs have been involved, it is possible that they remain at the breakpoint junctions in inverted chromosomes, where the absence of recombination hinders their elimination. In this situation where TE insertions are associated to the presence of the inversion, there is the possibility that any observed phenotypical effect can be caused not only by the inversion itself (through the inclusion of certain combination of alleles within the inverted segment or by position effects at the breakpoints), but also by the TEs, which as we will see can modify by themselves the expression of nearby genes through countless mechanisms.

TEs can cause a great variety of mutations derived from their insertion, excision or transposition mechanisms. As a result of these processes variability that might be useful to increase the population adaptive potential is generated. Just as with any other type of mutations, any changes originated by TEs can turn out to be detrimental, neutral or result advantageous for the individuals carrying them. If the effects are beneficial, the causal TE insertion will increase its frequency in the population and become fixed. This re-use of TEs to serve cellular functions is known as co-option or exaptation of TEs (MILLER *et al.* 1999, BOWEN and JORDAN 2007). The involvement of TEs in the origin of new genes is one of their most surprising contributions to host evolution and provides irrefutable evidence that these exaptation processes actually take place in nature. Genes can be created completely *de novo*, like the coding sequences of human genes *AD7C*, encoding a neuronal thread protein, that is 99% made up of a cluster of *Alu* sequences, or *BNIP3*, which encodes a protein involved in

Box 1 | Classes of transposable elements

TEs have been found in almost all eukaryotic species investigated so far, with *Plasmodium falciparum* being the only known exception, and they can represent a substantial portion of genomic DNA in some species (up to 80% of the genome in some plants). Given the abundance and diversity of TEs, WICKER *et al.* (2007) have recently proposed a hierarchical TE classification system that includes all types of TEs described so far (FIGURE 4).

Eukaryotic TEs are divided into two classes, according to whether their transposition intermediate is RNA (class 1) or DNA (class 2). Class 1 TEs (or retrotransposons) transpose via an RNA intermediate. A transcript encompassing the whole element serves as mRNA and is reverse-transcribed by a TE-encoded reverse transcriptase (RT) into a cDNA that integrates into the genome. Retrotransposons can be divided into five orders based on their features, organization and RT phylogeny: LTR retrotransposons, LINES, SINEs, *DIRS*-like elements, and *Penelope*-like elements.

LTR retrotransposons are often the major contributors to the repetitive fraction of large genomes, being the predominant order in plants. These elements can be as long as 25 kb and are flanked by long terminal repeats (LTRs) in direct orientation that range from a few hundred base pairs to more than 5 kb. Autonomous elements contain at least two genes: *gag*, that encodes a capsid-like protein, and *pol*, which encodes a polyprotein that includes an aspartic proteinase (AP), reverse transcriptase (RT), RNase H (RH) and DDE integrase (INT). Retroviruses encoding envelope proteins (ENV) and endogenous retroviruses (ERV), where the domains that enable extracellular mobility have been inactivated or deleted, are also included as superfamilies into this category.

LINES (long interspersed elements) can reach several kb in length and are the most abundant TEs in many animals. Only the *L1* family represents about 20% of the human genome. Autonomous LINES encode in their *pol* open reading frame (ORF) an RT and an endonuclease (EN or APE depending on the type of activity) that are required for transposition. A *gag*-like ORF is sometimes found 5' to *pol*, but its role remains unclear. At their 3' end LINES display either a polyA tail, a tandem repeat or an A-rich region. SINEs (short interspersed elements) are small (80-500 bp) and originate from accidental retrotransposition of various RNA polymerase III transcripts. They possess an internal RNA pol III promoter that allows them to be transcribed, but they rely on LINES for transposition functions such as RT. SINE superfamilies are defined

based on their origin (tRNA, 7SL RNA or 5S RNA). The best known SINE is the *Alu* sequence, present in at least 500000 copies in the human genome. LINES and SINEs amplify by a mechanism called target primed reverse transcription (TPRT), in which the element-encoded transcript (that is also used to translate the proteins encoded by the TE) re-enters the nucleus where DNA nicked by the element endonuclease is used to prime the reverse transcription of the transcript directly into genomic DNA.

In addition to classical class 1 TEs, two new groups of TEs have recently been described. *DIRS*-like elements contain a tyrosine recombinase (YR) instead of an integrase (INT), and they do not form target site duplications (TSDs) upon insertion. Their unusual termini also suggest an integration mechanism different from LTR and LINE TEs, but the presence of an RT places them in this class. Finally, *Penelope*-like elements (PLE) were first found in *D. virilis* but present a patchy distribution. These TEs encode an RT that is more closely related to a telomerase than to the RT of other TEs.

Class 2 DNA transposons lack an RNA intermediate during transposition. They are divided into two subclasses based on different transposition mechanisms. Subclass 1 comprises TEs of the order TIR, that mobilize through a “cut-and-paste” mechanism and are characterized by the presence of terminal inverted repeats (TIRs) of variable length (ranging from a few base pairs to several hundred). Transposition is mediated by a transposase enzyme that recognizes the TIRs and cuts both DNA strands at each end. Thus, the element is excised from its initial position and inserted into a new site in the genome. The excision can be precise or imprecise leaving a footprint of the element in the donor site or causing a deletion of part of the adjacent DNA sequences. Increase in copy number of these TEs occurs by transposition during chromosome replication from a position already replicated to another one that the replication fork has not reached yet, or by restoring the transposon sequence to an empty original site after its excision using the DNA sequence of the sister chromatid as a template during gap repair. These TEs are found in almost all eukaryotes usually in low to moderate numbers. The nine known superfamilies (FIGURE 4) are distinguished by the TIR sequences and their characteristic TSD size (2-11 bp). Some families include additional ORFs of unknown function and some show a target site preference. Crypton elements, which are found only in fungi, contain a tyrosine recombinase (YR) and form

Box 1 | Classes of transposable elements (continued)

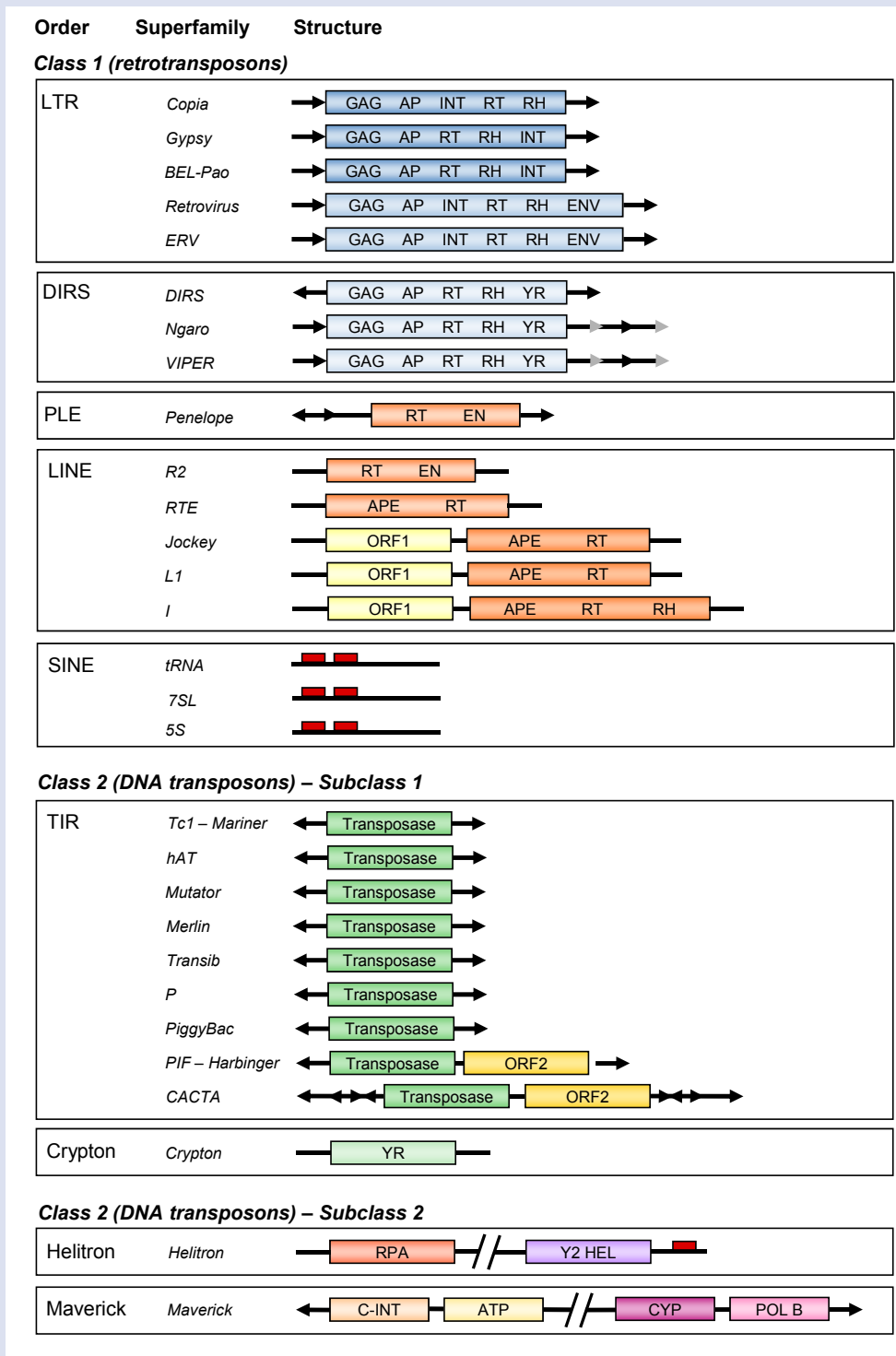


FIGURE 4 | Classification and structure of TEs. Colored boxes represent coding regions contained within TE sequences. See text in Box 1 for details on the coding regions of each type of TEs. Arrows indicate the orientation of repeats (direct or inverted) of TE ends and internal sequences. Small red boxes denote distinctive features in non-coding regions. Figure redrawn from WICKER *et al.* (2007).

Box 1 | Classes of transposable elements (continued)

an independent order also included in this subclass because they lack an RT, which suggests that they transpose via a DNA intermediate.

Subclass 2 includes TEs that undergo transposition processes that do not entail double-stranded DNA cleavage. Elements in the order Helitron appear to replicate via a rolling-circle mechanism with only one strand cut and do not generate TSDs. They encode a Y2-type tyrosine recombinase with a helicase domain and replication initiator activity (Y2 HEL) and can also encode other proteins. This type of TEs has been described mainly in plants but they are also found in animals and fungi. TEs of the order Maverick are

large, reaching 10-20 kb and have long TIRs. They encode up to 11 proteins that vary in number and order which include a DNA polymerase B (POL B) and an integrase (C-INT) related to those of some Class 1 TEs, but they lack an RT.

Typically, each group of TEs contains autonomous elements with ORFs that encode the products required for transposition, and defective non-autonomous elements, which have lost their coding capability but are still able to transpose using the machinery provided by other autonomous elements because they retain the *cis*-sequences necessary for transposition.

controlling apoptosis and whose coding region is 97% formed by sequences derived from a human endogenous retrovirus (BRITTEN 2004). In both cases, sequence evolution of a group of TEs allowed the formation of a coding sequence that could be translated into a functional protein of some importance for human biology. However, new host genes can also be acquired by adopting TE's own genes used for their replication and transposition. That is the case of syncytin (ERVWE1 *locus*), a gene derived entirely from the *env* protein of the endogenous retrovirus *HERV-W* that is expressed in human placenta, where it induces the formation of syncytia due to its fusogenic properties. It is possible that syncytin has a role in the formation of the trophoblast and, therefore, that a retroviral protein is involved in the normal development of the placenta (MALLETT *et al.* 2004).

TEs can insert in different positions within preexistent genes and in general, their effects on gene expression will be different depending on the place of insertion of the TE (TABLE 2). If their insertion occurs inside a coding sequence, it will most likely cause the inactivation of the gene due to the introduction of a new sequence into the transcript that could alter the reading frame and produce a truncated non-functional protein. Gene disruption is usually associated to detrimental effects. For example, in humans, new integrations of *Alu* and *L1* elements have been reported to cause the inactivation of genes in patients with hemophilia and neurofibromatosis (KAZAZIAN 1998, DEININGER and BATZER 1999). However, gene disruption can also result in neutral or even potentially advantageous mutations. That might be the case of the resistance to some *Bacillus thuringiensis* toxins in the

TABLE 2 | Effects on gene expression of TE insertions within genes. The first column indicates the place of insertion of the TE with respect to the affected gene. Further details about each case are given in the main text and in the corresponding references.

Insertion	Effect	Examples	References
Coding sequence	Gene disruption	<i>Alu</i> and <i>L1</i> insertions cause hemophilia and neurofibromatosis in humans	KAZAZIAN 1998 DEININGER and BATZER 1999
		Resistance to <i>Bacillus thuringiensis</i> toxins in insect <i>Heliothis virescens</i>	GAHAN <i>et al.</i> 2001
		Absence of sialic acid Neu5Ac in humans	HAYAKAWA <i>et al.</i> 2001
		Resistance to an organophosphate pesticide in <i>D. melanogaster</i>	AMINETZACH <i>et al.</i> 2005
	New intron	<i>Dissociation</i> insertion in maize <i>sh2-m1</i> gene	GIROUX <i>et al.</i> 1994
Intron	Splicing defects	Abnormal splicing of gene <i>KPL2</i> causes a sperm defect in pigs	SIRONEN <i>et al.</i> 2006
	Exonization	Short form of human leptin receptor is encoded by an LTR	KAPITONOV and JURKA 1999
Regulatory regions	Disruption of promoter sequences	Reduced <i>Hsp70</i> expression in African <i>D. melanogaster</i> populations	ZATSEPINA <i>et al.</i> 2001 LERMAN and FEDER 2005
	Generation of alternative or main promoters	Alternative promoters for the <i>EDNRB</i> and <i>APOC1</i> genes in humans provided by HERV LTRs	MEDSTRAND <i>et al.</i> 2001
		Different and independently acquired LTRs provide promoter functions for gene <i>NALP</i> in primates and rodents	ROMANISH <i>et al.</i> 2007
	Contribution of new regulatory sequences	Up-regulation of <i>Cyp6g1</i> gene is associated to insecticide resistance in two <i>Drosophila</i> species	SCHLENKE and BEGUN 2004 CHUNG <i>et al.</i> 2007
	Ectopic expression	Salivary amylase expression in humans	TING <i>et al.</i> 1992
	New hormone-response elements	Androgen response of mouse gene <i>Slp</i>	STAVENHAGEN and ROBINS 1988
<i>Alu</i> is responsible for hormone dependence of <i>BRCAl</i> gene		BRITTEN 1996a	
UTRs	Changes in transcription or translation efficiency	Incorporation of <i>Alu</i> , <i>L1</i> and <i>HERV</i> segments into the 5' UTR of human gene <i>ZNF177</i> transcripts	LANDRY <i>et al.</i> 2001
	New polyA signals	<i>L1</i> insertion downstream thymidylate synthase gene in mice	HARENDZA <i>et al.</i> 1990
Adjacent sequences	Generation of new transcripts	Tissue-specific transcripts generated from the outward-reading antisense promoter of <i>LINE1</i> express neighboring genes	NIGUMANN <i>et al.</i> 2002

insect *Heliothis virescens* caused by the TE-induced loss of toxin receptors (GAHAN *et al.* 2001). Another possible example is the inactivation of human *CMAH* gene early in human evolution due to an *Alu* insertion that deleted a 92-bp exon, and that could have conferred protection against certain infectious pathogens by preventing the synthesis of sialic acid Neu5Ac, which is needed to form the cell surface glycoproteins that are bound to initiate infection (HAYAKAWA *et al.* 2001). Also, in *D. melanogaster*, the adaptive insertion of a *Doc* retroelement truncates gene *CHKov1*, but at the same time generates a functional protein that confers increased resistance to an organophosphate pesticide (AMINETZACH *et al.* 2005) (FIGURE 5a). In fact, this ability of TEs to inactivate genes has been used as a tool to experimentally eliminate a particular gene function. The Gene Disruption Project is trying to disrupt each gene in the *D. melanogaster* genome through TE insertions (mostly *P* elements) to create *Drosophila* lines carrying a single mutated gene that can be very useful for the study of the biological roles of each individual gene (BELLEN *et al.* 2004). As a far less frequent event, TE insertions within a coding region have also been reported to be able to create a new intron in maize without affecting the protein sequence (GIROUX *et al.* 1994).

TEs can also insert in intronic sequences where, theoretically, they would be harmless for the host because they should be spliced out with the rest of the intron. However, intronic TE insertions can interfere with the normal splicing of the gene or be recruited as new alternatively spliced exons by using splicing signals located within the TE. Both the inclusion of TE-derived exons or abnormal splicing events will usually be deleterious changes because they will likely cause the premature termination of the peptide sequence, as happens in the pig gene *KPL2*, where the insertion of a retrotransposon in an intron results in the skipping of the upstream exon or the inclusion of intronic sequences in the transcript. In both cases, translation terminates prematurely and the protein is truncated, which causes immotile short-tail sperm defect (SIRONEN *et al.* 2006). Nevertheless, there are cases of TE-derived exons that may be functionally important. For example, the short form of the human leptin receptor is generated by an alternative splicing event taking place within the LTR of an endogenous retrovirus *HERV-K*, which also encodes the terminal 67 amino acids of the protein (KAPITONOV and JURKA 1999). Some evidences indicate that these exonization processes can be quite frequent, at least in humans. For example, *Alu* elements containing splice sites are often oriented in the opposite transcriptional direction with respect to the gene, which has

been interpreted to be because most motifs similar to splice sites are located on their minus strand, and more than 75000 antisense *Alu* elements in introns carry potential functional splice acceptor sites in the human genome (LEV-MAOR *et al.* 2003). Also, NEKRUTENKO and LI (2001) estimated that as much as 4% of human genes included retroelement sequences within their coding regions, usually as distinct exons recruited by splicing.

1.3.1 Regulatory changes

However, it is not necessary to target the coding region to affect gene function and the ability of TEs to alter gene regulation without interfering with the protein sequence itself is well-documented. Numerous different mechanisms have been described by which TEs can up- or down-regulate gene expression or modify the tissues or the timing of expression of a gene (FIGURE 3 and TABLE 2). TEs can provide regulatory elements able to increase the level of expression of a gene. For example, the up-regulation of the cytochrome P450 gene *Cyp6g1* has been associated with insecticide resistance in *D. simulans* (SCHLENKE and BEGUN 2004) and *D. melanogaster* (CHUNG *et al.* 2007) (FIGURE 5b). In both species, increased transcription correlates with the presence of TEs (retroelement *Doc* in *D. simulans* and LTR retrotransposon *Accord* in *D. melanogaster*) inserted ~200-300 bp upstream of the transcription start site (TSS). In *D. melanogaster*, it has been demonstrated that the *Accord* LTR carries tissue-specific enhancers (CHUNG *et al.* 2007) and additional natural alleles have been discovered involving two more TE insertions (*HMS Beagle* and *P*) that also contribute to DDT resistance (SCHMIDT *et al.* 2010). But TEs can also cause the down-regulation of gene expression by disrupting promoter sequences. In some African *D. melanogaster* populations that live at high temperatures, expression of the chaperone *Hsp70* is reduced because of *jockey* and *HMS Beagle* TE insertions in the promoter of two of the *Hsp70* genes that alter promoter architecture (FIGURE 5d). This down-regulation could avoid the detrimental effects of the constitutive expression of this protein involved in thermotolerance, which could seriously reduce fitness (ZATSEPINA *et al.* 2001, LERMAN and FEDER 2005).

TEs can also modify spatial patterns of expression by inducing the ectopic expression of genes. In the human amylase *locus* a *HERV-E* LTR acts as a tissue-specific enhancer that

controls the expression of this protein in saliva. Mice lack salivary amylase and also the TE, so its insertion in the hominid lineage could have induced expression in a novel tissue (TING *et al.* 1992). Besides, TEs contain potential hormone-responsive sites able to control expression of adjacent genes. For example, the 5' LTR of an endogenous provirus inserted 2 kb upstream of gene *S β* in mouse includes a hormone-responsive enhancer that has conferred androgen response to this gene (STAVENHAGEN and ROBINS 1988), and motifs within *Alu* sequences seem to be responsible for the hormone dependence of several human genes like *BRCA1* (BRITTEN 1996b).

These regulatory elements can also come from single-copy genomic sequences from the host that have been relocated by TEs to a new position. The transduction or movement of cellular sequences along with the element during transposition (especially in non-LTR retroelements because of their transposition mechanism) is another consequence of TE activity able to alter the expression of genes nearby the new insertion site and can even contribute to the creation of new genes (XING *et al.* 2006). For example, a sequence adjacent to an *Alu* insertion that was transposed with it into the human interferon γ gene promoter provides two functional transcription factor binding sites (TFBSs) required for the transcriptional activation of this gene (ACKERMAN *et al.* 2002).

Besides TFBSs and enhancers, TEs can also provide alternative promoters for those genes located close to their insertion sites. That is the case of human genes *EDNRB* and *APOC1*, where LTRs from *HERV-E* retrovirus act as alternative promoters. Transcripts from both the native and LTR promoters have been detected in both genes, but in gene *EDNRB*, the LTR-driven transcripts appear limited to placenta and represent a significant proportion of the total transcription, which indicates that the LTR acts as a strong promoter (MEDSTRAND *et al.* 2001). In some other situations, the main promoter of a gene seems to be derived from TEs. For example, gene *NAIP* has multiple promoters sharing no similarity between human and rodents that are derived from independently acquired LTRs (FIGURE 5e). In humans, a *HERV-P* LTR serves as a tissue-specific promoter, active primarily in testis. However, in rodents, where multiple copies of this gene are present, an ancestral *ORR1E* LTR common to all rodent genes seems to be the major constitutive promoter, whereas a second LTR found in two of the mouse genes functions as a minor promoter (ROMANISH *et al.* 2007). Particularly

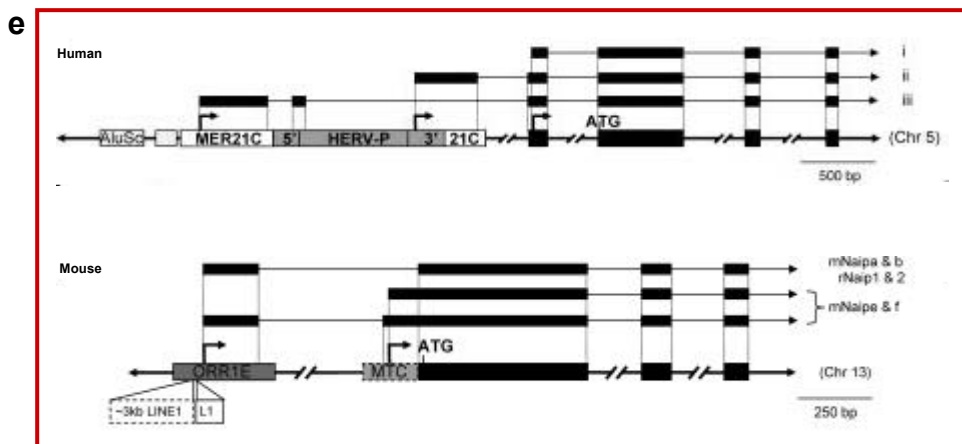
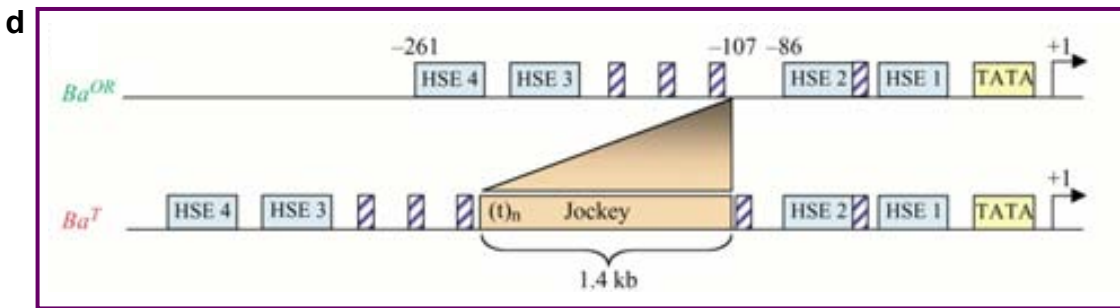
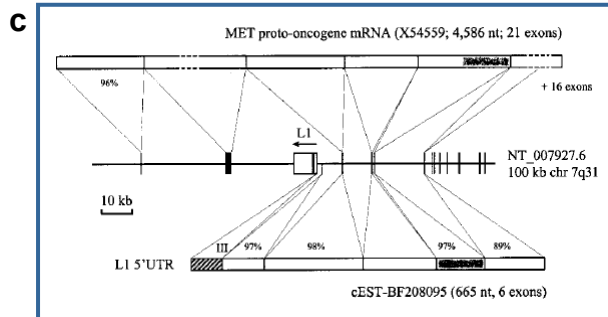
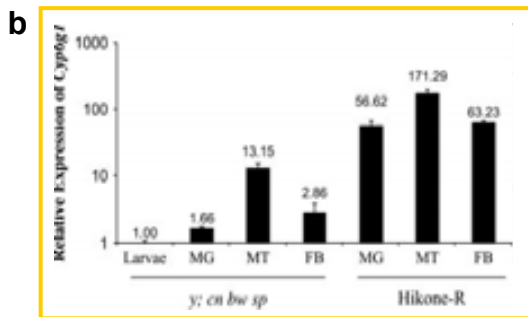
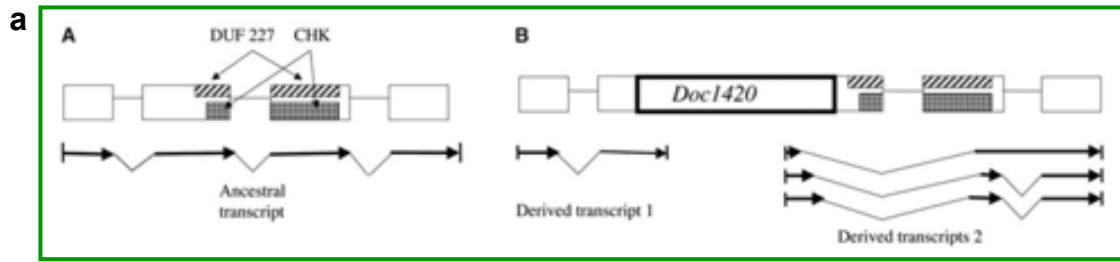


FIGURE 5 | Specific cases of genes whose expression is affected by TEs. (a) *D. melanogaster* gene *CHKov1* before (A) and after (B) a *Doc* insertion disrupts its coding region. The transcripts produced in this *locus* in each case are represented below as black arrows (AMINETZACH *et al.* 2005). (b) An increase in the expression levels of *D. melanogaster* gene *Cyp6g1* in several tissues is detected when a normal line and a strain carrying an *Accord* LTR upstream of this gene (Hikone-R) are compared using real-time RT-PCR (CHUNG *et al.* 2007). (c) Normal transcript of human gene *MET* is shown above a representation of genomic DNA where exons are marked as black rectangles. An RNA transcribed in the same direction that *MET* mRNA but from the antisense promoter of an *L1* copy inserted in an intron of the gene is shown below. This new transcript includes part of the element, some downstream exons of the gene and a new exon. Most of these transcripts originated in *L1* elements do not seem to contain ORFs but they could be involved in transcriptional interference or post-transcriptional silencing of the corresponding cellular genes (NIGUMANN *et al.* 2002). (d) A 1.4 kb fragment of retroelement *Jockey* (brown) is inserted 107 bp upstream of *hsp70Ba* TSS in *D. melanogaster* African T strains disrupting the promoter of this gene by displacing several regulatory elements, represented here as blue or hatched boxes (ZATSEPINA *et al.* 2001). (e) An alternative promoter for human gene *NAIP* is provided by the 3' LTR of an *HERV-P* endogenous retrovirus, while in mice the major promoter of the same gene (*Naip*) derives from another TE insertion, in this case from the LTR of an *ORR1E* element. Arrows indicate TSSs and the final transcripts are indicated above as black bars (ROMANISH *et al.* 2007). Figures extracted in each case from the corresponding cited articles.

striking is the case of human gene *IRGM*, which encodes an immunity-related GTPase that has been recently identified as a risk factor for Crohn's disease. The coding region of this single-copy gene was disrupted by an *Alu* retrotransposition event ~40 million years ago, but the open reading frame (ORF) was reestablished in the common ancestor of humans and apes thanks to an *ERV9* insertion that serves as the functional promoter for the present human *IRGM* gene (BEKPEN *et al.* 2009).

Finally, TEs can insert in the untranslated regions (UTRs) of genes, where they can affect mRNA stability and induce changes in transcription or translation efficiency by introducing new polyadenylation signals, microRNA binding sites or non-coding exons that modify the length of the UTRs. In humans, *Alu* and *L1* insertions form a 5' UTR exon in gene *ZNF177* that exerts a positive transcriptional enhancer effect, but represses translation of this gene (LANDRY *et al.* 2001). In mouse, the thymidylate synthase transcript is polyadenylated at the stop codon because of a *L1* insertion downstream of a cryptic polyadenylation signal (HARENDZA and JOHNSON 1990). Also, the insertion of TEs in the adjacent non-coding sequences can generate new transcripts of a gene. For example, human *L1* retrotransposons have an antisense promoter that drives transcription into adjacent cellular sequences yielding chimeric transcripts that are highly represented in expressed sequence tag (EST) databases (FIGURE 5c). The authors of this study suggests that these *L1*-driven chimeric transcripts can

be a common phenomenon that may involve transcriptional interference or epigenetic control of different cellular genes (NIGUMANN *et al.* 2002).

1.3.2 Epigenetic effects

On the other hand, there are other effects of TEs that derive from the fact that in order to combat the potentially harmful impact of active TEs, the genome has evolved epigenetic defense mechanisms to suppress their activity, that is, to avoid the production of the proteins necessary for their mobilization. These mechanisms include post-transcriptional silencing of TEs by RNAi, chromatin modifications involving methylation of histones or cytosine residues and germline silencing through the Piwi/Aubergine pathway (SLOTKIN and MARTIENSSSEN 2007). However, the ability of TEs to recruit this silencing machinery means that they serve as building blocks for epigenetic phenomena, at the level of single genes or across larger chromosomal regions. Therefore, the silencing mechanisms intended for TEs can affect not only their expression but also the expression of nearby genes. For example, yellow (A^y) mice have an *LAP* retrotransposon inserted 100 kb upstream from the *agouti* (*A*) gene (FIGURE 6b). When this *LAP* element is epigenetically silenced, the agouti protein is expressed specifically in the hair follicle with a determined timing and mice have dark agouti hair. In those cells where *LAP* is active, gene *A* is transcribed from a cryptic promoter in the LTR of the TE and the agouti protein is expressed ectopically, which results in yellow fur, as well as obesity, diabetes and an increased incidence of tumors (MORGAN *et al.* 1999). This phenotype is variegated and most mice present patches of yellow and agouti fur depending on the activity of the TE. Position-effect variegation (PEV) is another phenomenon where the spreading of heterochromatin into adjacent genes silences their expression through an epigenetic mechanism (GIRTON and JOHANSEN 2008). For example, insertions of a transgene containing gene *white* in different sites along *D. melanogaster* chromosome 4, which is mainly heterochromatic, reveal that those insertions located near genes are expressed normally and result in a red eye, while those that fall near or into TEs have a variable expression resulting in variegated eyes (FIGURE 6a). Besides, changes in transgene expression were shown to correlate with switches in chromatin structure (SUN *et al.* 2004).

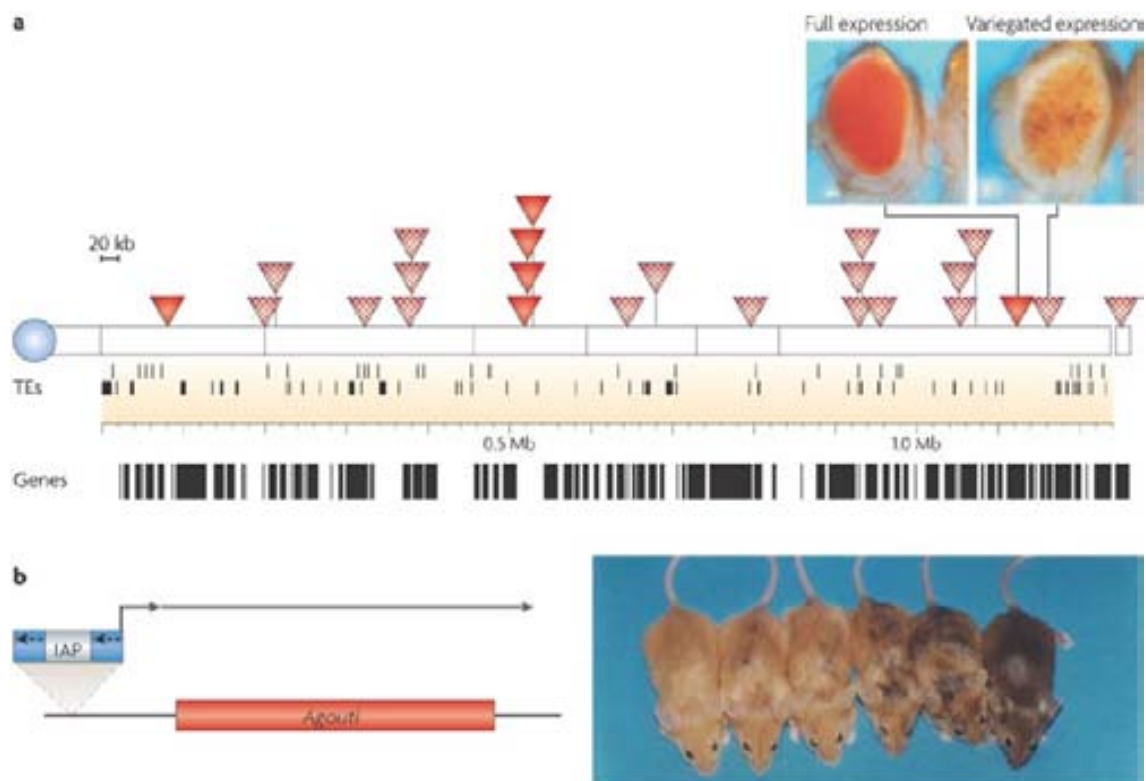


FIGURE 6 | Epigenetic regulation of TEs is able to influence gene expression. (a) PEV on *D. melanogaster* chromosome 4 correlates with the proximity of a transgene to a TE (SUN *et al.* 2004). Horizontal lines represent chromosome 4. Transgene insertions are indicated by triangles showing the resulting phenotype: normal red eye or variegated eye (which is indicative of a silenced gene in those cells that originate the white patches). At the bottom black lines represent both the TE content and gene content along this chromosome. **(b)** In yellow *A^y* mice, an active *LAP* retrotransposon produces an outward-reading transcript that extends into the *agouti* coat-color gene, whose proper expression is subject to the epigenetic status of the TE (silenced or active) (MORGAN *et al.* 1999). The variegated phenotype caused by this partial inactivation of the TE is shown in the picture at the right corner. A mouse with *agouti* hair (and a completely silenced TE) is shown at the right and a yellow mouse (with an active TE) at the left. Extracted from SLOTKIN and MARTIENSSEN (2007).

In fact it is possible that these epigenetic silencing mechanisms also contribute to the function of heterochromatic regions of telomeres and centromeres, which are formed mainly by repetitive DNA and TEs (SLOTKIN and MARTIENSSEN 2007). For example, in *D. melanogaster*, retrotransposons *HeT-A* and *TART* transpose specifically to chromosome ends to maintain telomeres (PARDUE *et al.* 2005). Moreover, TEs can act as insulators that establish and separate chromatin domains with different gene activity. These insulators are able to block enhancer-promoter interactions when located between these two elements. That is the case of *gypsy* endogenous retrovirus in *Drosophila*, whose insulator properties are responsible for the

mutant phenotypes caused by its insertion (GDULA *et al.* 1996). TEs and its epigenetic silencing mechanisms have also been associated to other processes important for the host, like imprinting or V(D)J recombination (SLOTKIN and MARTIENSSEN 2007). Therefore, not only TEs can assume host functions by themselves, but some of them, together with the epigenetic silencing mechanisms, have been co-opted and perform essential roles for the host, especially regarding chromosome structure and function.

All these studies show that TEs already have become a part of the genome, developing important regulatory functions at different levels. However, the global impact of TEs on gene regulation is difficult to assess. Two recent studies have estimated that as many as 25% of all known human (JORDAN *et al.* 2003) and mouse genes (VAN DE LAGEMAAT *et al.* 2003) contain TE elements within their promoter regions, which suggests that TEs have an important role in the evolution of gene regulation. It is important to note that this kind of studies face the additional difficulty that many TE insertions that occurred a long time ago and were selected and fixed can not be recognized as such anymore, because any distinctive sequences (TIRs, TSDs, internal ORFs) have decayed with time in the absence of selective pressure and only those motifs that provided an advantage remain. Recently, the analysis of conserved non-coding sequences in vertebrates allowed the discovery of an ancient previously unknown SINE retroposon family. One copy of this TE originated a neural-specific distal enhancer for gene *ISL1* while another copy, a ~200-bp ultraconserved region that is 100% identical in mammals, contains a 31-aa alternatively spliced exon of gene *PCBP2* (BEJERANO *et al.* 2006). So it is possible that in the future the origin of more conserved non-coding sequences or regulatory elements can be traced back to TEs (LOWE *et al.* 2007).

The striking number of cases of TEs affecting gene expression in multiple species (BRITTEN 1996a,1996b, KIDWELL and LISCH 1997, BRITTEN 2004, MEDSTRAND *et al.* 2005, FESCHOTTE 2008) makes it impossible to regard all these examples as curiosities or strange and infrequent events. Evidence clearly demonstrates that TEs are a highly used and extremely versatile mechanism of evolution. They are able to cause DNA breaks and copy themselves to multiple locations within a genome, which are activities that open up many possibilities to create variation. Besides, they do not represent just random sequences that simply provide raw material on which new regulatory elements can emerge by mutation, but TEs are already

functional sequences carrying their own promoters and regulatory signals as well as coding regions with protein domains. Actually, TEs have even been proposed to be involved in the establishment of regulatory networks due to their multiple copies, the fact that they contain regulatory elements and the DNA-binding properties of the proteins they encode (FESCHOTTE 2008). It has to be kept in mind that evolution is an opportunistic process that will use anything at hand as long as it works, and the activity of TEs is a fast mechanism able to generate variants that might be functional and have an adaptive value.

1.4 The inversion 2j of *D. buzzatii*

Within the genus *Drosophila*, inversion polymorphism has been studied in many species. However, one particular group that has received special attention for many years is the *repleta* group within *Drosophila* subgenus, to which the species *D. buzzatii* belongs. *D. buzzatii* is a cactophilic species originated in South America that colonized the south of Europe, north of Africa, and Australia (FONTDEVILA *et al.* 1981, BARKER 1982, FONTDEVILA *et al.* 1982). This species, whose karyotype consists of five pairs of telocentric chromosomes and one pair of dot chromosomes, has been extensively analyzed at a cytogenetic level, which led to the discovery of several polymorphic inversions, most of them in chromosome 2 (WASSERMAN 1992). During many years numerous studies of inversion distribution in numerous locations (FONTDEVILA *et al.* 1981, FONTDEVILA *et al.* 1982, BARKER *et al.* 1985, HASSON *et al.* 1995), habitat (HASSON *et al.* 1992) and selective effects (RUIZ *et al.* 1991, NORRY *et al.* 1995, BETRÁN *et al.* 1998, FERNÁNDEZ IRIARTE and HASSON 2000), have provided valuable information about *D. buzzatii* ecology and inversion polymorphism. In more recent years several inversion breakpoints of both polymorphic (CÁCERES *et al.* 1999, CASALS *et al.* 2003, DELPRAT *et al.* 2009) and fixed (PRAZERES DA COSTA *et al.* 2009, CALVETE 2010, PRADA 2010) inversions have been sequenced. A BAC genomic library has also been constructed, together with a physical map anchoring contigs assembled by fingerprinting of BAC clones to salivary gland polytene chromosomes (GONZÁLEZ *et al.* 2005). In addition, the *D. buzzatii* genome is currently being sequenced (Alfredo Ruiz, personal communication) using the 454 sequencing technology (MARGULIES *et al.* 2005). All this makes this species an excellent genetic model for in-depth studies of inversions.

In particular, inversion $2j$ is the most common natural polymorphic inversion of *D. buzzatii* (WASSERMAN 1954) and has been studied for a long time. It is a paracentric inversion (it does not include the centromere) that spans one fourth (27%) of chromosome 2, which means that size of the inverted fragment corresponds approximately to 9 Mb (FIGURE 7). Individuals with $2j$ chromosomes can be found in many South-American populations, as well as in Europe and Australia, continents that have been also colonized by *D. buzzatii* (FONTDEVILA *et al.* 1981, FONTDEVILA *et al.* 1982). Inversion $2j$ frequency varies between 0 in some Brazilian populations and 1 in locations in the north of Argentina (HASSON *et al.* 1995), with an average frequency around 60%. This wide geographical distribution together with its high frequency in natural populations suggest that it is a successful inversion that must have some effect on the fitness of individuals carrying it. In fact, some studies have revealed that inversion $2j$ forms a balanced polymorphism with $2st$ non-inverted arrangement that is maintained in natural populations by natural selection acting on adult body size and developmental time. It has been known for a long time that individuals that carry chromosomes with inversion $2j$ exhibit an increased body size and a longer developmental time when compared to flies with $2st$ chromosomes (RUIZ *et al.* 1991, NORRY *et al.* 1995, BETRÁN *et al.* 1998, FERNÁNDEZ IRIARTE and HASSON 2000). Both chromosomal arrangements are thought to be present in natural populations thanks to a trade-off where $2st$ individuals have a shorter developmental time and result in smaller adults, while $2j$ carriers

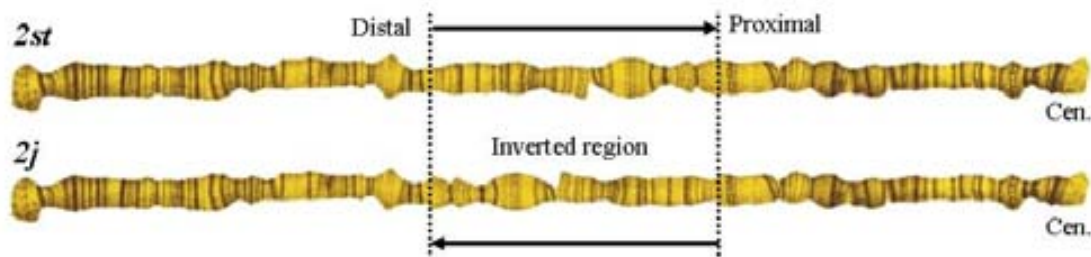


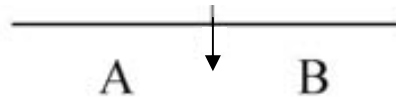
FIGURE 7 | Inversion $2j$ in *D. buzzatii* chromosome 2. The chromosome on top displays the $2st$ non-inverted arrangement and the one at the bottom corresponds to the inverted $2j$ arrangement. Dotted lines indicate inversion $2j$ distal and proximal breakpoints as they can be cytologically observed in polytene chromosomes. Arrows indicate the change of orientation of the intervening segment. The left side of the chromosomes corresponds to the telomere and the right side to the centromere (Cen.).

have a longer developmental time but, in exchange, present a larger body size in the adult stage. Body size is positively correlated with important fitness variables such as mating success, female fecundity, longevity and even tolerance to heat, cold or starvation, and it is therefore subject to intense evolutionary selection (RUIZ *et al.* 1991, EDGAR 2006). On the other hand, a shorter developmental time gives individuals a demographic advantage in earlier stages of development.

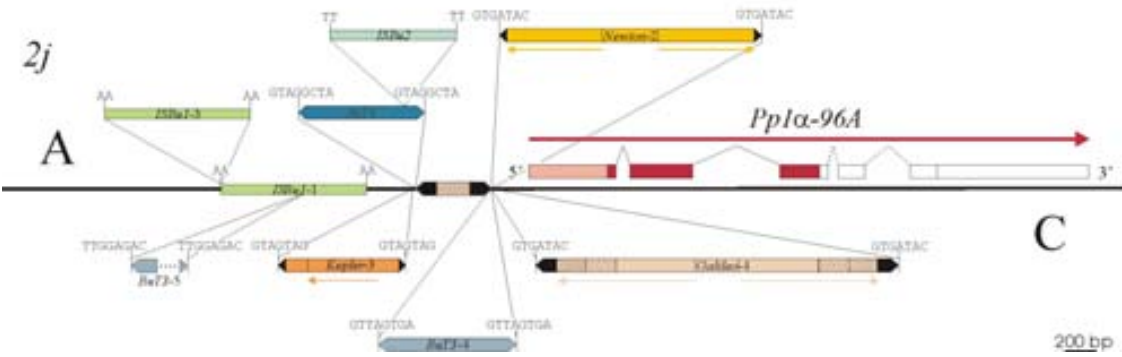
1.4.1 The origin of inversion *2j*

Inversion *2j* is one of the first natural inversions whose breakpoints have been studied at the molecular level and one of the few cases where the mechanism of formation could be elucidated. This inversion was originated by ectopic recombination between two oppositely oriented copies of a new TE named *Galileo* (CÁCERES *et al.* 1999). The two *Galileo* copies found at both breakpoints in *2j* chromosomes presented exchanged 7-bp target site duplications (TSDs), which indicated that the pairing and recombination between these repeated sequences caused the inversion of the chromosomal fragment separating them. No TEs were found in the corresponding sites in *2st* non-inverted chromosomes, where only one copy of the 7-bp TSD was present. Interestingly, the analysis of the breakpoint regions in 39 *D. buzzatii* lines showed that there is a great degree of structural variation in the TE insertions found in *2j* chromosomes (CÁCERES *et al.* 2001) (see APPENDIX I). In fact, these large insertions (1.2-6.3 kb long, depending on the line) are made up of several TEs inserted within or close to the *Galileo* copies responsible for the generation of the inversion. A total of 22 insertions of ten different TEs, 13 deletions, 1 duplication, and 1 small inversion were found in the inverted chromosomes. All this structural variation contrasts with the low variability observed at nucleotide level among *2j* chromosomes, which indicates that these alterations have occurred in a short period of time of only 84000 years (although other evidences postulate that coalescent times were grossly underestimated due to the inclusion of a large number of alleles sampled in recently colonizing populations, and that the age of the sampled *2j* alleles can be closer to 270000 years, LAAYOUNI *et al.* 2003), and suggests that these regions have become genetically unstable hotspots. In addition, the analysis of the nucleotide variation in the sequences adjacent to the breakpoints in several *D. buzzatii* lines with and without the

2st



Distal breakpoint



Proximal breakpoint

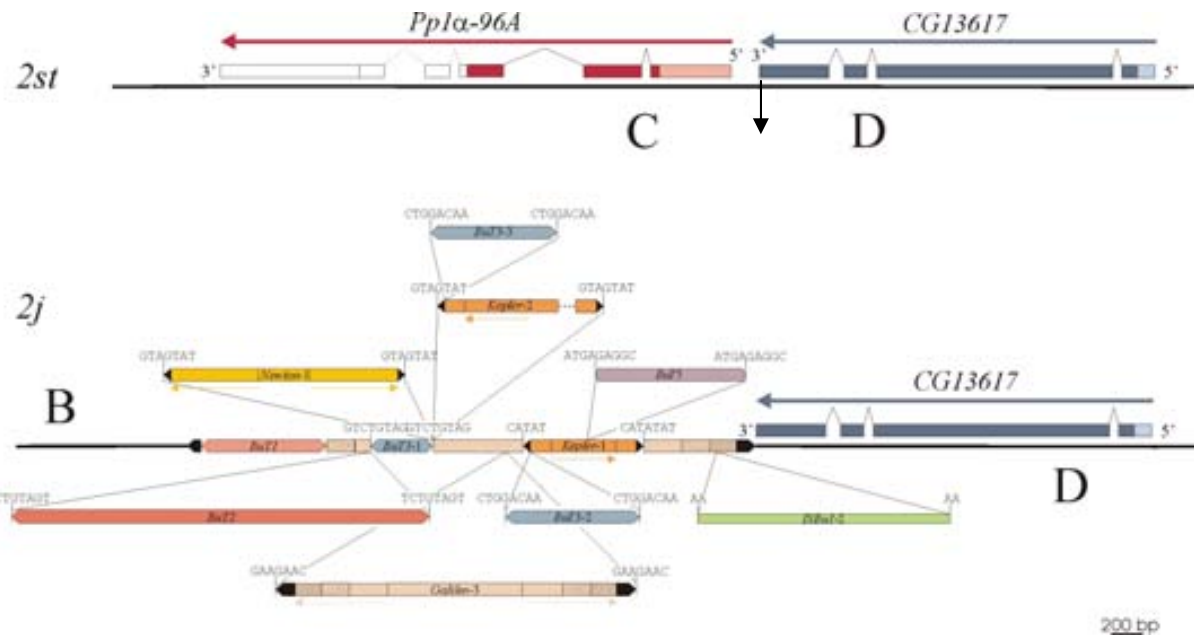


FIGURE 8 | TEs inserted at the distal and proximal breakpoint junctions in different 2j chromosomes. The upper part of each panel represents the breakpoint regions in the 2st arrangement. Sequences flanking the breakpoint sequence, marked by a vertical black arrow, are designated as AB and CD for the distal and proximal breakpoint, respectively. Colored boxes correspond to sequenced exons of genes adjacent to the breakpoint (red for gene *Pp1α-96A* and blue for *CG13617*, see below). Dark colors represent the coding regions of each gene while light colors indicate the 5' and 3' UTRs. In the lower part of each panel, the breakpoint regions in 2j chromosomes carrying the inversion are shown. The single-copy sequences around the inversion breakpoints are designated as A (outside the inverted segment) and C (inside) for the distal breakpoint, and B (inside the inversion) and D (outside) for the proximal breakpoint (CÁCERES *et al.* 1999, CÁCERES *et al.* 2001). In 2j chromosomes, regions B and D (ends of the inverted fragment) are exchanged with respect to the non-inverted arrangement, as can be seen in FIGURE 9. TEs inserted at the breakpoint junction are depicted as colored boxes with pointed ends. Black or patterned sections correspond to *Galileo*, *Kepler* and *Newton* TIRs. The TSD sequences generated upon insertion are shown for each TE. The group of TEs represented on the black line is found in all 2j chromosomes analyzed so far (CÁCERES *et al.* 2001) and it is formed by a small *Galileo* copy and an *ISBu-1* insertion at the distal breakpoint, and three TE insertions (*BuT1*, *BuT3-1* and *Kepler*) nested into a *Galileo* element at the proximal breakpoint. The numbers that appear after some TE names (e. g. *Galileo-3*) indicate the copy of the element of those found at inversion 2j breakpoints. In total, TE insertions at the distal breakpoint are 1.2-3.4 kb long, and at the proximal breakpoint 3.2-6.3 kb long depending on the 2j line.

inversion revealed that inversion 2j is monophyletic (it was generated only once) and that it is approximately 1 million years old (CÁCERES *et al.* 2001, LAAYOUNI *et al.* 2003).

Galileo is a TE that was first described when it was discovered at the inversion 2j breakpoints (CÁCERES *et al.* 1999). A total of four copies of this TE were subsequently sequenced at the breakpoint insertions in the different 2j lines, with two of the copies, the ones that generated the inversion, being present in all the analyzed 2j lines. Also, three copies of a related element that was named *Kepler* and two more of yet another similar TE called *Newton* were also found at the breakpoint insertions of different 2j lines (CÁCERES *et al.* 2001). All of these elements were defective non-autonomous copies that lacked any coding capability and their mechanism of transposition was not clear. Therefore, *Galileo* was initially classified as a *Foldback* element based on its unusual structure (not on sequence similarity) that included extremely long terminal inverted repeats (TIRs). *Kepler* and *Newton* also had long TIRs and presented a 84% nucleotide identity with *Galileo* elements, although the terminal 40 bp were almost identical to those of *Galileo* (CÁCERES *et al.* 2001, CASALS *et al.* 2005). In addition, these three TEs generate 7-bp TSDs upon insertion and show similar sequence insertion preferences (CASALS *et al.* 2005). Only one of the *Galileo* copies at the breakpoints seemed to include a 141-bp ORF with similarity to a transposase (CASALS *et al.* 2005). Recently, thanks to the sequencing of 12 *Drosophila* genomes, the transposase of *Galileo* could be identified in *D.*

mojavensis and a complete copy of *Galileo* could then be amplified, sequenced and assembled in *D. buzzatii* (MARZO *et al.* 2008) (see APPENDIX III) . This complete *Galileo* copy is 5406 bp long and has 1229-bp TIRs and a 2738-bp coding region capable of encoding a 912-aa transposase (even though the sequenced copy has several mutations that prevent the correct translation of a protein). The homology of the transposase with those of *D. melanogaster 1360* and *P* elements allowed the unequivocal classification of *Galileo* as a DNA Class 2 transposon belonging to the *P* superfamily (MARZO *et al.* 2008). The presence of different types of *Galileo* elements in the *D. mojavensis* genome prompted that the related TEs *Kepler* and *Newton* started to be considered as different subfamilies of *Galileo* within the *D. buzzatii* genome (instead of different TEs) and, accordingly, they were respectively renamed as *GalileoK* and *GalileoN* elements, while the original *Galileo* copies were included in the *GalileoG* subfamily.

1.4.2 Genes flanking the inversion *2j* breakpoints

Despite the gained information on the origin of the *2j* inversion, the underlying causes for the phenotypical differences between *2st* and *2j* individuals remain unknown. As mentioned above, one possibility is that the inversion breakpoints affect the expression of adjacent genes. In this particular case, the TE insertions at the breakpoint junctions could also modify by themselves the expression of flanking genes in *2j* chromosomes. Since the location of the inversion breakpoints is known at a molecular level, these nearby genes can be easily identified (FIGURE 9 and TABLE 3) and the search for expression changes can be attempted.

The distal breakpoint (AB) is located downstream of gene *Rox8*, an mRNA-binding protein that is part of the U1 ribonucleoprotein complex and that might be involved in the regulation of alternative mRNA splicing (MOUNT and SALZ 2000, KATZENBERGER *et al.* 2009). In region A, and therefore outside the inverted segment, the coding region of *Rox8* ends approximately 1.5 kb away from the breakpoint in the sequenced *2st* line and is located 2319 bp away from the *Galileo* inserted at the breakpoint in line j-1, even though 841 bp correspond to an *ISBu-1* element insertion (CÁCERES *et al.* 1999). *Rox8* mRNA extends closer to the breakpoint (FIGURE 10a). To try to determine the length of *Rox8* 3' UTR, *D. melanogaster* and *D. buzzatii* non-coding sequences downstream of *Rox8* coding region

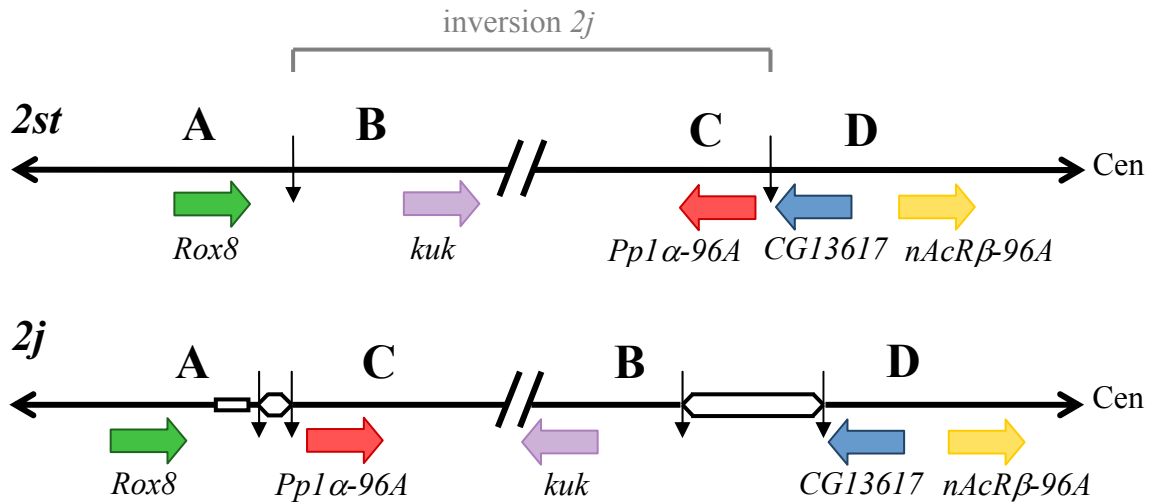


FIGURE 9 | Genes flanking inversion $2j$ breakpoints in $2st$ (top) and $2j$ (bottom) chromosomes. The distal breakpoint is represented at the left and the proximal breakpoint at the right. The centromere (Cen) of *D. buzzatii* chromosome 2 is also indicated. Vertical black arrows mark the breakpoints and white boxes on $2j$ chromosome correspond to the TE insertions found at the breakpoint junctions. The letters A, B, C and D indicate the single-copy sequences flanking each side of the two breakpoints (CÁCERES *et al.* 1999, CÁCERES *et al.* 2001). Note that the positions of B and C (located inside the inverted segment) are exchanged in the $2j$ chromosome with respect to the $2st$ non-inverted arrangement. Colored arrows represent the genes found close to the inversion breakpoints. The arrows indicate the direction of transcription of each coding region.

were aligned and a considerable sequence similarity was found between them, including an extraordinarily conserved 100-bp sequence that presents only five mismatches and one 1-bp indel between these two distantly related species. This relatively high similarity between species is lost abruptly 19 bp (in the *D. buzzatii* sequence) after the only conserved polyA signal (AAUAAA), which strongly suggests that this could be the end of the 3' UTR (FIGURE 10a). This would then result in a 875-bp long 3' UTR located 649 bp away from the distal breakpoint in the $2st$ arrangement, but separated only by 347 bp from the *ISBu-1* insertion present in region A in inverted chromosomes (CÁCERES *et al.* 2001). In region B, also at the distal breakpoint but inside the inverted segment, 1 kb of *D. buzzatii* single-copy DNA was sequenced during the course of this and previous works without finding any significant similarity to any known coding region. However, recent data provided by the ongoing *D. buzzatii* sequencing project have allowed the location of the inversion $2j$ distal breakpoint in the 18-kb long contig 104 and gene *kuk* (*CG5175*) has been identified as the closest gene in region B (Alfredo Ruiz, personal communication). More precisely, the initial methionine of *kuk* is located 2584 bp away from the breakpoint in the $2st$ arrangement. It is interesting that,

according to FlyBase [🔗](#), in *D. melanogaster* this gene has two alternative TSSs located 2468 and 307 bp upstream the coding region, even though the resulting 5' UTRs are shorter (161 and 151 bp, respectively) because both include introns that are spliced out. These two non-coding exons could not be found in the *D. buzzatii* sequence based on sequence conservation, and therefore, the approximate location of the TSS in this species has not been determined. However, the fact that in *D. melanogaster* the gene *kuk* possesses a promoter and a non-coding exon ~2.5 kb upstream of the coding region raises the possibility that the expression of this gene could be altered by the presence of the breakpoint or the TEs in chromosomes carrying inversion *2j* if such elements exist also in *D. buzzatii*. The protein encoded by *kuk* seems to be implicated in the cellularization and morphogenesis of the embryonic epithelium (PILOT *et al.* 2006), but its molecular function is unknown. Unlike the proximal breakpoint, which is located in a highly conserved block of genes, the gene downstream *Rox8* is a gene called *spas* in all the sequenced *Drosophila* species, except for *D. virilis* and *D. mojavensis* (CLARK *et al.* 2007), the two that are phylogenetically closest to *D. buzzatii*. In all these species the gene *kuk* is also located in chromosome arm 3R but in a more distant position with respect to *Rox8*. This change in gene order probably is due to one of the multiple inversions that have caused the complete reshuffling of genes within chromosome arms during *Drosophila* genus evolution (RANZ *et al.* 2001).

With respect to the proximal breakpoint (CD), it is surprising that it is located in a relatively gene-rich region of chromosome 2. Two genes, *Pp1 α -96A* and *nAcR β -96A*, were originally thought to be flanking the proximal breakpoint in the *2st* arrangement in regions C and D, respectively (CÁCERES *et al.* 1999). Gene *Pp1 α -96A* is located inside the inverted segment in region C and thus changes its position in those chromosomes carrying inversion *2j* (FIGURE 9). *Pp1 α -96A* encodes a protein serine/threonine phosphatase involved in amino acid dephosphorylation (DOMBRÁDI *et al.* 1990) and only 667 bp separate its initial methionine from the proximal breakpoint. Its TSS could be as close as 140 bp away from the breakpoint based on sequence homology with *D. melanogaster Pp1 α -96A* 5' UTR (data obtained from FlyBase [🔗](#)) (FIGURE 10b). However, in this case sequence conservation of the non-coding sequences upstream *Pp1 α -96A* coding region is much lower than for *Rox8* downstream sequences and, although the putative *D. melanogaster* TSS is found within a short stretch of sequence (10 bp) conserved in *D. buzzatii* (which might suggest that this sequence is

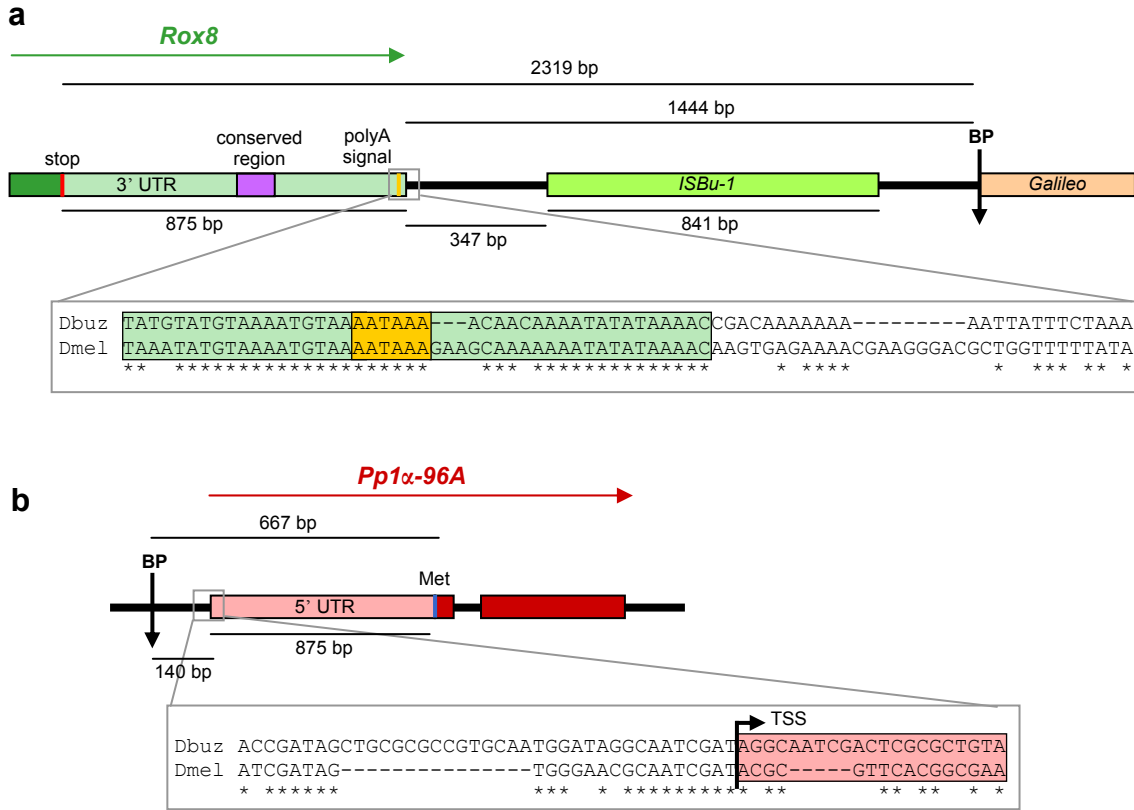


FIGURE 10 | Proximity of *Rox8* and *Pp1 α -96A* transcripts to inversion *2j* breakpoints. (a) End of *Rox8* coding region in *j-1* line. **(b)** Beginning of *Pp1 α -96A* coding region in *st-1* line. Coding sequences are depicted as bright colored boxes while lighter colors represent UTRs. The solid black line corresponds to single-copy non-coding sequences. Colored arrows above each image indicate the direction of transcription and vertical black arrows mark the inversion breakpoints. The lengths of different parts of these sequences are shown above or below a thin bar spanning the corresponding distance. For each gene, the line where this region has been more extensively sequenced has been represented. Both gene sequences are partial since only the last exon of *Rox8* and the first two exons of *Pp1 α -96A* have been sequenced in *D. buzzatii* and are therefore included in the figure. Part of the alignments (performed using MUSCLE [☛](#)) between *D. buzzatii* (Dbuz) and *D. melanogaster* (Dmel) non-coding sequences used to determine the end of *Rox8* 3' UTR and the TSS of *Pp1 α -96A* are shown below each diagram. Identification of the putative end of transcription for *Rox8* gene is based on a sudden decrease in the level of sequence conservation together with the presence of a conserved polyA signal, but a shorter 3' UTR is indicated for this gene in FlyBase [☛](#). For *Pp1 α -96A* we relied on the current information about the TSS available in FlyBase for this gene in *D. melanogaster* to determine a possible start point in *D. buzzatii*. See main text for additional details.

functionally important), the precise location of this gene TSS is not clear (FIGURE 10b). At the other side of the breakpoint, outside the inverted segment in region D, the initial methionine of gene *nAcR β -96A* coding region is located ~ 3.8 kb away from the breakpoint in the sequenced *2st* line and ~ 4 kb in the *2j* line (because of a 166-bp tandem duplication and several small indels). This gene encodes a subunit of a nicotinic acetylcholine-activated channel involved in ion transport (LITTLETON and GANETZKY 2000).

TABLE 3 | Genes adjacent to inversion *2j* breakpoints. Each column in the table indicates the following information: BP, breakpoint adjacent to each gene in the *2st* arrangement; Region, according to CÁCERES *et al.* (1999, 2001) single-copy sequence surrounding the breakpoints where each gene is situated; Position, location of each gene with respect to the inverted segment; Orientation, end of the gene closest to the corresponding breakpoint; Coding region, distance from the stop codon or initial methionine of a gene to the breakpoint; UTR, distance from the putative end of the 3' or 5' UTR to the breakpoint. When distances are different in the sequenced *2st* and *2j* lines the two values are given. The points of start and finish of the transcripts have been inferred from sequence conservation in alignments with *D. melanogaster* sequences (see main text and FIGURE 10). Question marks indicate unknown data.

Gene	BP	Region	Position	Orientation	Distance to BP		Function
					Coding region	UTR	
<i>Rox8</i>	Distal	A	Outside	3'	1549 bp (<i>2st</i>)	649 bp (<i>2st</i>)	mRNA binding protein involved in splicing
					2319 bp (<i>2j</i>) ¹	1444 bp (<i>2j</i>) ¹	
<i>kuk²</i>	Distal	B	Inside	5'	2584 bp (<i>2st</i>)	?	protein involved in morphogenesis of embryonic epithelium
<i>Pp1α-96A</i>	Proximal	C	Inside	5'	667 bp	140 bp	protein serine/threonine phosphatase
<i>nAcRβ-96A</i>	Proximal	D	Outside	5'	3785 bp (<i>2st</i>)	3646 bp (<i>2st</i>)	nicotinic acetylcholine-activated cation-selective channel involved in ion transport
					3998 bp (<i>2j</i>)	3860 bp (<i>2j</i>)	
<i>CG13617</i>	Proximal	D	Outside	3'	12 bp	?	unknown

¹ The sequence includes a 841-bp *IsBu-1* insertion. ² Gene identified in contig 104 of the draft assembly of the *D. buzati* genome sequencing project (Alfredo Ruiz, personal communication).

A preliminary analysis of the expression of the two genes initially considered to be the closest to the breakpoints, *Pp1α-96A* and *Rox8*, had been previously performed by semi-quantitative RT-PCR and Northern blot using one *2st* line and one *2j* line (CÁCERES *et al.* 1999). In spite of the proximity of the breakpoint, no expression differences were detected between the two lines for any of these two genes, which led to the conclusion that the adaptive value of the inversion did not seem to be related to mutations caused by the inversion breakpoints. Nonetheless, in 2000, the genome sequence of *D. melanogaster* was completed (ADAMS *et al.* 2000), and a new putative ORF, *CG13617*, was predicted between genes *Pp1α-96A* and *nAcRβ-96A*. This novel gene would be located in region D with respect to the proximal breakpoint in *D. buzati*, and therefore outside the inverted segment (FIGURE 9 and TABLE 3). The comparison of region D single-copy sequence with the *D. melanogaster* genome revealed that *D. buzati* region D indeed showed homology with the last exons of *CG13617* (CÁCERES *et al.* 2001). A careful annotation of the available sequence unveiled that

CG13617 stop codon was located only 12 bp away from inversion *2j* proximal breakpoint. Thus, gene *CG13617* became the closest gene to any of the breakpoints and its astonishing proximity made it a perfect candidate to search for position effects of either the inversion or the inserted TEs on its expression, reopening the possibility that one of the breakpoints affected somehow the expression of a gene and contributed to the evolutionary success of inversion *2j*.

1.5 Objectives

The sequencing of an increasing number of complete genomes has revealed a great degree of structural variation, a type of genetic variation whose importance had previously been overlooked in many species. In this work, we have focused on chromosomal inversions, a type of chromosomal rearrangement that is known to be maintained by natural selection in *Drosophila*. Molecular mechanisms underlying the effects of inversions remain unknown and, even though several theories are able to explain the spreading and maintenance of inversions in populations, evidence of what is actually happening in natural inversions is scarce. The main goal of this thesis is to investigate how inversions are able to affect phenotype and become the subject of natural selection. In order to do this, we have studied a particular case: inversion *2j* of *D. buzzatii*.

Inversion *2j* has been extensively studied in our research group: from the determination of its phenotypic effects on size (RUIZ *et al.* 1991) and developmental time (BETRÁN *et al.* 1998), to the elucidation of the mechanism that originated the inversion (CÁCERES *et al.* 1999) and the molecular characterization of the TE insertions at the breakpoint junctions in several inverted chromosomes (CÁCERES *et al.* 2001). To date, there is not a single case of an inversion for which the whole story can be told: from the mechanisms responsible for its origin to the genetic basis of how it is able to affect phenotypic traits. This work represents the next step in the study of this well-known polymorphic inversion: the assessment of the presence of expression changes in the genes adjacent to the breakpoints. These position effects could be caused either by the TEs inserted at the breakpoint junction

or by the inversion of a substantial portion of the chromosome. Specifically, we will focus on *CG13617*, a novel gene whose coding region is the closest to one of inversion *2j* breakpoints.

This work is divided in two parts. In the first part we characterize gene *CG13617* in *D. buzzatii*, we compare its expression between lines carrying inversion *2j* and those with the *2st* non-inverted arrangement to identify any differences that might be caused by the inversion, and finally we try to determine the molecular mechanisms responsible for the detected expression changes. In the first part we intend to answer the following questions:

1. Is there a position effect of inversion *2j* proximal breakpoint on *CG13617* expression?

Answering this question requires the complete characterization of *CG13617* coding region and transcripts in *D. buzzatii*, as well as a comparative study of *CG13617* expression levels and patterns in *2st* and *2j* lines to find any differences between them that could be associated to the presence of inversion *2j*.

2. Which are the molecular mechanisms that cause the expression change in inverted chromosomes?

The objective is to identify the change in DNA present in *2j* chromosomes that causes the differential expression with respect to the *2st* arrangement. It is important to distinguish between the effects originated by the TEs inserted at the breakpoint junction in inverted chromosomes or those generated by the inversion itself that changed the sequences downstream of *CG13617* coding region. Different possible mechanisms will be considered and investigated.

Once established that there is a difference between chromosomal arrangements in *CG13617* expression as well as the mechanism causing it, we try to determine the consequences this expression change might have for flies carrying inversion *2j*. In the second part of the work, we aim to answer these questions:

3. Which are the consequences of *CG13617* expression change at a molecular level?

The complexity of the characters affected in individuals carrying inversion *2j* (adult body size and developmental time) suggests that many genes could be contributing to these phenotypic effects. We intend to explore if other genes change their expression levels as a consequence of

CG13617 expression change in *2j* individuals, which could suggest a regulatory role for the protein encoded by this gene.

4. What is the function of gene *CG13617*? We are interested in determining the cellular processes in which *CG13617* protein is involved. Bioinformatic analyses of the protein sequence will be performed to address this question, which will be complemented by the information obtained in the previous objective.

5. Is there any causal relationship between *CG13617* expression change and the phenotypic differences between *2st* and *2j* individuals? Data about *CG13617* function, together with the analysis of the consequences of the expression change experimented by *2j* individuals, can provide valuable information to evaluate the possibility that the altered expression of this gene is the primary cause of the observed phenotypic effects of inversion *2j* affecting adult body size and developmental time, which could represent a mechanism able to explain the adaptive value of this inversion.

MATERIALS AND METHODS

Entonces me contó unas cosas increíbles. Me contó maneras de llegar a la verdad de cualquier tema, no sólo sentándote a pensar en ello como Aristóteles (un señor griego, listo pero confundido), sino saliendo a mirar con tus propios ojos; me habló de hacer hipótesis e idear experimentos, y de comprobar las cosas mediante la observación y llegar a una conclusión.

– JACQUELINE KELLY, *La evolución de Calpurnia Tate* (2009)

Materials and Methods

2.1 *Drosophila* lines

Thirteen *D. buzzatii* lines were used in the different stages of this work. Eleven of these lines (excluding st-13 and st-14, see below) are isogenic for chromosome 2 and all of them are homozygous for different chromosomal arrangements: $2st$ ($n = 7$), $2j$ ($n = 5$), or $2jz^3$ ($n = 1$). These lines represent the natural variability within the species' geographical range (TABLE 4). Also, the $2j$ lines differ in the size and TE content of the insertions at the $2j$ proximal breakpoint (CÁCERES *et al.* 2001).

During the course of the work, two new *D. buzzatii* $2st$ lines were isolated (st-13 and st-14) from different isolines founded with a single female collected from a natural population. Ten isolines from Mazán (Argentina), two from Wari (Peru) and two more lines from Atoqampa (Peru) were available. For each line ten male-female pairs were crossed and after leaving offspring, the DNA of each of the parents was isolated (see below) so that they could be genotyped by PCR using different primer pairs located close to the breakpoints that are specific for the $2st$ and $2j$ chromosome arrangements (see the PCR and RT-PCR section). In two cases (one line from Mazán and one from Wari), we found a pair where both male and female were $2st/2st$ homozygotes and the two new lines were established with their offspring.

Another two *Drosophila* species have been used for specific experiments during this work. The *D. martensis* line MA-4 (Guaca, Venezuela) was used to obtain the sequence of the gene *CG13617* in this species closely related to *D. buzzatii*. Finally, the RNA interference experiments were performed using the *D. melanogaster* wild-type line Canton-S (Bloomington Drosophila Stock Center, stock number 1).

TABLE 4 | *D. buzzatii* lines used in this study. Question marks indicate that the breakpoint junction has not been characterized for these lines. The size of the breakpoint insertions is indicated in base pairs.

Line	Chromosome 2 arrangement	Geographical origin	Type	Proximal breakpoint insertions	
				Size	TE content*
st-1	2st	Carboneras (Spain)	Isogenic	0	–
st-4	2st	Guaritas (Brazil)	Isogenic	0	–
st-7	2st	Termas de Río Hondo (Argentina)	Isogenic	0	–
st-8	2st	Ticucho (Argentina)	Isogenic	0	–
st-11	2st	Trinkey (Australia)	Isogenic	0	–
st-13	2st	Mazán (Argentina)	Isochromosomic	?	?
st-14	2st	Wari (Peru)	Isochromosomic	?	?
j-1	2j	Carboneras (Spain)	Isogenic	4313	<i>GalileoG</i> , <i>GalileoK</i> (2), <i>BuT1</i> , <i>BuT3</i> (2)
j-2	2j	Carboneras (Spain)	Isogenic	4313	<i>GalileoG</i> , <i>GalileoK</i> (2), <i>BuT1</i> , <i>BuT3</i> (2)
j-9	2j	Quilmes (Argentina)	Isogenic	3214	<i>GalileoG</i> , <i>GalileoK</i> , <i>BuT1</i> , <i>BuT3</i>
j-13	2j	Guaritas (Brazil)	Isogenic	3214	<i>GalileoG</i> , <i>GalileoK</i> , <i>BuT1</i> , <i>BuT3</i>
j-19	2j	Ticucho (Argentina)	Isogenic	4724	<i>GalileoG</i> , <i>GalileoK</i> , <i>BuT1</i> , <i>BuT3</i> , <i>GalileoN</i>
jz ³⁻⁴	2jz ³	Tilcara (Argentina)	Isogenic	6341	<i>GalileoG</i> (2), <i>GalileoK</i> , <i>BuT1</i> , <i>BuT3</i> (2)

* The number of copies of each TE is included in parenthesis when more than one copy is present. The exact nature of the breakpoint insertions is described in detail in Figure 2 of CÁCERES *et al.* 2001 (see APPENDIX I).

2.2 Nucleic acid isolation

Genomic DNA was extracted from 0.2 g of adult flies following the protocol described by PIÑOL *et al.* (1988). For the genotyping of the parental individuals to generate new 2st homozygous lines, a protocol based on a buffer containing cetyl trimethyl ammonium bromide (CTAB) (DOYLE and DICKSON 1987) that allows the isolation of DNA from a single fly was used. Plasmid DNA was extracted using standard methods (SAMBROOK *et al.* 1989).

Total RNA was isolated from embryos, larvae, pupae, and adults by using TRIzol[®] reagent (Invitrogen). The embryos were either 0-12 h old or 0-20 h old and were collected by incubating flies in 2% agar plates with yeast during 12 or 20 h, respectively. Larvae samples

included a mix of the three different stages, except in the RNA interference experiments where first instar larvae were collected after microinjection and RNA was extracted following a modified protocol for Trizol adapted to this kind of samples (CARTHEW 2003). This protocol includes a step to wash the halocarbon oil protecting the microinjected embryos with heptane and the use of reduced amounts of the different reagents involved in the RNA extraction.

2.3 RT-PCR and PCR

Around 2-4 μg of total RNA was treated with 1 unit of DNase I (DNA-free™, Ambion) for 30 min at 37 °C to eliminate DNA contamination. cDNA was synthesized from 500 ng or 1 μg of the DNase I-treated RNA by using an oligo(dT) primer able to bind the polyA tail of the mRNA molecules. The reagents used were the Transcriptor First Strand cDNA Synthesis Kit (Roche) for the *D. buzzatii* samples and the SUPERScript™ First-Strand Synthesis System for RT-PCR (Invitrogen) for the *D. melanogaster* ones. Negative reactions without retrotranscriptase were carried out for each sample to control for the presence of contaminant DNA. This step is especially important when the amplified fragments do not include an intron and generate PCR products of the same size whether they are amplified from genomic DNA or cDNA.

PCRs were performed in a total volume of 25 μl , including 1 μl of cDNA or 100-200 ng of genomic DNA, 10 pmol of each primer, 200 μM dNTPs, 1.5 mM MgCl_2 , and 1.5 units of *Taq* DNA polymerase. As mentioned above, whenever possible, primer pairs used in RT-PCRs were selected to span an intron of the gene to differentiate the size of the amplification products from cDNA and genomic DNA. Primer sequences are available in TABLE 5 and SUPPORTING TABLE 3 of PUIG *et al.* 2004 (see Results). Typical cycling conditions were 30 rounds of 30 sec at 94 °C, 30 sec at 55-60 °C (depending on the primer pair used), and 60 sec at 72 °C.

TABLE 5 | Primers used in this study. The table is divided in two parts, the first one containing the primer pairs used in *D. melanogaster* and the second one the primer pairs used in *D. buzzatii*. The T7 promoter sequence attached to the 5' end of the specific sequence of the primer is indicated in lowercase. The amplicon sizes that correspond to the RT-PCR products obtained from a cDNA template are shaded in blue. *D. buzzatii* PCR genotyping primers were designed and used in previous studies (CÁCERES *et al.* 1999, CÁCERES *et al.* 2001). Since these primers are located in non-coding sequences surrounding the breakpoints or inside TEs, the amplified inversion breakpoint is indicated in the Gene column for these primer pairs.

Primer pairs used in *D. melanogaster*

Primer name	Sequence (5'→3')	Primer name	Sequence (5'→3')	Amplicon (bp)	Gene	Use
DmE1	TGGAGGATCTGGAGCGCATA	DmE2	CGGAACTGGCCTCGAACTCA	1154	<i>CG13617</i>	cDNA cloning
T7DmE3	gcttctaatacactactatag CGGAAGCAGCACGAGAGGAT	T7DmE2	gcttctaatacactactatag CGGAACTGGCCTCGAACTCA	633	<i>CG13617</i>	dsRNA synthesis
DmE4	GGAAGATGGTCAACCGGAAG	DmE6	CGTCCTCCGTGGTGTGAA	531	<i>CG13617</i>	RT-PCR
DmH1	ACCGGAGTGTTCACCACCAT	DmH2	CCTCCTCGACCTTAGCCCTTG	499	<i>Gapdb1</i>	
DmU1	GCTCAATGAGGCAACCTTCG	DmU2	CAGCGAGACAAGCGACACAT	85	<i>mus209</i>	Real-time RT-PCR
DmM2-1	GGCCCAAGCTAACGAACATC	DmM2-2	TCTCAATGTGACGCACCGTAA	111	<i>Mcm2</i>	
Mcm5-1	GTCGCTGGCAAAGATTCGTT	Mcm5-2	CCAGTCATCGCAGCATCAAG	102	<i>Mcm5</i>	
Mcm7-1	GCTCAGATGATTCAGGGTTTGC	Mcm7-2	CGTCGTTCTTGTGATACAGATGAT	76	<i>Mcm7</i>	
RnrL1	GTGGGACTGGCAGAAATTGAA	RnrL2	TGGGAGCAACAAGCAGAGAGT	71	<i>RnrL</i>	
DmS1	CCCATCCAGTACCACGACATC	DmS2	ACAAATCCACCTCCTCGACAGT	79	<i>RnrS</i>	
CycE1	CCGCCATCAGTCATACATTTAGTC	CycE2	TCCATCCAGCGAGCACAAG	91	<i>CycE</i>	
Hsp83-1	CGCGCATGAAGGATAACCA	Hsp83-2	TCCACGAAGGCAGAGTTGCT	82	<i>Hsp83</i>	
DmH3	GCGTCATCGACCTGATCAAGT	DmH4	TTGCGGATTATGCAACAGTGA	111	<i>Gapdb1</i>	

Primer pairs used in *D. buzzatii*

Primer name	Sequence (5'→3')	Primer name	Sequence (5'→3')	Amplicon (bp)	Gene	Use	
Dbmus1	TGAGGCACGTTTGGGACAAG	Dbmus2	GCGTATGTGACCCAGATCCCT	718	<i>mus209</i>	Sequencing	
DbM2-1	ATACACGAGGCCATGGAGC	DbM2-2	GGATCGTAGGAGAAGCCATT	774	<i>Mcm2</i>		
DbM5-1	CACGAAGCAATGGAACAGCA	DbM5-2	GCGATATAGCATCTTCCGTTG	750	<i>Mcm5</i>		
DbM7-1	CGTCCCGATGGCATGAAGAT	DbM7-2	GTCGTCGATGCATTTGTCCAC	926	<i>Mcm7</i>		
DbRnrL1	GCCATCATCGAATCTCAAGT	DbRnrL2	GTGCGAAGATAATACATGCC	1003	<i>RnrL</i>		
DbRnrL3*	AGCATTCCAGTCCCAAAGAT						
DbRnrS1	ACGGAGAACGCCAACCAACG	DbRnrS2	AGATGATCTCAATGATGCGCTC	703	<i>RnrS</i>		
DbTs1	TCGTTCCCATTTGCTGACCAC	DbTs2	ATACGGCCATTTCCATTTGAAT	729	<i>Ts</i>		
DbCycE1	GCTAGACTGGCTGATCGAGGT	DbCycE2	TCCATTGTGGTTGTGTGCGT	765	<i>CycE</i>		
DbRG3	TGTATTCAATCTGGATGGCA	DbRG4	TCTGTTTTGATGGCGTAGTC	555	<i>RanGap</i>		
DbHsp1	GAGGACAAGGAGAACTACAAGA	DbHsp2	TAATCGACCTCCTCCATGTG	821	<i>Hsp83</i>		
Dbmus3	AGCGATTCCGGCATTACAG	Dbmus4	TGATCTCAACGTCAACGAAAACAA	72	<i>mus209</i>		Real-time RT-PCR
DbM2-3	GGAGGAGATACCACAGGACTTGTT	DbM2-4	CGATGTTTGTAGCTTAGGTCAAT	80	<i>Mcm2</i>		
DbM5-3	TGTAAATGAAGCACTGCGACTGT	DbM5-4	GCTCCAGCCAGGCTACCA	72	<i>Mcm5</i>		
DbM7-3	ATGTCCAAGGATTCGCTTAACC	DbM7-4	GCAAAGATGCGATCCGAAGT	77	<i>Mcm7</i>		
DbRnrL4	CGATCCAGCAACACGGTATTAG	DbRnrL5	TAAGGCTCGAACGACTCATTGT	100	<i>RnrL</i>		
DbRnrS3	GAAACAATGCCCGCTGTGA	DbRnrS4	AAGTTCGCCGCTTGGAT	71	<i>RnrS</i>		
DbTs3	TCCATACGCTAGGCGATGCT	DbTs4	TTGCGTTTTCAACTGTTCCA	71	<i>Ts</i>		
DbCycE3	GAGCCGTTCTTCCATGTCATATC	DbCycE4	TTTGTCACCTGCTCATTTTGCT	74	<i>CycE</i>		
DbRG5	TGACCACCAATGCTTATACCACTAA	DbRG6	ACGTCGGTAGCCGTATTGTTG	112	<i>RanGap</i>		
DbHsp3	GGAGACCTTGCGCCAGAA	DbHsp4	GAACAGCAGAATGACCAGATCCT	73	<i>Hsp83</i>		
E27	CCCTAAAGACTAAACTCCCACAAA	E25	TTAGCGGTTGTGTTGGATATGG	124	<i>CG13617</i>		
H8	CTGCCAACGGTCCATTGAA	H9	GAGTGAGTGTGCTGAGGAAGTC	82	<i>Gapdh</i>		
A1	ACCGAATCGATCTCAAAGGCT	B1	ATATTCGCGGAGTCAAAGGTTG	370	<i>2st</i> distal BP	PCR genotyping	
A1	ACCGAATCGATCTCAAAGGCT	C1	ACTTCGCCGCATCGCAATCTA	750	<i>2j</i> distal BP		
B1	ATATTCGCGGAGTCAAAGGTTG	G2	TGTGGCATCAACAACCGATCA	450	<i>2j</i> proximal BP		

* Primer located inside the DbRnrL1-DbRnrL2 fragment used only for a sequencing reaction to close a gap.

2.4 RACE

RACE (Rapid Amplification of cDNA Ends) experiments were done with DNase I-treated RNA from embryos of lines st-1 and j-1. 5' RACE was carried out using the 5'/3' RACE kit (Roche). 3' RACE was performed using cDNA synthesized with a primer that includes an oligo(dT) sequence with an extension added to its 5' end that provides an anchoring point for a more specific primer. The gene-specific primers used in each case are listed in SUPPORTING TABLE 3 of PUIG *et al.* 2004 (see Results). All of the amplification products spanned one intron of the gene to ensure that they originated from mRNA. The fragments corresponding to each end of the mRNA were amplified in two rounds using different nested gene-specific primers. In some cases, different amounts of the first amplification were tested as a template for the second one to obtain a specific band. RACE products were cloned into the pGEM-T vector (Promega). A minimum of eight different clones from each reaction were screened by restriction mapping to ensure that they contained the expected gene sequence. In the 3' RACE experiment, three clones were sequenced for each line. In the 5' RACE, two clones from line st-1 and three clones from line j-1 that contained slightly different inserts were selected for sequencing.

2.5 Real-time RT-PCR

Real-time RT-PCRs experiments were performed in an ABI PRISM 7900HT Sequence Detection System (Applied Biosystems) in the first part of the work and an ABI PRISM 7500 Sequence Detection System (Applied Biosystems) in the second part. The amplified products were detected with the dsDNA-binding dye SYBR Green (SYBR Green PCR Master Mix, Applied Biosystems, or iTaq SYBR Green supermix with ROX, Bio-Rad). Real-time RT-PCR reactions were performed in a total volume of 25 μ l in 96-well optical plates. Each reaction included 1-2 μ l of a 1/10 dilution of a cDNA sample, which was synthesized from 0.5-1 μ g of total RNA. Controls without retrotranscriptase were performed during cDNA synthesis for each sample and were included in the real-time RT-PCR plates as negative controls for DNA amplification. Primers were designed by using PRIMER EXPRESS 1.5 software (Applied

Biosystems) taking special care in that they are specific (so that only one PCR product is amplified) and that they do not form secondary structures or primer-dimers that could unspecifically bind SYBR Green and affect product quantification measures (TABLE 5 and SUPPORTING TABLE 3 in PUIG *et al.* 2004, see Results). Primers for the *CG13617* gene were designed in areas conserved between *2st* and *2j* lines. The housekeeping gene *Gapdh* (*Gapdh1* in *D. melanogaster*), which is expressed constitutively, was used as an internal control for differences in cDNA concentration among samples. For each sample, the gene of interest and *Gapdh* were both amplified in triplicate, and results were analyzed by using SEQUENCE DETECTION SOFTWARE versions 1.7 or 1.4 (Applied Biosystems) respectively in each part of the work. Cycling conditions were 95 °C for 10 min and 40 cycles of a denaturation step at 95 °C for 15 sec followed by annealing and extension at 60 °C for 1 min. A dissociation curve step was added at the end of each run to ensure that only one specific product was amplified in each reaction. Relative quantification was performed with the standard curve method, and gene amplification levels were normalized by dividing by *Gapdh* levels in each sample. Expression levels were compared by means of two-level nested ANOVA (mixed model) (SOKAL and ROHLF 2000).

2.6 DNA sequencing

During the cloning of inversion *2j* breakpoints (CÁCERES *et al.* 1999, CÁCERES *et al.* 2001), gene *CG13617* was partially sequenced in the *D. buzzatii* lines st-1 and j-1. The *CG13617* sequence was completed in line st-1 by subcloning the λ st9 phage into the pBluescript II SK vector (Stratagene), and in line j-1 by PCR amplification of different overlapping fragments covering the whole gene. The mRNA of the gene was also amplified in different fragments in line st-1 to verify the exonic structure of the gene, which had been predicted computationally by comparison with the *D. melanogaster* homologous gene (ADAMS *et al.* 2000). PCR and RT-PCR products were gel-purified with GeneClean[®] Spin kit (Qbiogene) and sequenced directly with the same primers. Cloned fragments were sequenced with M13 universal primers. Sequencing was carried out at the Servei de Genòmica of the Universitat Autònoma de Barcelona.

Since the genome of *D. buzcatii* is not yet sequenced, several genes had to be totally or partially sequenced in this species in order to design specific primers to be used in the real-time RT-PCR experiments (TABLE 5). Except for *Gapdh* gene, whose primers were designed in the *D. hydei* sequence, the sequencing primers for the rest of genes were designed in the corresponding homologous genes in *D. mojavensis*, the phylogenetically closest species to *D. buzcatii* with an available sequenced genome (CLARK *et al.* 2007). Each of the genes was identified in the *D. mojavensis* and *D. virilis* genomes, these sequences were aligned (see the Sequence analysis section) and primers were designed in *D. mojavensis* coding regions conserved between both species at the nucleotide level. Primers were chosen to try to amplify fragments approximately 800-1000 bp long in the *D. buzcatii* genome (a size that permits the complete sequencing of the PCR product performing one sequencing reaction from each end), and whenever possible, they were located in two different exons, so that amplification products originated from genomic DNA and mRNA have different sizes. For some genes, we chose to amplify cDNA instead of DNA depending on the distance between the two chosen primers taking into account that cDNA-derived products are shorter than those that come from DNA because they do not include intronic sequences. *D. buzcatii* st-1 genomic DNA or embryo cDNA were used as templates for the amplification, although j-19 genomic DNA or embryo cDNA was always included as a control in order to check that the gene can also be amplified in other *D. buzcatii* lines. All the st-1 amplified products were gel-purified using QIAquick® Gel Extraction (Qiagen) or GenElute™ Gel Extraction (Sigma) kits and sequenced directly with the same primers. In this case the sequencing was carried out at Macrogen Inc. (Seoul, Korea). When *D. buzcatii* sequences were obtained, BLASTN searches on the *D. mojavensis* and *D. virilis* genomes were performed to check that the best hits in each case corresponded to the expected homologous genes. Details about the different sequenced fragments are available in TABLE 8 (see Results), except for the *Gapdh* gene, of which a 1017 bp fragment including the coding region almost completely as well as the 79-bp intron that this gene has in *D. buzcatii*, were amplified and sequenced in order to be used as an internal control in real-time RT-PCR experiments (PUIG *et al.* 2004).

2.7 Sequence analysis

2.7.1 Sequence annotation and comparative DNA sequence analysis

CG13617 coding regions were identified in the *D. buzzatii* sequence using BLASTX (translated nucleotide query *vs.* protein database) (MCGINNIS and MADDEN 2004) and compared with the *D. melanogaster* predicted protein annotation according to FlyBase (TWEEDIE *et al.* 2009). DNA and protein sequence alignments were performed with CLUSTALW (LARKIN *et al.* 2007) and/or MUSCLE (EDGAR 2004) and visualized with BIOEDIT software (HALL 1999). *CG13617 Drosophila* orthologous sequences and annotations correspond to the CAF1 (Comparative Analysis Freeze 1) genome assemblies of the 12 sequenced genomes (CLARK *et al.* 2007) and were obtained from the DroSpeGe Browser (GILBERT 2007) that was accessed through the Assembly/Alignment/Annotation of 12 related *Drosophila* species website. The GLEANR consensus annotation (CLARK *et al.* 2007) was verified for each species by translating the protein encoded by the predicted coding region and comparing with those of *D. melanogaster* and *D. buzzatii*. *CG13617* similarity searches in species outside the genus *Drosophila* were performed by BLASTP (protein query *vs.* protein database) and TBLASTN (protein query *vs.* translated database) (MCGINNIS and MADDEN 2004) using in both cases *D. buzzatii 2st* protein as a query to interrogate the corresponding GenBank non-redundant databases. The search of homologous sequences in the available completed genomes of all organisms yielded no new addition to the list of putative orthologous proteins.

CG13617 upstream flanking sequences were analyzed using mVISTA and rVISTA (LOOTS *et al.* 2002, FRAZER *et al.* 2004) to search for conserved non-coding sequences (CNSs) among the different *Drosophila* species that could indicate the presence of promoter or regulatory elements. The DNA sequences of the five *Drosophila* subgenus species, including from the first two exons of gene *nAcR β -96A* to the first two exons of *CG13617*, were submitted to the VISTA server together with their respective annotations, and aligned with LAGAN (BRUDNO *et al.* 2003) using translated anchoring to improve the alignment of distant homologues. A phylogenetic tree showing the relationships among species was introduced manually. Criteria used to identify significant conserved sequences were 70% identity in a

window size of 100 bp. Both mVISTA and rVISTA tools were used to define the CNSs and rVISTA was also employed to try to identify potential TFBSs. MATCH™ public version 1.0 (Biobase) [↗](#) was also used to search for TFBSs using all matrices from TRANSFAC® database and a cut-off to minimize false positives. Promoter predictions were performed with the 2006 fly version of MCPROMOTER [↗](#) (OHLER *et al.* 2002, OHLER 2006) with the highest sensitivity level, or with the NEURAL NETWORK PROMOTER PREDICTION tool [↗](#) (REESE 2001) at the Berkeley *Drosophila* Genome Project site.

The other genes sequenced in *D. buzzatii* (TABLE 5, see Results) were identified in the *D. mojavensis* and *D. virilis* CAF1 genomes by searching the available annotations in the DroSpeGe Browser (GILBERT 2007) at the Assembly/Alignment/Annotation of 12 related *Drosophila* species website [↗](#). In case of doubt in defining the exact coding regions, the consensus GLEANR annotation (CLARK *et al.* 2007) was followed. Homology with the corresponding *D. melanogaster* gene was always verified by performing BLASTN searches against this genome. Finally, the nucleotide sequences of the two *Drosophila* subgroup species (*D. mojavensis* and *D. virilis*) were aligned for each gene with MUSCLE [↗](#) (EDGAR 2004) in order to identify conserved regions where interspecific primers could be designed.

2.7.2 Protein analysis

Protein sequences were obtained by conceptual translation of the predicted CG13617 coding regions in the different available *Drosophila* genomes or the sequences generated experimentally in this work. To determine the best alignment of the CG13617 proteins of 14 *Drosophila* species, we tried several alignment methods, including MUSCLE, CLUSTALW and T-COFFEE with different parameters. Results of the different alignment methods are very similar and just differ in the position and/or length of some gaps. A final alignment was generated based on the MUSCLE alignment [↗](#) (EDGAR 2004) with some minor modifications according to the regular T-COFFEE alignment [↗](#) (NOTREDAME *et al.* 2000). The multiple sequence alignment was visualized with BIOEDIT (HALL 1999). Identity and similarity values between CG13617 proteins in the different *Drosophila* species were calculated with MATGAT software

(CAMPANELLA *et al.* 2003) based on pairwise alignments performed by the same program using BLOSUM62 as scoring matrix.

The distinct putative protein domains and motifs were detected using different prediction programs. The C2H2-type zinc finger was identified in the different proteins using INTERPROSCAN [🔗](#) (ZDOBNOV and APWEILER 2001). The COILS software [🔗](#) (LUPAS *et al.* 1991) was used to predict and characterize the coiled coil regions in each sequence individually with the following parameters: MTIDK matrix (weighted and unweighted), 28-residue window to determine the presence or absence of a coiled coil structure, 21-residue window to define more accurately the ends of coiled coil segments, and considering residues with probabilities >50% to be part of a coiled coil. In the *D. buzzatii* protein sequence, the nuclear localization signal (NLS) was predicted with the PSORTII software [🔗](#) (NAKAI and KANEHISA 1992) and the nuclear export signal (NES) using the NETNES 1.1 server [🔗](#) (LA COUR *et al.* 2004). Finally, the presence of PEST sequences was determined with EPESTFIND [🔗](#) (RECHSTEINER and ROGERS 1996).

For the protein conservation analysis we used the web version of the AL2CO program [🔗](#) (PEI and GRISHIN 2001) on the multiple alignment of the CG13617 proteins of 14 *Drosophila* species (excluding the Dbuz_2j sequence). All positions with gaps in more than 50% of sequences (77 positions) were excluded from the analysis. To visualize the conservation patterns of the protein, we tried different window sizes (the optimum size for motif identification according to AL2CO documentation is 3) and selected a 10 aa window as the one giving the most interpretable results, although qualitatively similar results were obtained with other sizes. The different methods used to calculate conservation give very similar results overall (as pointed out by PEI and GRISHIN 2001), and we have chosen the Sum of Pairs measure with BLOSUM62 matrix because it takes into account the similarity between different amino acids (not only identity). To calculate the amino acid frequency we used the Henikoff & Henikoff modified method, which corrects for unequal distances between the different sequences (weight of similar sequences is lower) and has been widely used. Finally, to make conservation indices equal to each other for invariant positions we used the S' normalization method.

2.8 Northern blot analysis

Northern blot hybridization was performed as described in DE *et al.* (1990) with 20 µg of total RNA from each developmental stage (embryo, larva, pupa and adult) of lines st-1 and j-19. Ethidium bromide staining of the 1.2% agarose/formaldehyde gel, which allows the visualization of ribosomal RNA bands, was used to confirm that equal amounts of RNA had been loaded onto each lane. Hybridization was carried out for 16 h at 68 °C in a hybridization buffer containing 0.45 M NaCl, 0.09 M Tris HCl, 6 mM EDTA, 0.2% BSA, 0.2% Ficoll®, 0.2% PVP, 10% dextran sulfate and 250 µg/ml yeast tRNA. Washes were performed at 68 °C for 45 min – 1 h first with a 1x SSC 0.1% SDS solution and then with a 0.3x SSC 0.1% SDS solution. The probe used in the Northern blot analysis was an antisense [³²P]UTP-labeled riboprobe synthesized by *in vitro* transcription from a clone containing the 1128-bp E2-E10 RT-PCR product amplified from st-1 *CG13617* cDNA (FIGURE 1 of PUIG *et al.* 2004).

2.9 Whole-mount *in situ* hybridization in *Drosophila* embryos

Whole-mount *in situ* hybridization was performed as described in LEHMANN and TAUTZ (1994) using embryos from lines st-1 and j-19. The same 1128-bp *CG13617* antisense riboprobe as in the Northern analysis was used, but it was non-radioactively labeled with digoxigenine-UTP during synthesis by *in vitro* transcription. Embryos 0-24 h old were collected, dechorionized with a 50% bleach solution and fixed with heptane. Hybridization was performed overnight at 65 °C using a hybridization buffer containing 50% formamide, 5x SSC, 0.1% Tween 20, 100 µg/ml yeast tRNA and 50 µg/ml heparin. Staining times for *2st* and *2j* embryos were identical. Similar hybridizations with a sense probe (that can not bind the mRNA molecule because it has the same sequence) were also performed but they did not yield any clear and reproducible results.

2.10 dsRNA detection

Following the method described by ARAVIN *et al.* (2001), in order to detect the presence of double-stranded RNA (dsRNA) in st-1 and j-19 embryos, 10 µg of RNA was treated with RNase (RNase ONE, Promega) for 3 min at 37 °C to degrade single-stranded RNA. The formation of dsRNA protects the RNA molecules from RNase digestion. Next, the RNase-treated RNA was precipitated with ethanol, treated with DNase (DNA-free™, Ambion) to eliminate any remaining contaminant DNA, denatured, and retrotranscribed using random primers to allow the synthesis of any fragment that could have hybridized with another RNA molecule. Finally, 2 µl of these cDNAs were amplified by nested PCR using in the first place the primer pair E11-E12 located in the second exon of gene *CG13617* and then the internal primer pair E33-E34 (FIGURE 1 of PUIG *et al.* 2004). It is essential that the primers used in the amplification reactions are located inside the same exon because intronic sequences have been eliminated from the mRNA molecule and, although they are present in the antisense RNA, they are not protected from RNase digestion. For each sample, a negative control without the denaturation step previous to cDNA synthesis was performed.

2.11 RNA interference

2.11.1 Synthesis of a dsRNA molecule complementary to *D. melanogaster* gene *CG13617*

In order to synthesize a specific dsRNA molecule for the *D. melanogaster* gene *CG13617*, a 1154 bp cDNA fragment comprising exons 3-5 was amplified by RT-PCR with primers DmE1-DmE2 (this gene has five exons in *D. melanogaster*, see Results) (FIGURE 11). We chose the final part of the gene to design the dsRNA because it is much less conserved among species and it does not contain any of the characterized functional domains of the protein. This is important to avoid that the small RNA fragments that will be generated from the long dsRNA molecule bind additional mRNAs, which could cause unspecific silencing of other genes. This RT-PCR product was then cloned into the pGEM-T Easy vector (Promega) and the plasmid DNA was used as a template in a PCR reaction with primers

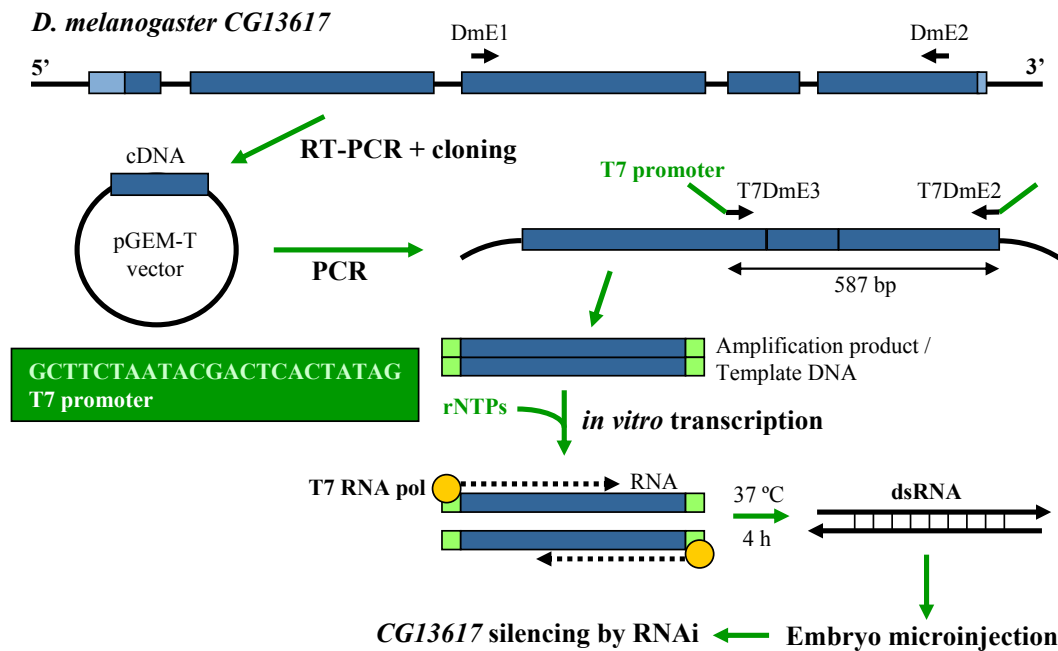


FIGURE 11 | dsRNA synthesis in *D. melanogaster*. Blue rectangles represent *CG13617* exons and green rectangles correspond to the T7 promoter sequences introduced in the PCR product. Primers are depicted as short black arrows. Yellow circles represent T7 RNA polymerase and two oppositely oriented parallel arrows symbolize the synthesized dsRNA molecule. Green arrows indicate the order of the different procedures.

T7DmE2-T7DmE3, which have a T7 promoter sequence (GCTTCTAATACGACTCACTATAG) added at their 5' end, just after the corresponding *CG13617* specific sequence (TABLE 5). We used cloned cDNA as a template in the PCR reaction (instead of cDNA directly) to be able to obtain easily large amounts of the amplification product (since the cDNA is amplified in the bacteria, the quantity of template no longer depends on the expression level of the gene). As only half of the primer sequence is able to bind the template DNA at the beginning of the amplification reaction, the annealing temperature was 57 °C during the first 10 cycles and 60 °C during the remaining 25 cycles. A PCR product containing 587 bp of *CG13617* cDNA and a T7 promoter sequence at each end was generated. This PCR product was purified from an agarose gel using the QIAquick® Gel Extraction Kit (Qiagen), cleaned with Phase Lock Gel™ Light 1.5 ml tubes (Eppendorf), precipitated with ammonium acetate and ethanol, and resuspended in RNase-free water to be used as template in an *in vitro* transcription reaction with T7 RNA polymerase (MEGAscript®

T7 Kit, Ambion). In this case, transcription starts at both ends of the DNA template because they both possess a promoter recognized by T7 RNA polymerase. The two complementary RNA molecules bind each other in the same synthesis reaction at 37 °C to form dsRNA. After the *in vitro* transcription reaction, we used RNeasy[®] Mini Kit (Qiagen) to remove proteins, ribonucleotides and template DNA and the newly synthesized dsRNA was dissolved in RNase-free injection buffer (0.1 mM sodium phosphate pH 7.8, 5 mM KCl). Finally, the dsRNA molecule was run in an agarose gel where it is expected to migrate in a way similar to dsDNA. The general guidelines for this procedure were obtained from CARTHEW (2003) and are illustrated in FIGURE 11.

2.11.2 Microinjection

D. melanogaster embryos for microinjection were collected when they were less than 1 h old in plates containing 1.5% agar dissolved in apple juice. It is essential to collect the embryos at a very early stage of development to ensure that they are in the syncytial blastoderm stage (a phase in which incomplete cell division causes the embryo to have many nuclei contained within a common cytoplasm), so the whole embryo acts as a single cell and the injected material can reach all the future cells of the individual. Embryos were dechorionized manually using double-sided adhesive tape adhered on a slide, were desiccated for 10 min at room temperature (to allow the introduction of the volume we want to inject), and were covered with halocarbon oil to avoid any posterior desiccation. Dechorionized embryos were then injected with the dsRNA complementary to the gene *CG13617* at a concentration of 430 ng/μl, which was experimentally determined to be a dsRNA concentration that produced an effective gene silencing without having an extremely toxic effect for the embryos. Control embryos were also microinjected using a solution only with buffer to avoid introducing differences between samples caused by the injection process (rather than by the effects of the dsRNA itself), which could result in changes when expression profiles of the embryos are compared. After microinjection, embryos were kept in a humid chamber at room temperature covered in halocarbon oil until collection. It is important to take into account that a variable fraction of the microinjected embryos died during the process due to any of the manipulations they endured (dechorionation, desiccation or microinjection).

Samples of microinjected individuals were initially collected at three different developmental stages to check for the effectiveness of *CG13617* silencing: ~20 h old embryos (including individuals that are already dead as well as alive embryos because it is impossible to distinguish them at this stage), first instar larvae that have just hatched, and third instar larvae that had been transferred to a vial with medium after hatching to continue their development. For embryos and first instar larvae, all the embryos in the same slide (20-30 individuals) were collected as one sample. For third instar larvae, 2 individuals were enough to be able to extract sufficient RNA. For further analysis, we chose to use only the first instar larvae samples because we can ensure that we are working with RNA from tissues of individuals that had survived the microinjection process and that were alive at the moment of collection. Also, these larvae are closest to the embryonic stage, where *CG13617* is known to be silenced in *D. buzzatii*. Since first instar larvae samples were intended to be used for different gene expression experiments (microarrays, real-time RT-PCR) a larger amount of RNA was needed and 50-100 larvae were pooled together and collected as a single sample. After hatching, first instar larvae were collected following a protocol described in CARTHEW (2003) that includes a wash with heptane to eliminate halocarbon oil. First instar larvae were stored at -80 °C until enough quantity was obtained to proceed with RNA isolation.

BOX 2 | RNA interference

Double-stranded RNA (dsRNA) has been revealed in recent years to be an important regulator of gene expression in many eukaryotes. It triggers different types of gene silencing that are collectively referred to as RNA silencing or RNA interference. The main feature that characterizes this mechanism is the presence of small ~20-30 nt non-coding RNAs able to regulate gene expression at different levels. As a general rule, these small RNAs serve as specificity factors that direct bound effector proteins to target nucleic acid molecules via base-pairing interactions. The result of this process is an inhibitory effect on the gene expression of the target gene. The discovery of these mechanisms has led to the development of experimental techniques to knock out specific genes based on the introduction of dsRNAs to silence their expression. These procedures are known as RNA interference (RNAi) techniques and are widely used in multiple model organisms such as *Caenorhabditis elegans*, *Drosophila* or even human cells.

There are three main categories of short non-coding RNAs: short interfering RNAs (siRNAs), microRNAs (miRNAs) and piwi-interacting RNAs (piRNAs). siRNAs and miRNAs come from double-stranded RNA precursors and are broadly distributed both phylogenetically and within the tissues of an organism. Instead, piRNAs are found primarily in animals, they only function in the germ line, and seem to be derived from single-stranded precursors. It is also important to take into account that these two groups of molecules bind to distinct sets of effector proteins. There are also some differences between siRNAs and miRNAs: miRNAs are processed from stem-loop precursors with incomplete double-stranded character that are purposefully expressed. They derive from a type of regulatory genes known as microRNA genes that produce non-coding transcripts with an imperfectly-paired hairpin structure. On the other hand, siRNAs are primarily exogenous in origin (they come from TEs, viruses, or transgenes) although they can be

BOX 2 | RNA interference (continued)

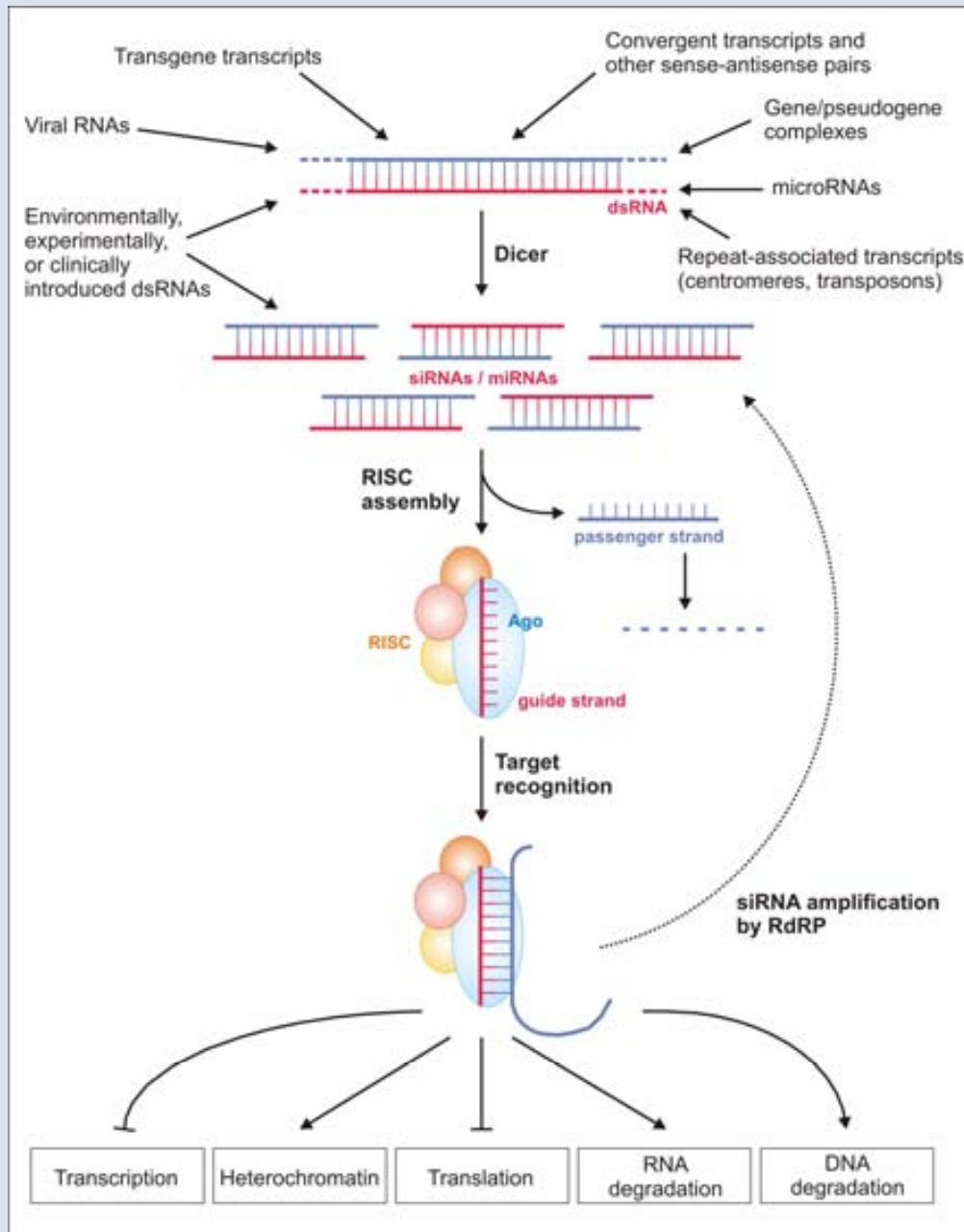


FIGURE 12 | RNA interference mechanism. Several different categories of transcripts can adopt dsRNA structures that can be processed by Dicer into short (~21-23 nt) siRNAs (or miRNAs, if microRNA genes are the origin of the hairpin dsRNAs). To the left side are the exogenous sources of dsRNA molecules and to the right the endogenous ones. These RNA duplexes can be intra or intermolecular and although most are perfectly base-paired, some are not (for example, those coming from gene/pseudogene complexes or microRNA hairpin precursors). A siRNA or miRNA consists of a guide strand (red) which assembles into a functional RISC, and a passenger strand (blue), which is ejected and degraded. All forms of RISC contain the small RNA bound to an Ago protein and some additional factors. Target RNAs are then recognized by base-pairing, and silencing occurs through one of several mechanisms. In many species, the siRNA populations that engage a target can be amplified by the action of RNA-dependent RNA polymerase (RdRP) enzymes, strengthening and perpetuating the silencing response. Adapted from CARTHEW and SONTHEIMER (2009).

BOX 2 | RNA interference (continued)

endogenous too (see below) and they are excised from long, fully complementary dsRNAs. Here we will focus our attention on siRNAs and how they regulate gene expression in *Drosophila*.

The signature components of the RNA silencing machinery are: the Dicer enzymes, the Ago proteins and the ~21-23 nt siRNAs. The trigger is the presence of dsRNA molecules that can arise from multiple sources, which can be exogenous or endogenous. The exogenous dsRNA sources include transgene transcripts (transgenes can insert in the genome forming arrays made up of multiple copies with different orientations and their transcription can result in dsRNA formation), viral RNAs and experimentally introduced dsRNAs. The endogenous dsRNA molecules can be originated from convergent transcripts or natural sense-antisense pairs, the pairing of gene/pseudogene transcripts, microRNA genes and other hairpin RNA structures, or from repeated sequences that can be transcribed, like TEs or centromeres (FIGURE 12, top).

All these dsRNA precursors need to be processed by the Dicer enzymes to form the small ~21-23 nt siRNAs. These enzymes have a PAZ domain that binds the end of the dsRNA molecule and two RNase III domains that cleave one strand each. The resulting product is a short dsRNA molecule ~21-23 nt long with ~2 nt 3' overhangs. Many different siRNAs can be excised from a single long dsRNA molecule. Some organisms have only one Dicer enzyme, but others are capable of producing several of them. In *D. melanogaster*, there are two distinct Dicers with functional specialization: Dicer-1 is required for miRNA biogenesis, while Dicer-2 is devoted mostly to the siRNA pathway. Dicer enzymes are usually associated to another protein with dsRNA-binding domains. In the case of *Drosophila* Dicer-2 this protein is R2D2.

The next step consists in the incorporation of these small RNAs into an RNA-induced silencing complex (RISC). The Argonaute (Ago) proteins are the central defining components of the various forms of RISC. They also have a PAZ domain and possess a PIWI domain exclusive of this protein family. In *Drosophila* there are five Argonaute proteins with functional specialization, but numbers vary in different organisms. However, double-stranded siRNAs generated by Dicer cannot load directly into Argonaute proteins. These siRNAs enter into a RISC assembly pathway that involves duplex unwinding and culminates in the stable association of only one of the two strands with the Ago effector protein. This Ago-

associated strand will become the guide strand that directs target recognition by base pairing, while the other passenger strand is discarded. In *Drosophila*, in the first place, the R2D2/Dicer-2 heterodimer binds a siRNA duplex and then unknown factors are added to form the RISC-loading complex (RLC). This complex assembles with Ago2 to form pre-RISC and finally, Ago2 cleaves the passenger strand, that is ejected, and a functional RISC complex is formed. Several dsRNA-binding proteins are involved in this process. Strand selection is dictated by the relative thermodynamic stabilities of the two duplex ends: whichever strand has its 5' terminus at the less stably base-paired end will be favored as the guide strand. siRNAs with equal base-pairing stabilities at their ends will incorporate either strand into RISC with approximately equal frequency.

The main mechanism of action of siRNAs is to cause gene silencing at the post-transcriptional level through the degradation of target mRNAs. The identities of the genes to be silenced are specified by this small RNA component of RISC. The siRNA guide strand directs RISC to perfectly complementary mRNA targets, which are then degraded by the PIWI domain of the Ago protein that cleaves the linkage between target nucleotides paired to siRNA nucleotides 10 and 11 (counting from the 5' end). Then cellular exonucleases attack the fragments to complete the degradative process. The mRNA target dissociates from RISC after cleavage, leaving the protein complex free to cleave additional target molecules. Mismatches at or near the center of the siRNA/target duplex suppress the endonucleolytic cleavage, but the gene can still be silenced at post-transcriptional level by other mechanisms, such as translational repression, a pathway generally used by miRNAs. Silencing of imperfectly matched mRNAs in a way similar to how miRNAs act appears to account for most "off-target" effects of siRNAs and is therefore of considerable importance.

In some organisms like *C. elegans*, primary siRNAs can induce the synthesis of secondary siRNAs through the action of an RNA-dependent RNA polymerase (RdRP) that uses siRNAs as primers to synthesize dsRNA with the target transcript as a template. This secondary siRNAs can amplify and sustain the silencing response, making it very strong. This amplification process involves the appearance of siRNAs corresponding to regions not included in the initial dsRNA trigger, so the lack of RdRP in insects and vertebrates makes the silencing mechanisms more specific in these species.


BOX 2 | RNA interference (continued)

As illustrated in FIGURE 12 (bottom), there are multiple mechanisms for siRNAs or miRNAs to cause the silencing of the target genes. Another important pathway known to take place in many species is the induction of heterochromatin formation with the consequent silencing of the affected genes. This transcriptional silencing was first reported in *Saccharomyces pombe* and also in plants as transcriptional gene silencing (TGS). It has been better characterized in *S. pombe*, where a RITS (RNA-induced transcriptional silencing) complex containing Ago1 is bound to specific *loci* such as centromeric repeats by double-stranded siRNAs where it recognizes nascent transcripts thanks to an interaction between RITS and RNA pol II. RITS association promotes histone H3 methylation on lysine 9 by histone methyltransferases, which leads to the recruitment of Swi6 protein and chromatin condensation. Engagement of RITS to nascent

transcripts also activates RdRP, which generates secondary siRNAs able to spread the silencing. Heterochromatinization causes DNA to be inaccessible for the transcription machinery and therefore, any genes included in the heterochromatin will be silenced to some extent.

In summary, in just a few years it has become clear that RNA interference pathways provide not only a completely new and unexpected mechanism to regulate gene expression and to defend the genome from invasive nucleic acids, but also have proven to be a very useful tool for biological research. These silencing mechanisms acting at some of the most important levels of genome function constitute a very active area of research and as more details of how this pathways work are discovered, new applications, such as their clinical use, will also be developed.

2.12 Microarrays

Gene expression levels of first instar larvae samples microinjected with the dsRNA that silences gene *CG13617* expression (DSRNA) and control samples injected only with buffer (CONTROL) were compared using *D. melanogaster* oligonucleotide microarrays (GeneChip® Drosophila Genome 2.0 Array, Affymetrix ) . These microarrays contain 18880 probe sets (14 probes 25-nt long each) able to interrogate ~18500 transcripts. Labeling and hybridization were performed following the instructions of the manufacturer, starting from 3 µg of total RNA of four different samples: DSRNA1, DSRNA2, CONTROL1, CONTROL2 (FIGURE 13). Each sample was processed independently and hybridized to a different array. For the four arrays quality measures fell within the usual limits and were similar between them.

To take into account the diversity of methods available to process the array information and calculate expression values, array results were analyzed using three different programs: GENECHIP® OPERATING SOFTWARE (GCOS) version 1.4 (Affymetrix), RMA (IRIZARRY *et al.* 2003) as implemented in Bioconductor (GENTLEMAN *et al.* 2004), and DCHIP version 2004 (LI and WONG 2001). In the GCOS analysis, first all arrays were normalized

separately to a same average intensity of 500 and pair-wise comparisons between the arrays of the control and embryos injected with dsRNA were generated. Then, probe sets showing expression differences between them were identified using the BULLFROG 5.3 program (ZAPALA *et al.* 2002) with the following criteria: a consistent call of increase/marginal increase or decrease/marginal decrease in the four pair-wise comparisons and an average fold change greater than 1.8. For the Bioconductor analysis, arrays were normalized by quantile normalization and expression values were calculated using the RMA method. The resulting expression values were then analyzed with the SAM (Significance Analysis of Microarrays) program (TUSHER *et al.* 2001) as two class unpaired data using default parameters. The list of differentially expressed genes was obtained by fixing a false discovery rate (FDR) of 10% ($\delta = 0.5843$) and all those with an average fold change between DSRNA and CONTROL arrays greater than 1.8 were selected. In the DCHIP analysis, arrays were normalized to that with median intensity and expression values for each probe set were calculated as a model-based expression index using the PM/MM difference model and the program default parameters. The criteria used to identify probe sets with signal differences between experimental conditions were a fold-change greater than 1.5 using the lower bound of the 90% confidence interval, absolute difference between means greater than 50, and a *t*-test *P*-value lower than 0.05. The lists obtained with each method were combined and those probe sets differentially expressed in at least two of the three independent analyses were considered to be significant. Cluster analysis was carried out by average linkage hierarchical clustering using the CLUSTER and TREEVIEW programs (EISEN *et al.* 1998). Prior to the clustering, the hybridization signal from each probe set was median-centered and normalized across the samples, and the uncentered Pearson correlation was used as similarity metric. The gene ontology analysis of the differentially expressed genes was performed using the Functional Annotation Clustering tool at the DAVID Bioinformatics Resources NIAID/NIH webpage [🔗](#) (DENNIS *et al.* 2003, HUANG *et al.* 2009) using the 41 differentially expressed probe sets and medium classification stringency (see Results).

Eukaryotic Target Labeling for GeneChip® Probe Arrays

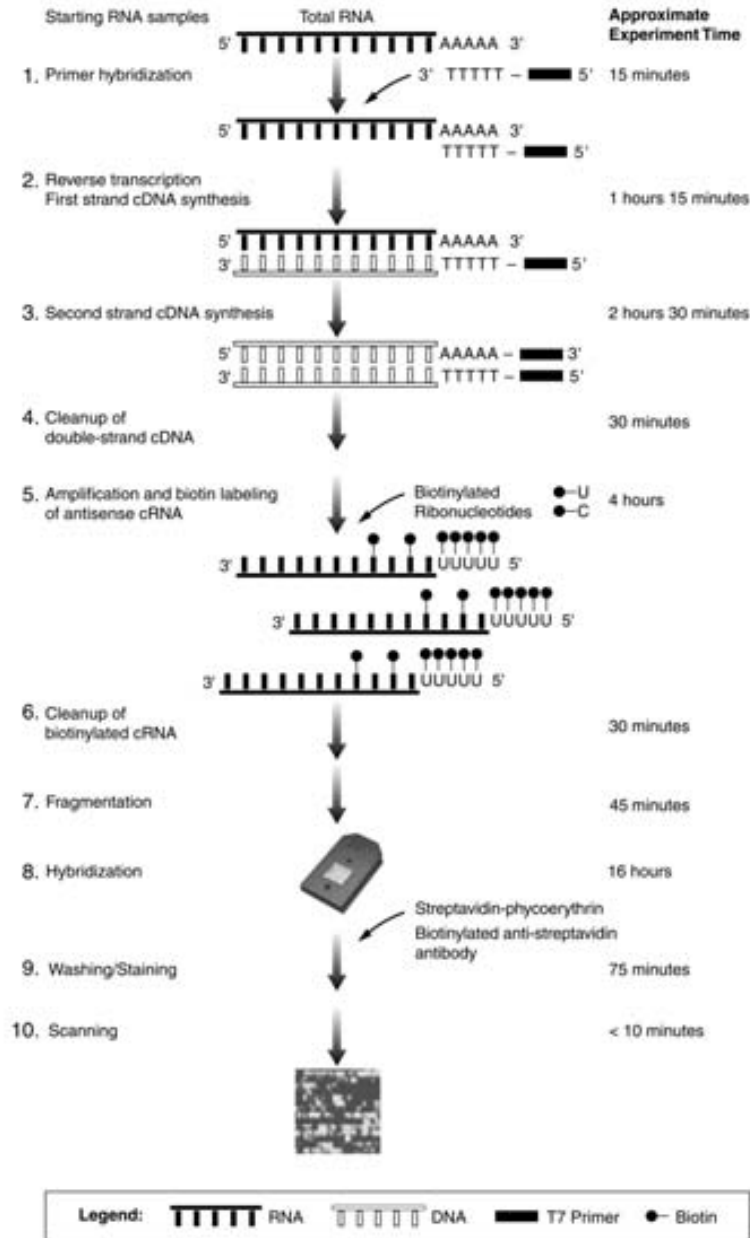


FIGURE 13 | Labeling of RNA samples to be hybridized in Affymetrix microarrays for gene expression profiling. The different procedures are indicated at the left of the figure. Starting from total RNA samples, double stranded cDNA is synthesized adding a T7 promoter at the 3' end. This double stranded cDNA is then used as a template in an *in vitro* transcription reaction with biotin-labeled ribonucleotides. In this reaction, biotin-labeled antisense complementary RNAs (cRNA) are generated. These cRNA molecules are the complementary strand to the initial cellular RNAs. After fragmentation, cRNAs are able to hybridize with the oligonucleotide probes contained in the microarray (which possess the original mRNA sequence). Figure from <http://www.affymetrix.com>.

BOX 3 | Oligonucleotide microarrays

DNA oligonucleotide microarrays are a technology used in molecular biology to assess the expression levels of thousands of genes simultaneously in a given tissue. This microarray-based gene expression profiling can be used to identify genes whose expression is different in two or more samples. Affymetrix GeneChip® arrays consist of a glass surface containing hundreds of thousands of oligonucleotide probes packed at extremely high densities (FIGURE 14). These probes are synthesized *in situ* using photolithography and combinatorial chemistry techniques. Each oligonucleotide is located in a specific area on the array called a probe cell and each probe cell contains hundreds of thousands to millions of copies of a given oligonucleotide. These probes are 25 nucleotides long and they are complementary to the thousands of annotated transcripts in a genome (FIGURE 15a). Probes are designed to maximize sensitivity, specificity, and reproducibility, allowing consistent discrimination between specific and background signals, and between closely related target sequences. Probes are arranged in probe pairs formed by a perfect match and a mismatch probe where the central nucleotide has been substituted by a different one (FIGURE 16a). This single nucleotide change affects hybridization of the target labeled mRNA and provides a measure for background signals against which the true signal obtained with the perfect match probe can be compared. Several probe pairs are used to measure the level of transcription of each transcript forming what is called a probe set (FIGURE 16b). In

order to perform the experiment, biotin-labeled RNA fragments of the sample we want to analyze are hybridized to the array (FIGURE 13). The hybridized microarray is then stained with streptavidin phycoerythrin conjugate and scanned. The amount of light emitted at 570 nm is proportional to the bound target at each location on the probe array (FIGURE 15b).



FIGURE 14 | Affymetrix GeneChip® microarray. Probes are located in the central glass surface enclosed within a sealed cartridge.

The GeneChip® *Drosophila* Genome 2.0 Array (Affymetrix) is a microarray tool for studying expression of *D. melanogaster* transcripts. It comprises 18880 probe sets analyzing over 18500 transcripts, which provides a considerable coverage of the transcribed part of this species' genome. The microarray contains 14 probe pairs per probe set. The probes in this microarray were designed based on the FlyBase release 3.1 annotation of the *D. melanogaster* genome, but other published gene predictions from the *Drosophila* Research community were also included on the array.

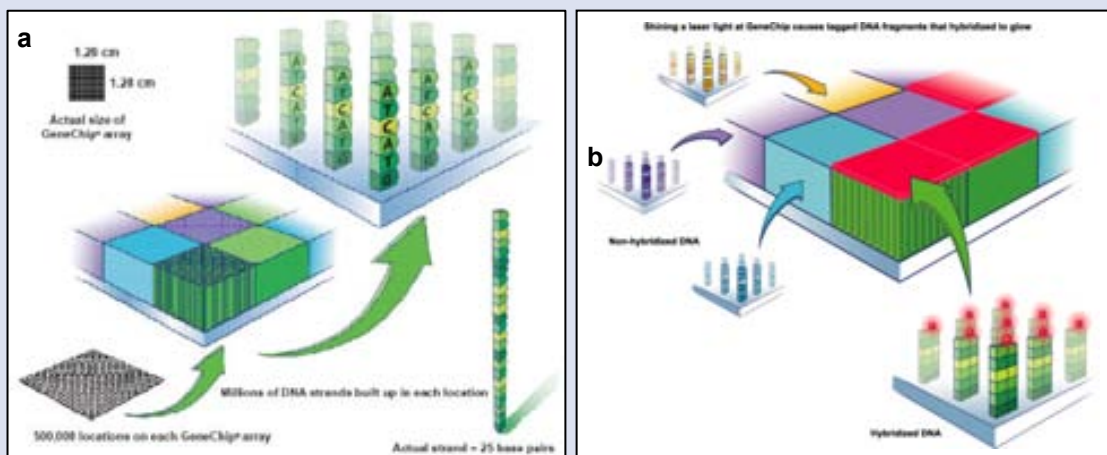


FIGURE 15 | Structure and spatial organization of Affymetrix microarrays (a) and hybridization of the labeled mRNAs to their corresponding complementary probes (b). If a larger quantity of a determined transcript is available, more copies of the probe will bind an mRNA molecule and a stronger signal will be detected when the microarray is scanned. The intensity of the measured signal will be always proportional to the initial amounts of each transcript in the analyzed sample. Figures from <http://www.affymetrix.com>.

BOX 3 | Oligonucleotide microarrays (continued)

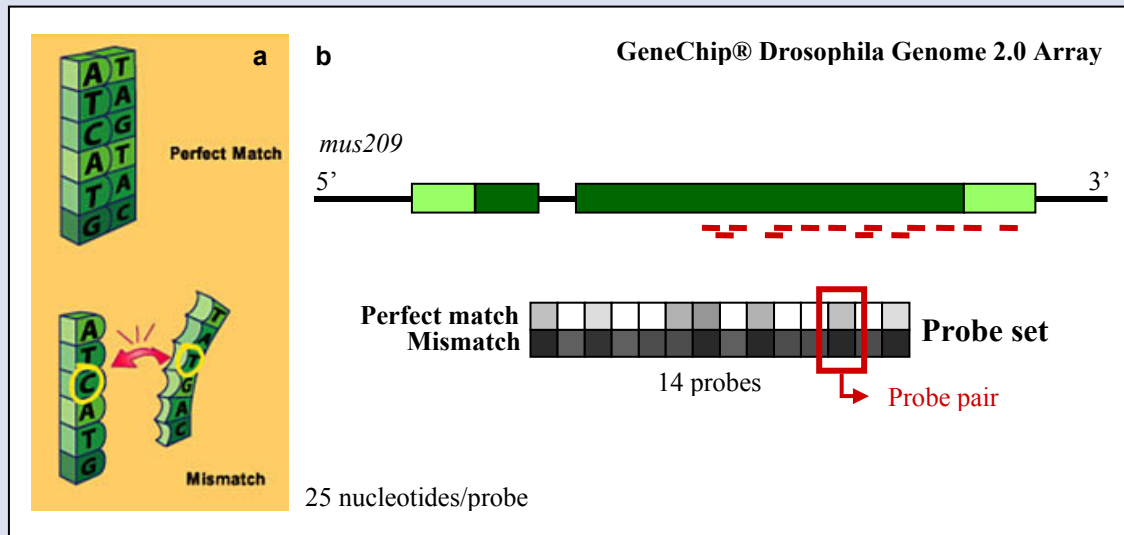


FIGURE 16 | Each transcript is analyzed by a specific probe set formed by 14 probe pairs. (a) A probe pair is formed by a perfect match probe and a mismatch probe, where the central nucleotide of the 25-nt long oligonucleotide probe is substituted by a different one. This mismatch does not allow the hybridization of the labeled mRNAs. **(b)** A schematic representation of *D. melanogaster mus209* gene structure (dark green rectangles correspond to the coding region and light green are the UTRs) shows the location of the 14 probes (red bars) contained in the microarray to interrogate the expression level of this gene. In each of the 14 probe pairs, the perfect match probe gives strong expression signals (clear squares) in the hybridized, stained and scanned microarray, while the mismatch probe (where hybridization of the target mRNA does not occur) provides a background measure and serves as a control for non-specific hybridization. Panel (a) from <http://www.affymetrix.com>.

RESULTS

Arrayed in their regiments, my genes carry out their orders. All except two, a pair of miscreants – or revolutionaries, depending on your view – hiding out on chromosome number 5.

– JEFFREY EUGENIDES, *Middlesex* (2002)

3.1 Position effect of inversion *2j* on *CG13617* gene expression in *D. buzzatii*

The first part of the results consists of an article published in *Proceedings of the National Academy of Sciences of the USA* in 2004. In this article we describe an expression difference in gene *CG13617*, adjacent to the proximal breakpoint of the polymorphic inversion *2j* of *D. buzzatii*, between inverted and non-inverted chromosomes. Also, we propose an unusual silencing mechanism seemingly triggered by the TEs inserted at the breakpoint junction in inverted chromosomes as the cause of this difference, which constitutes one of the few examples of position effects of inversion breakpoints on the expression of nearby genes found in nature.

MARTA PUIG, MARIO CÁCERES and ALFREDO RUIZ (2004) Silencing of a gene adjacent to the breakpoint of a widespread *Drosophila* inversion by a transposon-induced antisense RNA. *Proc. Natl. Acad. Sci. USA* **101**: 9013-9018.

Silencing of a gene adjacent to the breakpoint of a widespread *Drosophila* inversion by a transposon-induced antisense RNA

Marta Puig*, Mario Cáceres[†], and Alfredo Ruiz**

*Departament de Genètica i de Microbiologia, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain; and [†]Department of Human Genetics, Emory University School of Medicine, 615 Michael Street, Atlanta, GA 30322

Communicated by Margaret G. Kidwell, University of Arizona, Tucson, AZ, May 4, 2004 (received for review February 5, 2004)

Adaptive changes in nature occur by a variety of mechanisms, and *Drosophila* chromosomal inversions was one of the first studied examples. However, the precise genetic causes of the adaptive value of inversions remain uncertain. Here we investigate the impact of the widespread inversion *2j* of *Drosophila buzzatii* on the expression of the *CG13617* gene, whose coding region is located only 12 bp away from the inversion proximal breakpoint. This gene is transcribed into a 2.3-kb mRNA present in all *D. buzzatii* developmental stages. More importantly, the expression level of *CG13617* is reduced 5-fold in embryos of lines homozygous for the *2j* inversion compared with lines without the inversion. An antisense RNA that originates in the *Foldback*-like transposon *Kepler* inserted at the breakpoint junction in all of the *2j* lines and that forms duplexes with the *CG13617* mRNA in *2j* embryos is most likely responsible for the near silencing of the gene. Few examples of RNA interference caused by transposable elements (TEs) have been previously described, but this mechanism might be prevalent in many organisms and illustrates the potential of TEs as a major source of genetic variation. In addition, because chromosomal rearrangements are usually induced by TEs, position effects might be more common than previously recognized and contribute significantly to the evolutionary success of inversions.

In the genus *Drosophila*, many species are polymorphic for inversions in one or more chromosomes, providing the most extraordinary and best studied example of chromosomal variation in nature (1). Compelling evidence for the adaptive significance of inversion polymorphisms has been drawn from seasonal and long-term frequency changes as well as latitudinal and altitudinal clines (2, 3). However, the molecular mechanisms underlying inversion maintenance in natural populations are still unclear. According to the coadaptation hypothesis, the reduction of recombination in the inverted chromosomal segment of heterokaryotypes keeps together favorable allele combinations (4, 5). Alternatively, the position effect hypothesis proposes that the localization of the inversion breakpoints near or inside genes could affect their function or expression profile (6). Although the available evidence favors coadaptation as the most plausible explanation for the adaptive value of chromosomal inversions (2, 7, 8), little is known of the mutational outcome of breakpoints in natural inversions and their consequences on the expression patterns of nearby genes (9).

A variety of position effects have been observed in spontaneous mutations generated by inversions in diverse organisms. Inversions can disrupt the coding region of a gene, causing the loss of its function, as in the *Dpp6* gene in mice (10) or the hemophilia A factor VIII gene in humans (11). The inversion of a chromosomal segment can also remove or exchange the regulatory sequences of a gene and alter its expression pattern, as in the *Antp*^{73b} mutation of *Drosophila melanogaster* (12) or the *nivea* gene of *Antirrhinum* (13). Finally, inversions can move genes to distant sites where their expression is silenced by the proximity of centromeric heterochromatin (14). Another factor that could contribute to the position effects of inversions is the presence of transposable elements (TEs)

at their breakpoints. TEs constitute a large fraction of most eukaryotic genomes and have been implicated in the generation of inversions in *Drosophila* and other organisms (15, 16). In addition, TEs have been shown to have diverse effects on gene expression (17, 18), as a result not only of the modification of functional regions (19–21) but also of interference with the transcription of adjacent genes by read-through transcription, antisense transcripts, or insulation (22).

The analysis of position effects of natural inversions has been hindered by the lack of molecular studies of their breakpoints and flanking genes. In this work we have studied the case of inversion *2j* of *Drosophila buzzatii*, which was originated by a unique event of ectopic recombination between two oppositely oriented copies of the *Foldback*-like element *Galileo* and which contains at its breakpoints blocks of different TEs absent from noninverted chromosomes (15, 23). In *D. buzzatii* chromosome 2, the ancestral *2 standard* (*2st*) arrangement and the derived *2j* arrangement are found in most natural populations at high frequencies (24). Both arrangements seem to be maintained as a balanced polymorphism by a fitness tradeoff in which the carriers of the *2j* and *2st* arrangements are characterized by a larger adult body size and a shorter developmental time, respectively (25), although the causes of these differences are not known. Two genes, *Pp1α-96A* and *nAcRβ-96A* (15), were originally found flanking the *2j* proximal breakpoint in the *2st* arrangement. However, the sequencing of the *D. melanogaster* genome (26) revealed a putative ORF named *CG13617* located outside the inversion and just a few nucleotides away from the breakpoint (Fig. 1). This observation prompted us to ascertain whether the generation of the inversion caused any change in *CG13617* expression. The results show that this gene is nearly silenced in embryos of lines homozygous for inversion *2j* and that this silencing is not caused by the inversion itself but by the transcription of an antisense RNA from the transposon *Kepler* inserted at the proximal breakpoint junction.

Materials and Methods

Flies. Eight isogenic *D. buzzatii* lines homozygous for chromosomal arrangements *2st* ($n = 4$), *2j* ($n = 3$), or *2jz*³ ($n = 1$) were used. Arrangement *2jz*³ is derived from *2j* by an additional inversion, *2z*³ (27). These lines represent the natural variability within the species' geographical range (Table 1), and the *2j* lines differ in the size and TE content of the insertions at the *2j* proximal breakpoint (23).

DNA Sequencing and Sequence Analysis. During the cloning of inversion *2j* breakpoints (15), the *CG13617* gene was partially sequenced in the *D. buzzatii* lines *st-1* and *j-1*. The *CG13617* sequence was completed in both lines by subcloning the *λst9* phage

Abbreviation: TE, transposable element.

Data deposition: The sequences have been deposited in the GenBank database (accession nos. AY551073–AY551076).

[†]To whom correspondence should be addressed. E-mail: alfredo.ruiz@uab.es.

© 2004 by The National Academy of Sciences of the USA

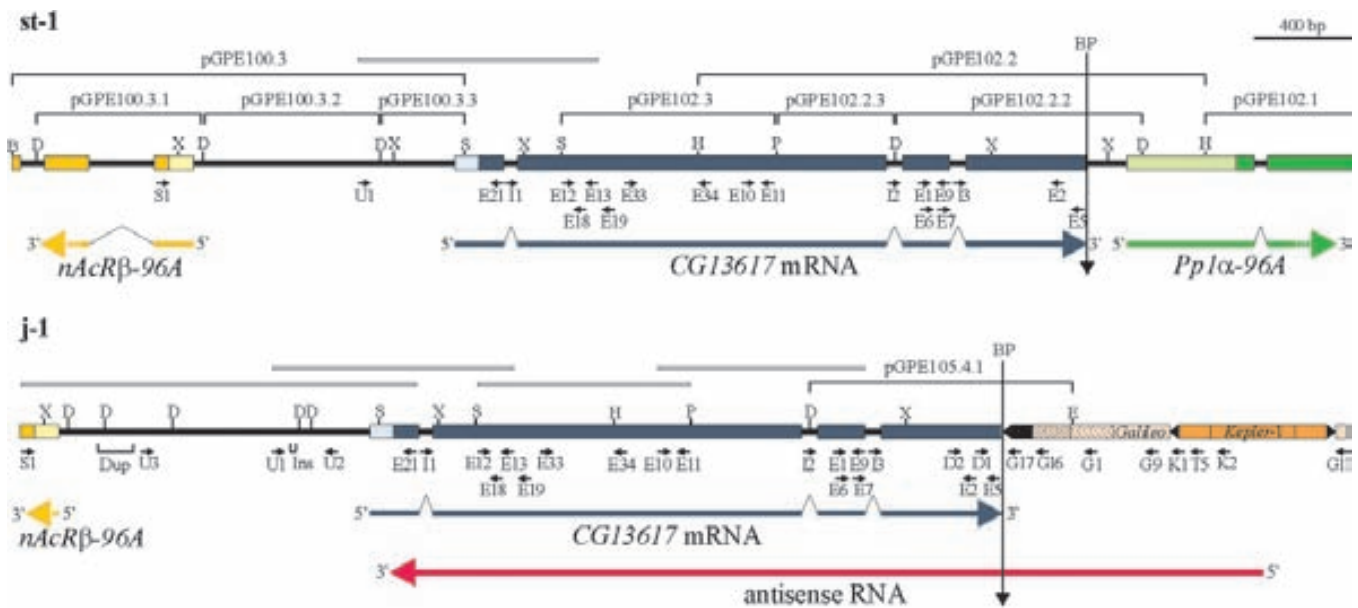


Fig. 1. Schematic representation of inversion *2j* proximal breakpoint sequence in lines *st-1* and *j-1*. Vertical arrows indicate the inversion breakpoint (BP) and separate the inverted and the noninverted segments (to the right and left of the vertical arrows, respectively). Exons of genes flanking the breakpoint are represented as colored boxes, with coding sequences in dark and UTRs in light colors. Transcripts of each gene are shown below the diagram, and 5' and 3' denote their orientation. Small black arrows indicate PCR primers used in this study. Clones and PCR fragments used for sequencing are represented above each diagram by thin black lines and open bars, respectively. TEs inserted at the breakpoint in line *j-1* are depicted as colored boxes with sharp ends. The small duplication (Dup) and insertion (Ins) found at the *CG13617* upstream region in line *j-1* are also indicated. Primer G11 is absent from line *j-1* as the result of an internal rearrangement, but it is represented here for simplicity. Restriction sites: B, *Bam*HI; D, *Dra*I; X, *Xmn*I; S, *Sal*I; H, *Hind*III; P, *Pst*I; E, *Eco*RI.

into the pBluescript II SK vector (Stratagene) and by PCR amplification (Fig. 1). PCR and RT-PCR products were gel-purified and sequenced directly with the same primers. GenBank similarity searches were performed by using BLASTX to identify *CG13617* coding regions and BLASTP to find homologous proteins in other species. Sequences were aligned with CLUSTALW.

RNA Extraction and Northern Analysis. Total RNA was isolated from embryos, larvae, pupae, and adults by using TRIzol (Invitrogen). Northern blot hybridization was performed as described (28), with 20 μ g of total RNA from each developmental stage of lines *st-1* and *j-19*. Loading of the gel was verified by ethidium bromide staining. An antisense [³²P]UTP-labeled riboprobe was synthesized by *in vitro* transcription from a clone containing the 1,128-bp E2/E10 RT-PCR product from *st-1 CG13617* cDNA. Hybridization was carried out for 16 h at 68°C.

RT-PCR and PCR Amplification. Total RNA was treated with 1 unit of DNase I (Ambion, Austin, TX) for 30 min at 37°C to eliminate

DNA contamination, and cDNA was synthesized from 500 ng of the DNase I-treated RNA by using an oligo(dT) primer (First Strand cDNA Synthesis kit for RT-PCR, Roche Diagnostics). Negative reactions without retrotranscriptase were carried out to control for DNA contamination. PCRs were performed in a total volume of 25 μ l, including 1 μ l of cDNA or 100–200 ng of genomic DNA, 10 pmol of each primer, 200 μ M dNTPs, 1.5 mM MgCl₂, and 1.5 units of *Taq* DNA polymerase. To differentiate the size of amplification products from cDNA and genomic DNA, primer pairs used in RT-PCRs were selected to span an intron of the gene. Primer sequences are available in Table 3, which is published as supporting information on the PNAS web site.

RACE. RACE experiments were done with DNase I-treated RNA from embryos of lines *st-1* or *j-1*; 5' RACE was carried out by using the 5'/3' RACE kit (Roche Diagnostics), and 3' RACE was carried out by using cDNA synthesized with primer poly(T). The gene-specific primers used in each case are listed in Table 3. All of the amplification products spanned one intron of the gene to ensure

Table 1. *D. buzzatii* isogenic lines used in this study

Line	Chromosomal arrangement	Geographic origin	Proximal breakpoint insertions	
			Size, bp	TE content*
<i>st-1</i>	<i>2st</i>	Carboneras, Spain	0	—
<i>st-4</i>	<i>2st</i>	Guaritas, Brazil	0	—
<i>st-7</i>	<i>2st</i>	Termas de Río Hondo, Argentina	0	—
<i>st-8</i>	<i>2st</i>	Ticucho, Argentina	0	—
<i>j-1</i>	<i>2j</i>	Carboneras, Spain	4,313	<i>Galileo</i> , <i>Kepler</i> (2), <i>BuT1</i> , <i>BuT3</i> (2)
<i>j-13</i>	<i>2j</i>	Guaritas, Brazil	3,214	<i>Galileo</i> , <i>Kepler</i> , <i>BuT1</i> , <i>BuT3</i>
<i>j-19</i>	<i>2j</i>	Ticucho, Argentina	4,724	<i>Galileo</i> , <i>Kepler</i> , <i>BuT1</i> , <i>BuT3</i> , <i>Newton</i>
<i>jz³⁻⁴</i>	<i>2jz³</i>	Tilcara, Argentina	6,341	<i>Galileo</i> (2), <i>Kepler</i> , <i>BuT1</i> , <i>BuT3</i> (2)

*The number of TE copies is indicated in parenthesis when more than one copy is present. The exact nature of the breakpoint insertions is described in detail in figure 2 of ref. 23.

Table 2. Structure of the *CG13617* coding region in *D. buzzatii* and *D. melanogaster*

Structure	<i>D. buzzatii</i>		<i>D. melanogaster</i>		Identity, %
	st-1, bp	j-1, bp	Structure	bp	
Exon 1	94	94	Exon 1	94	72.3
Intron 1	60	60	Intron 1	58	—
Exon 2	1,513	1,513	Exon 2	785	68.1
			Intron 2	58	
Intron 2	56	53	Exon 3	731	—
			Intron 3	55	
Exon 3	157	157	Exon 4	157	64.3
Intron 3	83	81	Intron 4	58	—
Exon 4	438	438	Exon 5	444	52.9

that they came from mRNA. In some cases, different amounts of the first amplification were tested as a template for the second one to obtain a specific band. RACE products were cloned into the pGEM-T vector (Promega). A minimum of eight different clones from each reaction were screened by restriction mapping, and two or three clones were finally sequenced. In the 5' RACE, clones with different inserts were selected for sequencing.

Real-Time RT-PCR. Real-time RT-PCR was performed in an ABI PRISM 7900HT Sequence Detection System with the DNA-binding dye SYBR Green (Applied Biosystems). Primers were designed by using PRIMER EXPRESS VERSION 1.5 software (Applied Biosystems) in areas conserved between *2st* and *2j* lines (Table 3). The housekeeping gene *Gapdh* was used as internal control for differences in cDNA concentration. Previously, 1,017 bp of the *D. buzzatii* *Gapdh* gene were amplified and sequenced from st-1 genomic DNA. For each sample, the gene of interest and *Gapdh* were both amplified in triplicate, and results were analyzed by using SEQUENCE DETECTOR VERSION 1.7 and DISSOCIATION CURVE VERSION 1.0 software (Applied Biosystems). Relative quantification was performed with the standard curve method, and gene amplification levels were normalized by dividing by *Gapdh* levels in each sample. Expression levels were compared by means of ANOVA.

Results

The gene *CG13617* and its flanking regions were sequenced in two *D. buzzatii* lines, with (j-1) and without (st-1) the *2j* inversion (Fig. 1). Sequence comparison with the predicted *D. melanogaster* *CG13617* gene (26) allowed us to determine the *D. buzzatii* *CG13617* coding region, which is 2,202 bp long and shares a 65% nucleotide identity with *D. melanogaster*. The *D. buzzatii* gene is split into four exons and lacks the second intron of *D. melanogaster*, but the other three are very similar in size and location (Table 2). The *CG13617* sequence presents an overall nucleotide identity of 97.8% between st-1 and j-1. The most remarkable difference was the large TE insertions of the *2j* proximal breakpoint located only 12 bp from the stop codon of the gene in line j-1 (Fig. 1). Furthermore, sequence comparison revealed two other small structural changes in line j-1: a 166-bp tandem duplication and a 26-bp insertion located 1,010 and 408 bp upstream of the start codon, respectively (Fig. 1). To further characterize the structural variation between *2st* and *2j* lines, the entire region comprised by the *CG13617* gene and upstream sequences was analyzed in three additional lines with each arrangement (Table 1) by PCR amplification and restriction mapping (Table 4, which is published as supporting information on the PNAS web site). Only a few restriction site polymorphisms were detected. In addition, line j-19 showed the same duplication and small insertion as j-1, and jz⁻³-4 presented an ≈900-bp insertion (probably an *ISBu-1* element) in the middle of the intergenic region. Therefore, *2st* and *2j* sequences are very similar, and the TE

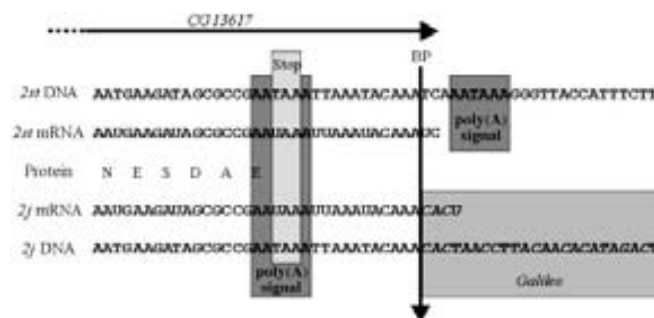


Fig. 2. Nucleotide sequence of the 3' end of *CG13617* in lines st-1 and j-1. The stop codon, the two putative polyadenylation signals, and the *Galileo* element inserted in line j-1 are included in shaded boxes. A vertical arrow indicates the *2j* inversion proximal breakpoint (BP). Sequences present only in lines with the inversion are shown in italic.

insertions at the *2j* proximal breakpoint are the only structural change that differentiates all *2j* and *2st* lines.

To validate the predicted structure of the gene, the *CG13617* mRNA was sequenced completely in line st-1 through RT-PCR amplification with primer pairs E9/E10 and E11/E12 (Fig. 1) and isolation of both ends by 5' and 3' RACE. In the 5' RACE, the longest of the sequenced clones resulted in a 118-bp 5' UTR and contained in its 5' end a G not present in the DNA sequence that likely corresponds to the cap and the transcription start site. Because of the proximity to the inversion breakpoint, the 3' RACE was carried out in both st-1 and j-1 lines, and the three clones sequenced for each line contained precisely the same sequence. In the *2st* line the *CG13617* 3' UTR was just 17 bp long and extended 2 bp after the breakpoint, whereas in the *2j* line it was 19 bp long and included the first 4 bp of the *Galileo* element inserted at the proximal breakpoint (Fig. 2). This is consistent with the use of the first of the two putative polyadenylation signals found at the end of the *D. buzzatii* *CG13617* gene during mRNA processing (Fig. 2). The total size of the *CG13617* mRNA is then 2,337 bp in line st-1 and 2,339 bp in line j-1.

CG13617 expression was analyzed in *D. buzzatii* by Northern blot hybridization and RT-PCR in embryos, larvae, pupae, and adults. Northern analysis detected an ≈2.3-kb transcript expressed at different levels in embryos, pupae, and adults, but the signal in larvae was inappreciable (Fig. 3A). However, RT-PCR amplification with primer pair E2/E7 produced the expected 387-bp band in all stages, including larvae (Fig. 3B). The expression level was similar in both arrangements in all stages except embryos, which showed lower RNA amounts in the *2j* than in the *2st* line. Semi-quantitative RT-PCR in embryos from four lines with each arrangement (Table 1) confirmed this difference (data not shown), and *CG13617* expression levels were more accurately quantified by real-time RT-PCR (Fig. 4). The real-time RT-PCR analysis did not detect differences across the lines with the same arrangement, but a significant 5-fold reduction in the average *CG13617* expression level was found in *2j*, compared with *2st*, embryos ($P = 0.010$; Table 5, which is published as supporting information on the PNAS web site). A similar real-time RT-PCR analysis was performed with *PpIα-96A*, the next closest gene to the proximal breakpoint (Fig. 1). However, no significant differences were detected in this case between *2st* and *2j* embryos ($P = 0.952$; Fig. 4 and Table 5), corroborating previous results (15). Finally, *in situ* hybridization in whole embryos was carried out to study spatial differences in *CG13617* expression between *2st* and *2j* lines. Stage 11–12 embryos showed a similar expression pattern in both lines, which consisted of several signals in the head and in each metamer, forming two lines along the embryo (Fig. 7, which is published as supporting information on the PNAS web site). However, the pattern clearly

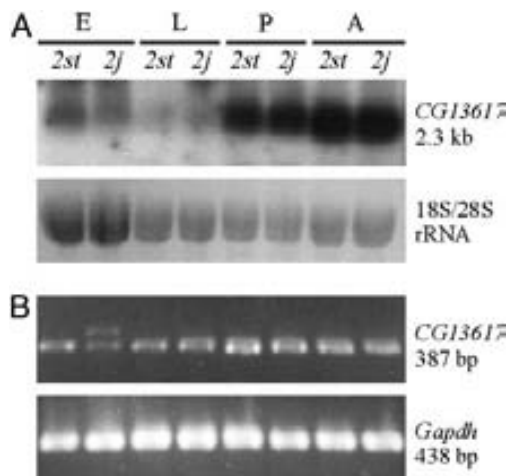


Fig. 3. Expression analysis of *CG13617* in four different developmental stages of two lines with (*2j*) and without (*2st*) the inversion. (A) Northern blot analysis of *CG13617*. (Upper) An \approx 2.3-kb transcript was detected in embryos, pupae, and adults. (Lower) rRNA used as a loading control is shown. The RNA amount was higher in embryos, but this fact is not relevant for the comparison between arrangements. (B) Semiquantitative RT-PCR analysis of *CG13617*. Primers E2/E7 and H1/H2 were used to amplify, respectively, 387 bp of *CG13617* mRNA (Upper) and 438 bp of *Gapdh* mRNA (Lower) as a reference. In *2j* embryos, there is an extra band containing intron 3 that corresponds to the *CG13617* antisense transcript. E, embryos; L, larvae; P, pupae; A, adults.

was fainter in *2j* than in *2st* embryos, possibly reflecting the expression level difference between them.

An unexpected result was that in the E2/E7 RT-PCR, besides the 387-bp band from *CG13617* mRNA, an additional band of equal size to the genomic DNA (470 bp) appeared in embryos of the four *2j* lines studied (Fig. 3B and data not shown). To confirm this result, two RT-PCR amplifications with primer pairs I2/E9 and I3/E2, in which one primer is placed in an intron sequence, were carried out by using RNA from embryos of the eight lines. Products with the expected size (211 and 398 bp, respectively) were obtained in all *2j* lines but not in the *2st* lines or any of the negative controls without retrotranscriptase, ruling out genomic DNA contamination (data not shown). Two explanations were possible for this additional transcript containing at least the last two introns of *CG13617*: an unprocessed pre-mRNA or an antisense RNA transcribed from the other DNA strand, which lacks the appropriate splicing signals. To test the second option, a strand-specific RT-PCR (29) was performed. In the sense-specific cDNAs, a single band of 387 bp, corresponding to the mRNA after the elimination of the third intron, was amplified in all lines, with clearly lower levels in *2j* than

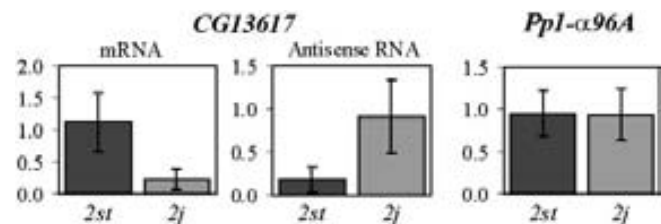


Fig. 4. Relative quantification of *CG13617* mRNA and antisense RNA and *Pp1 α -96A* mRNA by real-time RT-PCR. Average expression levels for four *2st* and four *2j* lines are represented in the graphs. Standard deviation within each arrangement is indicated by error bars. Expression levels are shown in relation to those of lines *st-1* (*CG13617* and *Pp1 α -96A* mRNA) and *j-1* (antisense RNA). A primer spanning the third and fourth exons (E25) and a primer located in the second intron (AS1) were used to ensure that, respectively, only the sense or the antisense transcripts of *CG13617* were amplified.

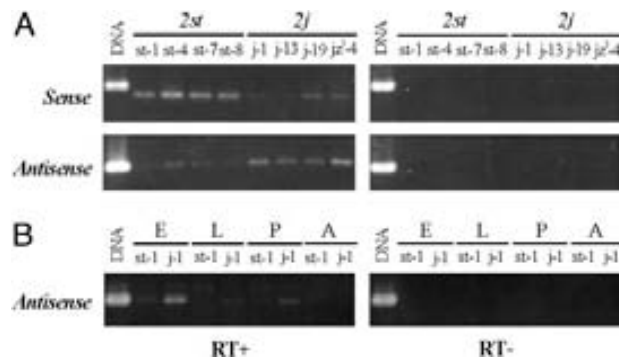


Fig. 5. Strand-specific RT-PCR of *CG13617*. Two different cDNAs were synthesized for each sample by using primers specific for the sense (E5) and antisense (E6) transcript and were then amplified with primer pair E2/E7 (RT+). Negative controls for each sample without retrotranscriptase (RT-) are also shown. (A) Amplification of sense and antisense RNAs in embryos from eight different lines. The 387-bp band in the sense mRNA is smaller than that of the DNA and the antisense transcript (470 bp) because of the splicing of the 83-bp intron. (B) Amplification of antisense RNA in four developmental stages of lines *st-1* and *j-1*. E, embryos; L, larvae; P, pupae; A, adults.

in *2st* embryos (Fig. 5A). No extra band was observed in any case. In the antisense-specific cDNAs, a 470-bp product was obtained in *2j* lines, but only a weak signal appeared in some of the *2st* lines (Fig. 5A). Thus, an antisense transcript of *CG13617* is expressed in all *2j* embryos but is not found or found only in very low amounts in *2st* embryos. This antisense RNA is apparently absent in other developmental stages, except for a barely visible band in *2j* larvae and pupae (Fig. 5B).

The level of the *CG13617* antisense RNA in embryos was quantified by real-time RT-PCR by using a primer located in the second intron. The same eight lines were analyzed, and no differences were found among those with the same arrangement. However, the average antisense RNA expression level was 5 times higher in *2j* than in *2st* lines (Fig. 4), a difference that is highly significant ($P = 0.016$; Table 5). In addition, sense and antisense RNA expression was analyzed by real-time RT-PCR in embryos heterozygous for *2j* inversion produced by the two reciprocal crosses of lines *st-1* and *j-1*. In both analyses, expression levels were again significantly different between the parental *st-1* and *j-1* lines, even though a new primer pair from the third intron region was used to quantify the antisense RNA, and no significant differences were found between the two reciprocal heterokaryotypes (Table 6, which is published as supporting information on the PNAS web site). In the sense RNA, the expression level in both heterokaryotypes was similar to that of line *j-1* and was significantly lower compared with the average of *st-1* and *j-1* homozygotes (Table 6). Conversely, the average antisense RNA level in the heterokaryotypes was not significantly different from the average of the two homozygous parental lines (Table 6), pointing to an intermediate expression of this transcript in heterozygotes.

The proximity of *CG13617* to the TEs inserted at the proximal breakpoint in the *2j* lines (Fig. 1) suggests that the antisense RNA is transcribed predominantly from these elements. In the lines without the inversion there may be a low level of antisense transcription in the absence of TEs, but the origin of this RNA is still unclear. The 5' end of the antisense RNA in the *2j* lines was characterized by RT-PCR from the strand-specific cDNAs using eight primers (G17 to G11) located in the part of the TE insertions shared by most *2j* lines, in combination with three primers located in the last exons of *CG13617* (E7, D1, and D2) (Fig. 1). RT-PCR products of the expected size were obtained with the five primers closest to the coding region (G17 to K1) in all four *2j* lines. Primers T5 and K2 yielded some amplification only in lines *j-13* and *jz³-4*. Finally, no amplification was observed in any *2j* line with the

outermost primer (G11), placing the origin of the antisense RNA inside the *Kepler* element inserted within the *Galileo* copy that generated the *2j* inversion. Repeated attempts to clone the 5' end of the transcript by 5' RACE failed, perhaps because of the complex secondary structures formed by these *Foldback*-like TEs. RT-PCR with j-13 and j-19 oligo(dT)-synthesized cDNA was used to determine the approximate extension of the antisense RNA. The expected amplification product was obtained with primers I1, located in the first intron, and E11, but not with primers E13 and U1, which is located in the upstream region (Fig. 1). A 3' RACE carried out to characterize precisely the 3' end of the antisense RNA revealed that it coincides with an A-rich region located in the *CG13617* 5' UTR in the four *2j* lines. Unfortunately, this A-rich region may provide a binding site for the oligo(dT) primer used in the cDNA synthesis, and we cannot be certain that the antisense transcript really ends there. The antisense RNA thus has an estimated length of ≈ 3 kb and includes the complete coding region of *CG13617*, as well as all its introns (Fig. 1). Conceptual translation of the *2j* antisense RNA sequence revealed 23 small ORFs of 53–143 aa, but none shared significant homology with any known protein, suggesting that this transcript does not have coding capability.

Discussion

CG13617 was described as a potential ORF in the genome of *D. melanogaster* (26). We have shown that the sequence and structure of *CG13617* are conserved in *D. buzzatii* and that *CG13617* is a fully functional gene expressed through the entire life cycle. *D. buzzatii* *CG13617* is transcribed into a 2.3-kb mRNA that encodes a 734-aa protein that presents a 59.7% identity (75.2% similarity) with the 737-aa protein predicted in *D. melanogaster*. Both proteins share several domains and structural characteristics typical of transcription factors (30): (i) a C2H2 zinc finger, (ii) four regions of 27–44 aa able to form coiled coils (one of the principal protein oligomerization motifs), (iii) a putative nuclear localization signal, and (iv) two PEST sequences (a motif involved in targeting proteins for rapid destruction) (Fig. 8, which is published as supporting information on the PNAS web site). An exhaustive search of homologous proteins in other species outside *Drosophila* detected similar proteins in *Anopheles gambiae*, *Rattus norvegicus*, *Mus musculus*, and *Homo sapiens* (Fig. 8). However, none of these proteins has a known function yet.

In *D. buzzatii*, *CG13617* is found adjacent to the proximal breakpoint of the widespread *2j* inversion, providing a unique opportunity to investigate possible position effects. Our results show that the stop codon of this gene is only 12 bp from the breakpoint and that the TE insertions responsible for the generation of the inversion (15, 23) took place inside the 3' UTR of the gene and altered the end of the *CG13617* transcriptional unit in the *2j* lines (Fig. 2). More importantly, the expression level of *CG13617* was compared between *2st* and *2j* homozygous lines in embryos, larvae, pupae, and adults. Despite the proximity of the inversion breakpoint, no differences in *CG13617* expression were detected in any developmental stage other than embryos, in which the average expression level was 5 times lower in *2j* than in *2st* lines (Fig. 4).

Several causes could explain the decrease in *CG13617* expression, such as the change of position of a downstream enhancer (31), a silencing effect of the repetitive DNA blocks at the breakpoint similar to that of heterochromatin (14), or problems in the mRNA processing due to the modification of the 3' UTR (32). However, the specific down-regulation affecting only *CG13617* expression in embryos (but not *Pp1 α -96A*, which is also located very close to the breakpoints, or in any other developmental stage) makes most of these explanations unlikely. In addition, other evidences indicate that the silencing is caused by an antisense transcript overlapping the whole *CG13617* coding region. First, the reduction of *CG13617* expression levels occurs only in *2j* embryos, which have the highest amounts of antisense RNA. No differences in the expression level of this gene were observed in other developmental stages, in which



Fig. 6. Detection of *CG13617* double-stranded RNA (dsRNA) in *st-1* and *j-19* embryos. RNA was treated with RNase to degrade single-stranded RNA, denatured, retrotranscribed, and amplified according to the methods of ref. 44. Amplification was carried out with two successive PCRs using nested primer pairs E11/E12 and E33/E34. A band of 318 bp corresponding to the dsRNA is observed in *2j*, but not in *2st*, embryos. For each sample, a negative control (NC) without the denaturation step was performed. The first two lanes correspond to genomic DNA of each line.

the antisense transcript was not found or was found in only very small quantities (Fig. 5B), probably insufficient to cause gene silencing (33). Second, the expression levels of the sense and antisense transcripts are negatively correlated in embryos (Fig. 4). In *2j* lines, a 5-fold increase in the level of antisense transcript is accompanied by a decrease in *CG13617* mRNA level compared with *2st* lines, where the antisense RNA is almost undetectable. Third, the intermediate expression of the antisense RNA in heterozygotes for *2j* inversion, together with the low levels of *CG13617* expression similar to those of *2j* lines, are consistent with a dominant effect acting in trans to silence both copies of the gene. Finally, the higher expression of the antisense RNA in *2j* embryos is, apart from the TE insertions at the breakpoints and the inversion, the only characteristic common to all *2j* lines that differentiates them from *2st* ones.

The use of antisense transcripts as a mechanism of expression down-regulation acting at the posttranscriptional level has been reported in an increasing number of genes in many species ranging from prokaryotes to humans, in which the occurrence of sense-antisense transcriptional units is a more common phenomenon than previously thought (29, 34). Antisense RNAs may control gene expression at various levels (35) and have been implicated in diverse processes including genomic imprinting (36), DNA methylation (37), X-inactivation (38), RNA editing (39), and transposon silencing (40), aside from being part of the natural regulation system of some genes (41). The mechanism of action of antisense RNAs is triggered by the formation of double-stranded RNA (dsRNA) duplexes with the sense mRNA. These duplexes are then cleaved into small pieces of 21–23 nucleotides that target the degradation of the complementary transcripts in a process known as RNA interference (42, 43). According to this model, the increased expression of the antisense RNA could be responsible for the elimination of part of the *CG13617* mRNA in embryos carrying the *2j* inversion. We investigated this possibility by testing the presence of *CG13617* dsRNA in embryos (44). As expected, after the dsRNA isolation, a denaturation-dependent amplification product was obtained in *2j*, but not in *2st*, embryos (Fig. 6), suggesting the existence of sense-antisense duplexes only in embryos carrying the *2j* inversion.

In *2j* lines, the antisense transcript extends from the TEs inserted at the inversion breakpoint (Fig. 1). Although the exact localization of its 5' end could not be determined, transcription appears to start inside the *Kepler* element inserted at the proximal breakpoint in all *2j* lines (23). This *Kepler* element could be contributing a promoter that drives the synthesis of the antisense transcript and causes the silencing of *CG13617* in *2j* embryos. A similar situation has been recently described in wheat, where retrotransposons inserted throughout the genome generate sense or antisense transcripts of the adjacent genes that increase or silence, respectively, the expression of these genes (33). Previously, the *Drosophila* TE *Hoppel* had been shown to silence the *Stellate* gene through a similar mechanism (44). In the *2j* inversion, *Kepler* is a DNA transposon with structural similarity to the *Foldback* family (23). Although no ORF encoding a putative transposase has been found and the element is probably

defective, a functional promoter might still be present. Besides, TE activity can be tightly regulated, and it is not unusual that TE promoters are active only in some specific tissues or developmental stages (45, 46).

Our results provide a clear example of a position effect associated with the breakpoints of a *Drosophila* natural inversion. The *2j* inversion of *D. buzzatii* is widespread in natural populations (24) and can be considered quite successful evolutionarily. Two events may have contributed to the formation of an antisense RNA overlapping the *CG13617* gene: the *Kepler* insertion and the *2j* inversion. If the TE insertion happened first, the ectopic recombination, which probably took place in the fragment of *Galileo* adjacent to *CG13617*, situated *Kepler* in its current position and orientation, allowing the antisense transcription of the gene. Alternatively, *Kepler* could have inserted in an already inverted chromosome assisted by the reduction of recombination at the breakpoints (47). The generation of antisense RNA and consequent silencing of *CG13617* might then have resulted in a favorable mutation that swept all other *2j* chromosomes in the population, causing perhaps an increase in inversion frequency. This would explain the observed discrepancy between the younger age of the different *2j* alleles at the breakpoints compared with the inversion *2j* itself, which is much older (23, 48). The consequences of the 5-fold decrease in *CG13617* expression level in *2j* embryos are hard to guess, because the function of the protein encoded by this gene remains unknown. However, it seems legitimate to speculate that the change in *CG13617* expression could be related to the pheno-

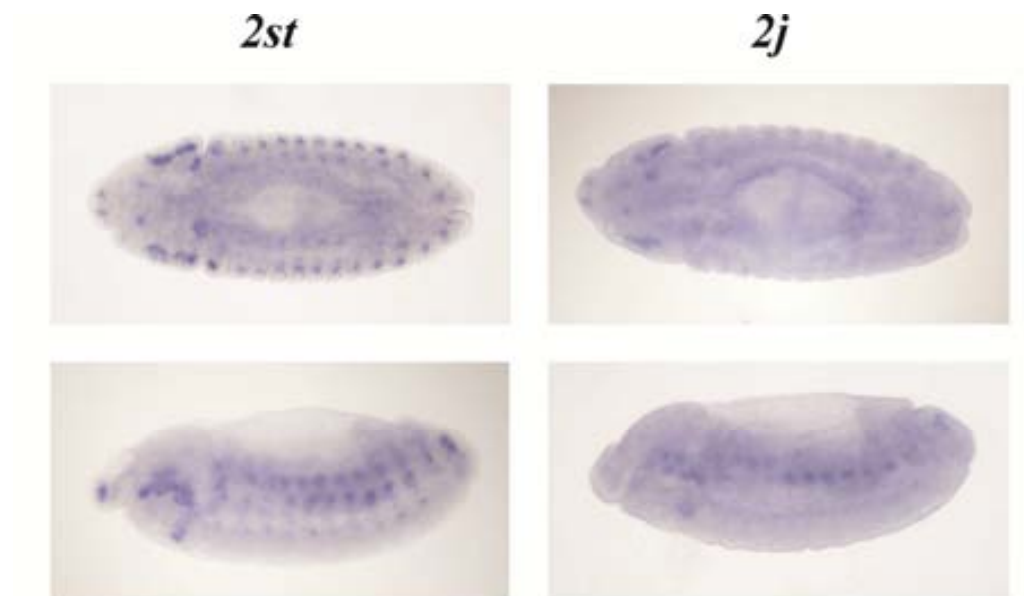
typic differences observed in adult body size and developmental time between individuals with the *2st* and *2j* arrangements (25) and the adaptive effect of the inversion. The *CG13617* expression pattern in embryos and the presence of motifs characteristic of transcription factors suggest a role in development consistent with significant fitness effects.

An important aspect in this case is that the silencing effect is not caused by the inversion itself, but by one of the TEs inserted at the breakpoint junctions. TEs have been largely recognized as an important source of genetic variation in the shaping of genomes and the adaptation of organisms to the environment (18, 49). Here, we show that they were not only involved in the origin of the *2j* inversion, but also in the regulation of the expression of a gene adjacent to its breakpoints by a mechanism of transcriptional interference (22). Because most inversions in *Drosophila* and other organisms are probably generated by TEs (15, 16), this kind of position effect may be much more common than previously thought.

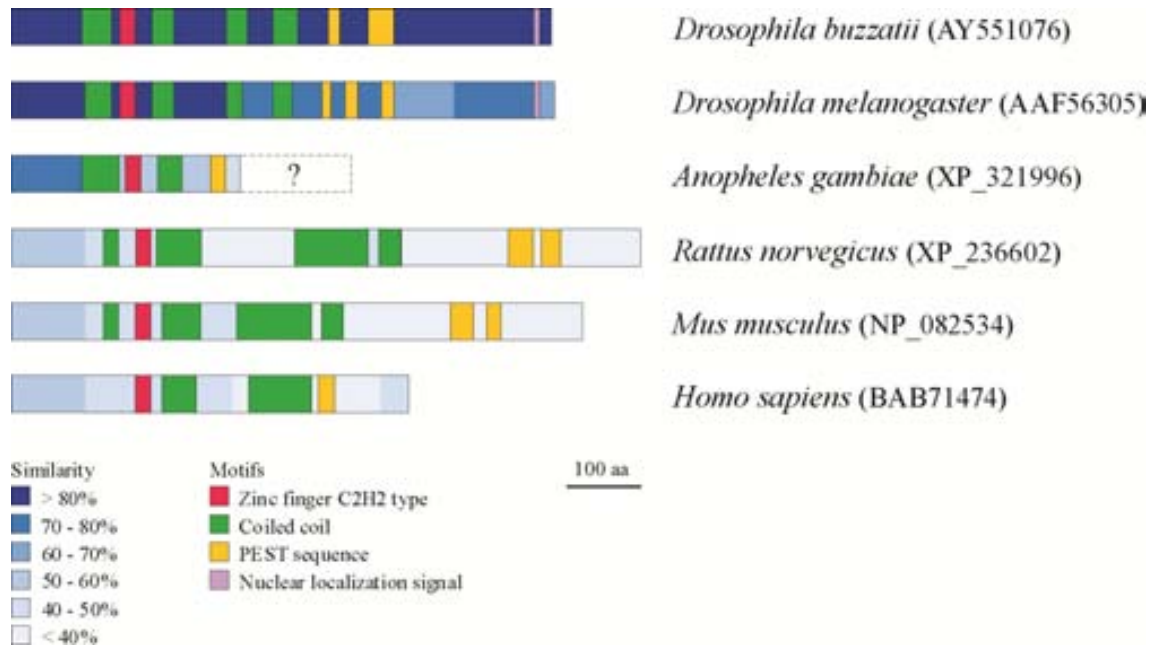
We thank J. Hidalgo and J. Carrasco for assistance with the Northern blotting, J. Casanovas and M. Llimargas for help with *in situ* hybridization in embryos, and M. Ashburner, S. Campuzano, J. Modollet, M. L. Pardue, and J. M. Ranz for valuable comments. This work was supported by Dirección General de Investigación, Ministerio de Ciencia y Tecnología (Spain) Grant BMC2002-01708 (to A.R.) and a doctoral fellowship from the Departament d'Universitats, Recerca i Societat de la Informació (Generalitat de Catalunya, Spain; to M.P.).

- Krimbas, C. B. & Powell, J. R., eds. (1992) in *Drosophila Inversion Polymorphism* (CRC, Boca Raton, FL), pp. 1–52.
- Lewontin, R. C., Moore, J. A., Provine, W. B. & Wallace, B. (1981) *Dobzhansky's Genetics of Natural Populations, I-XLIII* (Columbia Univ. Press, New York).
- Prevosti, A., Ribó, G., Serra, L., Aguadé, M., Balanyà, J., Monclús, M. & Mestres, F. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 5597–5600.
- Dobzhansky, T. (1970) *Genetics of the Evolutionary Process* (Columbia Univ. Press, New York).
- Charlesworth, B. (1974) *Genet. Res.* **23**, 259–280.
- Sperlich, D. (1966) *Genetics* **53**, 835–842.
- Hartl, D. L. (1977) in *Measuring Selection in Natural Populations*, ed. Levin, S. (Springer, Heidelberg), pp. 65–82.
- Cáceres, M., Barbadilla, A. & Ruiz, A. (1999) *Genetics* **153**, 251–259.
- Wesley, C. S. & Eanes, W. F. (1994) *Proc. Natl. Acad. Sci. USA* **91**, 3132–3136.
- Hough, R. B., Lengeling, A., Bedian, V., Lo, C. & Bucan, M. (1998) *Proc. Natl. Acad. Sci. USA* **95**, 13800–13805.
- Lakich, D., Kazazian, H. H., Jr., Antonarakis, S. E. & Gitschier, J. (1993) *Nat. Genet.* **5**, 236–241.
- Frischer, L. E., Hagen, F. S. & Garber, R. L. (1986) *Cell* **47**, 1017–1023.
- Lister, C., Jackson, D. & Martin, C. (1993) *Plant Cell* **5**, 1541–1553.
- Henikoff, S. (1990) *Trends Genet.* **6**, 422–426.
- Cáceres, M., Ranz, J. M., Barbadilla, A., Long, M. & Ruiz, A. (1999) *Science* **285**, 415–418.
- Casals, F., Cáceres, M. & Ruiz, A. (2003) *Mol. Biol. Evol.* **20**, 674–685.
- Britten, R. J. (1997) *Gene* **205**, 177–182.
- Kidwell, M. G. & Lisch, D. (1997) *Proc. Natl. Acad. Sci. USA* **94**, 7704–7711.
- Rubin, G. M. (1983) in *Mobile Genetic Elements*, ed. Shapiro, J. A. (Academic, Orlando, FL), pp. 329–361.
- KloECKener-Gruisssem, B., Vogel, J. M. & Freeling, M. (1992) *EMBO J.* **11**, 157–166.
- Lerman, D. N., Michalak, P., Helin, A. B., Bettencourt, B. R. & Feder, M. E. (2003) *Mol. Biol. Evol.* **20**, 135–144.
- Whitelaw, E. & Martin, D. I. (2001) *Nat. Genet.* **27**, 361–365.
- Cáceres, M., Puig, M. & Ruiz, A. (2001) *Genome Res.* **11**, 1353–1364.
- Hasson, E., Rodríguez, C., Fanara, J. J., Naveira, H., Reig, O. A. & Fontdevila, A. (1995) *J. Evol. Biol.* **8**, 369–384.
- Betrán, E., Santos, M. & Ruiz, A. (1998) *Evolution* **52**, 144–154.
- Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F., et al. (2000) *Science* **287**, 2185–2195.
- Ruiz, A. & Wasserman, M. (1993) *Heredity* **70**, 582–596.
- De, S. K., McMaster, M. T. & Andrews, G. K. (1990) *J. Biol. Chem.* **265**, 15267–15274.
- Shendure, J. & Church, G. M. (Aug. 22, 2002) *Genome Biol.* **3**, RESEARCH0044.1–0044.14.
- Banham, A. H., Beasley, N., Campo, E., Fernandez, P. L., Fidler, C., Gatter, K., Jones, M., Mason, D. Y., Prime, J. E., Trougouff, P., et al. (2001) *Cancer Res.* **61**, 8820–8829.
- Wray, G. A., Hahn, M. W., Abouheif, E., Balhoff, J. P., Pizer, M., Rockman, M. V. & Romano, L. A. (2003) *Mol. Biol. Evol.* **20**, 1377–1419.
- Proudfoot, N. J., Furger, A. & Dye, M. J. (2002) *Cell* **108**, 501–512.
- Kashkush, K., Feldman, M. & Levy, A. A. (2003) *Nat. Genet.* **33**, 102–106.
- Yelin, R., Dahary, D., Sorek, R., Levanon, E. Y., Goldstein, O., Shoshan, A., Diber, A., Biton, S., Tamir, Y., Khosravi, R., et al. (2003) *Nat. Biotechnol.* **21**, 379–386.
- Knee, R. & Murphy, P. R. (1997) *Neurochem. Int.* **31**, 379–392.
- Sleutels, F., Zwart, R. & Barlow, D. P. (2002) *Nature* **415**, 810–813.
- Tufarelli, C., Stanley, J. A., Garrick, D., Sharpe, J. A., Ayyub, H., Wood, W. G. & Higgs, D. R. (2003) *Nat. Genet.* **34**, 157–165.
- Lee, J. T., Davidow, L. S. & Warshawsky, D. (1999) *Nat. Genet.* **21**, 400–404.
- Kumar, M. & Carmichael, G. G. (1997) *Proc. Natl. Acad. Sci. USA* **94**, 3542–3547.
- Sijen, T. & Plasterk, R. H. A. (2003) *Nature* **426**, 310–314.
- Kramer, C., Loros, J. J., Dunlap, J. C. & Crosthwaite, S. K. (2003) *Nature* **421**, 948–952.
- Hammond, S. M., Caudy, A. A. & Hannon, G. J. (2001) *Nat. Rev. Genet.* **2**, 110–119.
- Pickford, A. S. & Cogoni, C. (2003) *Cell. Mol. Life Sci.* **60**, 871–882.
- Aravin, A. A., Naumova, N. M., Tulin, A. V., Vagin, V. V., Rozovsky, Y. M. & Gvozdev, V. A. (2001) *Curr. Biol.* **11**, 1017–1027.
- Laski, F. A., Rio, D. C. & Rubin, G. M. (1986) *Cell* **44**, 7–19.
- Bronner, G., Taubert, H. & Jackle, H. (1995) *Chromosoma* **103**, 669–675.
- Bartolomé, C., Maside, X. & Charlesworth, B. (2002) *Mol. Biol. Evol.* **19**, 926–937.
- Laayouni, H., Hasson, E., Santos, M. & Fontdevila, A. (2003) *Mol. Biol. Evol.* **20**, 931–944.
- McDonald, J. F. (1995) *Trends Ecol. Evol.* **10**, 123–126.

Supporting information



SUPPORTING FIGURE 7 | Expression pattern of *CG13617* in *Drosophila buzzatii* embryos analyzed by *in situ* hybridization. Dorsal (*Upper*) and lateral (*Lower*) views of stage 11-12 embryos are shown for two lines with the *2st* and *2j* arrangement. In both cases, several signals in the head and in all metamers, forming two lines at each side along the embryos, can be observed, with higher intensity in the *2st* than in the *2j* line. This pattern resembles that of the genes expressed in the chordotonal organs of the peripheral nervous system in *Drosophila* (JARMAN et al. 1993). Whole-mount *in situ* hybridization was performed as described in LEHMAN and TAUTZ (1994) by using embryos from lines st-1 and j-19. The same 1128-bp *CG13617* antisense riboprobe as in the Northern analysis was used, but it was nonradioactively labeled with digoxigenine-UTP. Hybridization was performed overnight at 65 °C, and staining times for *2st* and *2j* embryos were identical. Similar hybridizations with a sense probe did not yield repeatable results (data not shown).



SUPPORTING FIGURE 8 | Diagram of the *Drosophila* CG13617 protein and similar proteins of other species. In each case, percent similarity with the *Drosophila buzzatii* protein sequence is indicated every 100 aa by different color intensities. Putative functional domains and motifs found in each of the proteins are colored according to the key. The *Anopheles* protein is incompletely annotated in the database, and only the first 312 aa are shown. Given the homology with the other proteins, an alternative initial methionine located 181 aa downstream of the predicted start codon has been represented here for the rat protein. The different protein motifs were identified by using the following software: INTERPROSCAN (ZDOBNV and APWEILER 2001) for the C2H2 zinc finger, COILS (LUPAS *et al.* 1991) for the coiled coils, PESTFIND (RECHSTEINER and ROGERS 1996) for the PEST sequences and PSORTII (NAKAI and KANEHISA 1992) for the nuclear localization signal.

SUPPORTING TABLE 3 | Primers used in this study.

A		Experiments	Primer pairs	B	
				Primer name	Sequence (5' - 3')
CG13617 DNA sequencing and variation study			S1-E21	AS1	TCCCTAAAGACTAAGTAAGTAACCATATTACATT
			U1-E13	AS2	AAGAATACATCCATACATTTCCGTTCT
			E11-E12	D1 §	CCTCGTAAGCGTGTATGTTTC
			E9-E10	D2 §	CCGGCGGCTCATGTTGTTTCTAAGC
			E1-E5	E1	AAGACCATATCCAACACAACC
			U2*	E2	AGGCITTTGCTTGTGTATTTTG
			U3*	E5	AATTTATTCCGGCGCTATCTTC
				E6	CCAACACAACCGCTAAACGTT
	CG13617 mRNA sequencing		E11-E12	E7	CAAAGGGCCAGACTGAAAATTG
			E9-E10	E9	ATTTTCAGTCTGGCCCTTTGC
	CG13617 RACE	5' sense mRNA	E19 †	E10	CTTGAGGACTTGGAGCGTAT
			E13-oligodT ‡	E11	CGCCTTAGGTCTCGTTCACAG
			E18-anchor ‡	E12	CGAGAGGCGCACGAAAATATC
		3' sense mRNA	E1-poliT	E13	GGGAAATACAACCTCGCGTA
			E6-E6	E18	TGCAACTGCACCACCTCATTG
		3' antisense RNA	E13-poliT	E19	GCACATTCAGAGCCTCGTTGC
	E18-E6		E21	GTCCAGAGGCCCATCTCGATA	
	CG13617 probe cloning			E25	TTAGCGGTTGTGTTGGATATGG
	Semi-quant. RT-PCR	CG13617	E2-E7	E27	CCCTAAAGACTAAACTCCCACCAAA
		Gapdh	H1-H2	E30	TCCTTGCGAGTCAGTGGCTT
	Real time RT-PCR	CG13617 mRNA	E25-E27	E31	CAACAAAATCTGAGCACGTTCTG
		Antisense RNA	AS1-E30	E33	TGCAGTCATCAATACCATCA
			AS2-E31	E34	TGAAGTCCTCATTGCTCTGCT
Pp1α-96A		P5-P6	G1 §	CGCTCAGAAGGGAACCAATGGGA	
Gapdh	H8-H9	G9 §	CGTCGAGTATCACTTGTATAGG		
Gapdh sequencing			G11 §	GTCAACCTAACTGAGCAAGTG	
		H1-H2	G16 §	CGAAGCGGAATGATTTTGCCA	
		H1-H7	G17	GTCCAGTCTATGTGTGTAAG	
Antisense analysis			H1 §	ATGTCGAAGATTGGTATTAATGG	
			U1-E13	H2 §	GTTTCGACACGACCTTCATGT
			I1-E11	H7	TTAATCCTTGCTCTGCATGTA
			I2-E9	H8	CTGCCAACGGTCCATTGAA
			I3-E2	H9	GAGTGAGTGTGCTGAGGAAGTC
			G17-E7	I1	AGTTGTGATGCCTTGTAAAATG
			G16-E7	I2	GTAAGTAACCATATTACATTA
			G1-D2	I3	ACATCCATACATTTCCGTTCT
			G9-D1	K1	TCCAGTCTATGTGTTGTATGG
			K1-D1	K2	GAATAGCACACAGCGGACTTC
			T5-D2	P5	GCAGGTGTCAACAGCAGCAA
K2-D1	P6	GGCACCACGTACTTCCAACA			
G11-D1	poliT	CCAACACAACCGCTAAACGTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT			
dsRNA detection			S1	CAACCAAAGCCAACGACACAT	
		E11-E12	T5 §	CGAGCAAATACGAAGATGACT	
		E33-E34	U1	TGAGATGAGGGAGGCAGATA	
			U2	TTTGATCAGAGACATAAGAAC	
			U3	TAATAATGCTGGATAGAACAA	

* Primers used only for sequencing
† Primer used for cDNA synthesis
‡ Primers supplied with the RACE kit
§ Primers used previously in other studies

SUPPORTING TABLE 4 | Structural variation in the *CG13617* gene region analyzed by PCR amplification and restriction enzyme digestion of PCR products.

Line	Amplification products (kb)*				
	S1-E21	U1-E13	E11-E12	E9-E10	E1-E5
st-1	1.47	0.93	0.84	0.81	0.60
st-4	1.47	0.93	0.84 ^H	0.81 ^D	0.60
st-7	1.47	0.93	0.84	0.81 ^D	0.60
st-8	1.47	0.93	0.84 ^H	0.81 ^D	0.60
j-1	1.69 ^D	0.96 ^S	0.84	0.81	0.60
j-13	1.47 ^D	0.93	0.84	0.81 ^P	0.60
j-19	1.69 ^D	0.96 ^S	0.84	0.81 ^P	0.60
jz ³ -4	2.35 ^D	0.93	0.84	0.81	0.60

*Amplification products of each line were digested with the following restriction enzymes: S1-E21, *DraI*; U1-E13, *SalI*; E11-E12, *HindIII*; E9-E10, *DraI* and *PstI*; E1-E5, *XmnI*. Polymorphisms in the restriction sites of these enzymes compared to the st-1 sequence are indicated: D, *DraI*; S, *SalI*; H, *HindIII*; P, *PstI*.

SUPPORTING TABLE 5 | Real-time RT-PCR relative measurements of *CG13617* mRNA, antisense RNA, and *Pp1 α -96A* mRNA expression levels in embryos of homozygous lines with and without inversion *2j*. In section A, *CG13617* and *Pp1 α -96A* expression levels are given relative to those of line st-1, whereas antisense RNA expression levels are shown in relation to those of line j-1.

A. Summary of data

Arrangement	Line	<i>CG13617</i> mRNA			Antisense RNA			<i>Pp1α-96A</i> mRNA		
		n	Mean	SD	n	Mean	SD	n	Mean	SD
<i>2st</i>	st-1	3	1.000	0.785	3	0.080	0.084	3	1.000	0.269
	st-4	3	1.785	1.596	3	0.369	0.261	3	1.309	0.166
	st-7	3	0.949	0.768	3	0.183	0.105	3	0.708	0.240
	st-8	3	0.742	0.319	3	0.073	0.088	3	0.770	0.381
<i>2j</i>	j-1	3	0.139	0.086	3	1.000	0.432	3	1.083	0.242
	j-13	3	0.076	0.060	3	0.335	0.058	3	0.822	0.212
	j-19	3	0.247	0.142	3	1.350	0.880	3	1.268	0.169
	jz ³ -4	3	0.438	0.248	3	0.969	0.425	3	0.563	0.184

B. ANOVA table

Source of variation	df	<i>CG13617</i> mRNA			Antisense RNA			<i>Pp1α-96A</i> mRNA		
		SS	MS	F	SS	MS	F	SS	MS	F
Between arrangements	1	4.795	4.795	13.62**	3.268	3.268	11.00*	0.001	0.001	0.004 n.s.
Among lines within arrangements	6	2.113	0.352	0.71 n.s.	1.780	0.297	1.92 n.s.	1.519	0.253	4.29**
Within lines (error)	16	7.897	0.494		2.481	0.155		0.938	0.059	

* $P \leq 0.05$; ** $P \leq 0.01$; n.s. = non significant.

SUPPORTING TABLE 6 | Real-time RT-PCR relative measurements of *CG13617* mRNA and antisense RNA in embryos homozygous and heterozygous for inversion *2j*. The heterozygous genotypes indicate the maternal/paternal lines used in the reciprocal crosses. In section A, *CG13617* mRNA expression levels are given relative to that of *2st/st* homozygotes (line st-1), whereas antisense RNA expression levels are shown in relation to that of *2j/j* homozygotes (line j-1).

A. Summary of data

Genotype	<i>CG13617</i> mRNA			Antisense RNA		
	n	Mean	SD	n	Mean	SD
<i>2st/st</i>	3	1.000	0.413	3	0.041	0.072
<i>2st/j</i>	3	0.333	0.072	3	0.276	0.057
<i>2j/st</i>	3	0.097	0.069	3	0.222	0.142
<i>2j/j</i>	3	0.146	0.155	3	1.000	0.469

B. ANOVA table

Source of variation	df	<i>CG13617</i> mRNA			Antisense RNA		
		SS	MS	F	SS	MS	F
Between groups	3	1.562	0.521	10.22**	1.605	0.535	8.63**
<i>2st/st vs. 2j/j</i>	1	1.094	1.094	21.45**	1.381	1.381	22.27**
<i>2st/j vs. 2j/st</i>	1	0.084	0.084	1.65 n.s.	0.004	0.004	0.06 n.s.
Heterozygotes <i>vs.</i> Homozygotes	1	0.384	0.384	7.53*	0.220	0.220	3.55 n.s.
Within groups	8	0.410	0.051		0.497	0.062	

* $P \leq 0.05$; ** $P \leq 0.01$; n.s. = non significant.

3.2 Functional consequences of *CG13617* silencing

To investigate the potential functional consequences of *CG13617* silencing, in this work we have tried to reproduce the reduction of *CG13617* gene expression found in *D. buzzatii* 2j embryos in a different species, *D. melanogaster*. The objective of using this model species is to take advantage of the available genome information and be able to carry out microarray experiments to compare the expression profiles between samples with and without expression of the *CG13617* gene. These experiments could not be performed entirely in

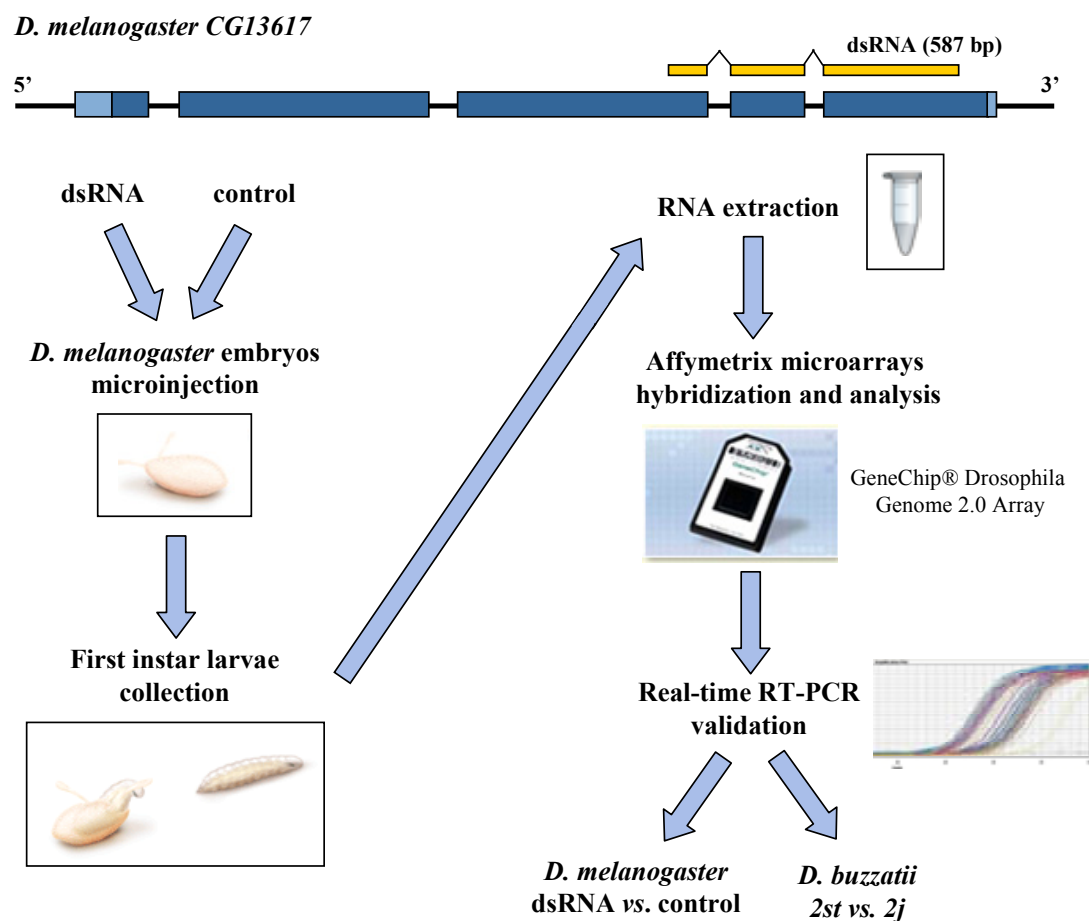


Figure 17 | Experimental design of the functional analysis of the consequences of *CG13617* silencing. In the upper part of the figure, blue rectangles represent *D. melanogaster CG13617* exons (coding regions are in dark blue and UTRs in light blue). Yellow rectangles above the gene picture mark the parts of the gene included in the dsRNA molecule. In the lower part, the different procedures used in this work are specified, with arrows indicating the order in which they were performed. See the text for details on each of these processes.

D. buzzatii because its genome sequence is needed to design the probes contained in the oligonucleotide microarrays and *D. buzzatii* genome has not been sequenced yet, so there are no available microarrays for this species. However, the detection of genes presenting expression changes when *CG13617* expression is eliminated in *D. melanogaster* can provide valuable clues about *CG13617* function and help to elucidate the consequences of its silencing in 2j chromosomes. FIGURE 17 shows an outline of the experimental design.

3.2.1 Silencing of *D. melanogaster* *CG13617* gene expression by RNAi

The silencing of gene *CG13617* was carried out in *D. melanogaster* using RNA interference techniques (see BOX 2). This method was chosen over other gene inactivation techniques because it uses the same silencing mechanism that is supposedly acting in *D. buzzatii*. In this case, the trigger of the RNAi pathway was not the dsRNA formed by the pair *CG13617* mRNA-antisense RNA (PUIG *et al.* 2004), but an exogenous dsRNA molecule with the sequence of *D. melanogaster* *CG13617* gene. In order to accomplish this, a 587-bp long dsRNA containing the *CG13617* coding sequence from the end of exon 3 to the end of exon 5 was synthesized (FIGURE 11, see Materials and Methods). The final part of the gene was chosen because it seems to provide a more specific target sequence for the dsRNA, since it does not include any functional domains and it is less conserved among species (see below). This is important in order to achieve a specific silencing of *CG13617* because the dsRNA molecule could unspecifically bind other mRNAs and cause their degradation if there are regions of sequence similarity between them.

The purified dsRNA molecule was injected in young *D. melanogaster* embryos (laid during the previous hour) that are still in the syncytial blastoderm stage to ensure that the injected material is able to diffuse through the whole embryo and reach all the future cells of the individual. It is important to note that this kind of silencing is transient because the amount of injected dsRNA gets diluted and lost as development progresses (CARTHEW 2003). Therefore, as the dsRNA disappears, the regular expression of the silenced gene is restored. Thanks to this property the artificial system created in *D. melanogaster* resembles even more what is naturally occurring in *D. buzzatii*, where *CG13617* gene expression is reduced only in

the embryonic stage in *2j* individuals but not in any of the other developmental stages (larvae, pupae and adults) where the gene is expressed normally and at similar levels to those of *2st* chromosome carriers.

In addition, control embryos (where *CG13617* is expressed normally) were microinjected following the same protocol, but only with injection buffer. This precaution was taken to try to avoid creating additional differences between the two types of samples that we want to compare (with and without *CG13617* expression) caused by the microinjection process. It has to be taken into account that only a certain fraction of the embryos survive microinjection independently of the introduced material. If all the analyzed embryos or larvae have undergone the same treatments (dechoriation, desiccation, injection, etc.) the samples are more comparable, since these processes could also lead to changes in gene expression that may be confused with those generated by the presence of the dsRNA.

After microinjection, we collected samples of three different developmental stages: embryos ~20 h old (this includes individuals that are already dead as well as alive embryos because it is impossible to distinguish them visually at this early stage), first instar larvae that have just hatched, and third instar larvae (which had been transferred to a vial with medium after hatching to allow them to continue their development). Total RNA was isolated from these three types of samples in control- and dsRNA-microinjected embryos. Semi-quantitative RT-PCR experiments were performed using primers DmE4-DmE6 to assess *CG13617* gene expression. These two primers amplify a 531-bp fragment from the mRNA close to the end of the *CG13617* coding sequence, the same region targeted by the dsRNA. In order to avoid the unspecific amplification of the injected molecule, primer DmE6 was designed outside the segment of the gene included in the dsRNA. The housekeeping gene *Gapdh1* was also amplified with primer pair DmH1-DmH2 in the same samples as an internal control for cDNA concentration. As expected, a band indicating the regular expression level of the *CG13617* gene was obtained in all the control samples (embryos and larvae) while no amplification product, or only a weak band suggesting a very low level of gene expression, could be observed in the samples that had been injected with the dsRNA (FIGURE 18). At the same time, a 499-bp *Gapdh1* band with similar intensity was amplified in all samples (FIGURE 18), suggesting that *Gapdh1* expression is unaffected by the fact that the samples were

microinjected with or without dsRNA and that the differences in *CG13617* gene expression levels can not be attributed to variation in cDNA concentrations among samples. Taken all together, these results show that *CG13617* silencing caused by the introduction of the dsRNA was effective and that it maintains its effects in developmental stages as advanced as third instar larvae.

Some individuals injected with dsRNA (and therefore with a silenced *CG13617* gene) were allowed to develop in vials with medium. They were able to complete development successfully and reproduce normally. So, a reduced expression level of this gene in the early stages of development does not seem to be critical for survival. This is consistent with the fact that *2j* inversion carriers in *D. buzzatii* do not present any visible defect or seem to have any problems in spite of exhibiting a lower *CG13617* expression level in embryos with respect to *2st* individuals.

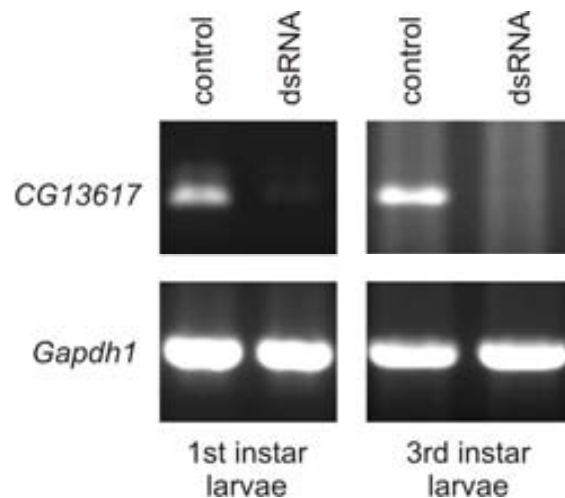


FIGURE 18 | Silencing of *CG13617* gene expression in *D. melanogaster* by RNAi. Semi-quantitative RT-PCR results for genes *CG13617* and *Gapdh1* in first and third instar larvae. The material used to microinject the embryos is indicated for each column.

3.2.2 Detection of gene-expression changes induced by *CG13617* silencing using microarrays

The next step of the work entailed comparing the expression profiles of the samples with a normal expression of gene *CG13617* and those with this gene silenced by the introduced dsRNA. In order to accomplish this, Affymetrix oligonucleotide microarrays (see Box 3) capable of analyzing ~18500 transcripts produced in the *D. melanogaster* genome were used. Our objective was to determine if there are other genes that change their expression levels when *CG13617* expression is greatly reduced, which could provide valuable clues about the cellular processes in which the protein encoded by *CG13617* might be involved and, more importantly, about the consequences of *CG13617* silencing in *D. buzzatii* 2j embryos.

Even though *CG13617* silencing in *D. buzzatii* was observed to happen specifically in embryos, for gene-expression analysis in *D. melanogaster* first instar larvae were used instead of embryos (FIGURE 17). This more advanced developmental stage allowed us to make sure that all the analyzed individuals were alive at the moment of collection and that no dead embryos that did not survive the microinjection process were included in the experiment. Dead

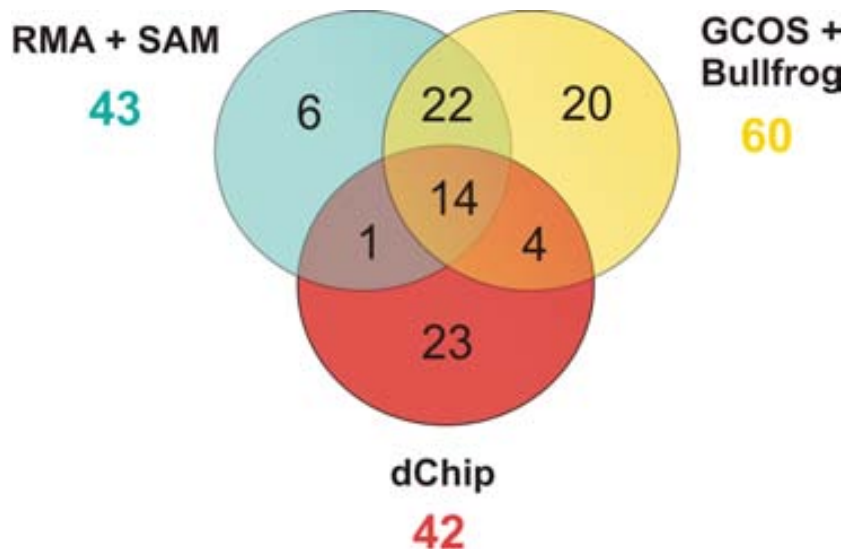


FIGURE 19 | Venn diagram showing the results of the microarray analysis using three independent methods. The total number of genes exhibiting expression changes is indicated in blue for each analysis. The 41 genes detected in at least two analyses (intersecting areas) constitute the list of differentially expressed genes (TABLE 6).

embryos may present different expression patterns or have degraded RNA, which could both alter the microarray results. A total of four microarrays were hybridized with the RNAs of four completely independent samples: two samples designated as CONTROL that express *CG13617* gene normally (injected with buffer), and two samples labeled as DSRNA where this gene has been silenced by RNAi (injected with the dsRNA). These four samples were tested by RT-PCR for *CG13617* expression level to check whether the silencing process had worked properly in the DSRNA samples before proceeding to their labeling and hybridization to the microarrays. So, two microarrays with expression data corresponding to CONTROL larvae and another two microarrays with expression profiles from DSRNA larvae were finally obtained.

To include all possible gene-expression changes caused by *CG13617* silencing and given the small number of hybridized microarrays, which complicates statistical analysis, microarrays were analyzed and compared with three commonly used independent methods: RMA and SAM, GCOS and BULLFROG, and DCHIP (see Materials and Methods for the specific criteria used in each case). The analysis with RMA and SAM yielded a list of 43 genes with expression differences. 60 genes exhibited expression changes according to the GCOS and BULLFROG analysis and 42 when DCHIP software was used (FIGURE 19). Genes were considered to be differentially expressed between the CONTROL and DSRNA samples if they were classified as presenting a significant expression change in at least two of the three analyses (FIGURE 19). As a result, a list containing 41 differentially expressed genes was obtained (TABLE 6). Surprisingly, all of them present a reduction of the expression level in the DSRNA samples with a silenced *CG13617* (FIGURE 20). There are no probe sets exhibiting a consistent increase in the expression level, except for the manipulated gene itself, *CG13617*, which shows a great increase (between ~25 and ~62 times, depending on the analysis) in the hybridization intensity in the DSRNA samples. This apparently contradictory result with respect to the silencing observed in the RT-PCR experiments, can be explained by the fact that 13 of the 14 25-nt probes that detect the *CG13617* transcript in the microarrays are located in the same region of the gene spanned by the dsRNA molecule. Therefore, what we are detecting in the microarray analysis is the large amount of injected dsRNA, which was also isolated and labeled together with the rest of the RNA and is capable of hybridizing to the microarray probes because they have the same sequence (FIGURE 21). This was tested by performing additional RT-PCRs with primer DmE5, that together with primer DmE4

TABLE 6 | List of differentially expressed genes determined by the combination of the results of the three different microarray analysis methods. Fold change values are given for the three analysis methods (RMA + SAM, GCOS + BULLFROG, and DCHIP). Grey boxes indicate the absence of the probe sets in the list of genes showing a difference in signal intensity obtained with that particular analysis. Colors represent the genes belonging to the different functional categories as shown in TABLE 7: dark yellow, DNA replication; light yellow, nucleic acid metabolism; orange, cell cycle; green, putative gene family with unknown function (see text for details). In some cases genes can be classified in more than one category. For the three first categories only the genes not included in any of the previous ones are highlighted in that specific color.

Gene	Probe set	Fold change		
		RMA + SAM	GCOS + BULLFROG	DCHIP
<i>lectin-24.A</i>	1637857_at	-2,45	-2,30	-2,57
<i>Ts</i>	1624747_at	-2,10	-3,14	-5,99
<i>mus209</i>	1623545_at	-3,16	-3,42	-3,60
<i>Obp56a</i>	1624074_at	-3,88	-4,14	-4,06
<i>DebB</i>	1628006_at	-2,12	-2,07	-2,14
<i>RnrS</i>	1635409_at	-2,32	-2,30	-2,34
<i>CG17974</i>	1640616_at	-2,26	-2,55	-3,04
<i>Hsp83</i>	1630688_at	-2,13	-2,14	-2,04
<i>pip</i>	1626952_at	-1,88	-2,30	-2,27
<i>Mcm2</i>	1632669_at	-2,50	-2,30	-2,53
<i>Mcm5</i>	1626647_at	-2,02	-2,26	-2,33
<i>CG8087</i>	1623467_at	-3,81	-4,92	-5,28
<i>CG7670</i>	1628316_at	-2,48	-2,73	-3,49
<i>dUTPase</i>	1634637_a_at	-2,10	-2,26	-2,32
<i>CG13135</i>	1625802_a_at	-5,63	-6,61	
<i>RnrL</i>	1631007_at	-2,98	-2,64	
<i>RanGap</i>	1623819_at	-2,48	-2,14	
<i>iclh</i>	1626871_at	-2,09	-2,30	
<i>CG6370</i>	1630630_at	-2,08	-2,00	
<i>CG13335</i>	1637275_a_at	-4,12	-3,93	
<i>Lcp4</i>	1630238_at	-2,13	-2,55	
<i>gp210</i>	1628986_at	-2,17	-2,00	
<i>CG32198</i>	1636183_at	-9,31	-14,42	
<i>Est-6</i>	1631224_at	-2,33	-2,22	
<i>Mcm7</i>	1631517_at	-2,37	-2,14	
<i>Hsp70.Ab</i>	1639571_s_at	-3,50	-2,93	
<i>Hsp70Bbb / Hsp70Ba</i>	1626821_s_at	-3,86	-3,36	
<i>Hsp70Ba</i>	1632841_x_at	-2,58	-2,73	
<i>CG14850</i>	1638815_at	-6,38	-7,34	
<i>CG10514</i>	1638424_at	-2,14	-2,55	
<i>CG15308</i>	1631616_at	-3,16	-3,54	
<i>CG7631</i>	1629098_at	-2,10	-1,93	
<i>Art1</i>	1632383_at	-1,94	-1,90	
<i>dnk</i>	1627691_at	-1,92	-1,93	
<i>CG3226</i>	1623702_at	-2,02	-1,93	
<i>CG40322</i>	1638720_x_at	-2,10	-5,37	
<i>CG18477 / CG31780</i>	1624619_s_at	-2,67		-3,11
<i>CG5602</i>	1630390_at		-2,03	-2,32
<i>CG31248</i>	1640854_at		-2,22	-2,45
<i>CycE</i>	1626249_s_at		-1,90	-1,99
<i>pont</i>	1635279_at		-1,87	-2,00

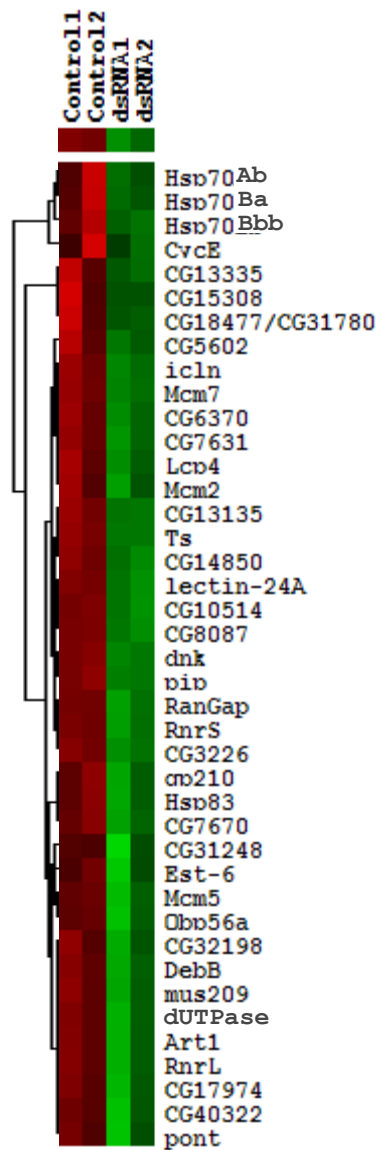


FIGURE 20 | Expression data of genes differentially expressed in samples with (CONTROL) and without (DSRNA) *CG13617* expression according to microarray analysis. Hybridization levels for 41 probe sets that show differences in signal intensity are represented in this figure. Columns correspond to the four analyzed samples. Rows represent the individual probe sets with the name of the corresponding gene indicated on the right (see TABLE 6 for more information). For each probe set, red, green, and black indicate increased, decreased, and equal hybridization levels relative to the median, respectively.

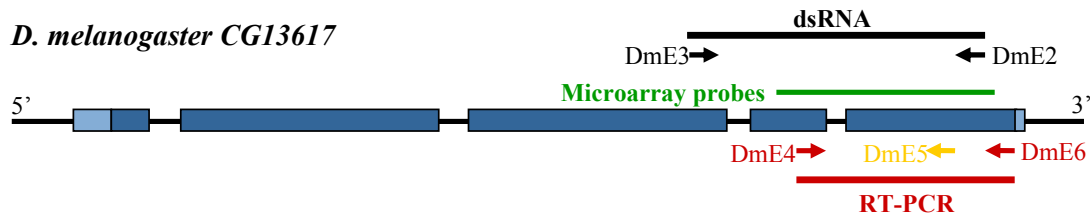


FIGURE 21 | Schematic representation of the location of microarray probes, dsRNA molecule, and amplification products in the *D. melanogaster CG13617* gene sequence. Blue rectangles represent *CG13617* exons. The black bar above corresponds to the sequence included in the injected dsRNA molecule together with the primers used to amplify that segment of the gene. The green bar signals the region where the 14 25-nt probes included in the microarrays hybridize. The red bar below the diagram of the gene represents the RT-PCR product used to assess the effectiveness of *CG13617* silencing by RNAi, with the primers used for the amplification also depicted in red. Primer DmE5 used to assess the presence of the injected dsRNA in the cDNAs used for RT-PCR amplification and microarray hybridization is indicated in yellow. Intronic sequences are not included in the dsRNA molecule, the microarray probes or the RT-PCR products (the bars comprise introns only for simplicity reasons).

generates a 425-bp product, which resulted in a more intense signal for *CG13617* expression than that obtained using primer DmE6, located outside the dsRNA sequence and therefore unable to amplify the injected dsRNA molecule (FIGURE 21).

Next, we performed a gene ontology analysis on the list of differentially expressed genes to see whether there are any overrepresented functional categories, or if, on the contrary, the proportion of genes classified in the distinct gene ontology categories does not differ significantly from what can be found considering the whole *D. melanogaster* genome. With respect to the biological process, a significant excess of genes belonging to functional categories related to DNA replication ($P = 2.6 \cdot 10^{-7}$) and cell cycle ($P = 2.97 \cdot 10^{-3}$) was found on our list (FIGURE 22 and TABLE 7). A total of 14 genes out of 26 in the differentially expressed list with assigned gene ontology categories (53.85%) can be classified in the category of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process, which includes genes involved in the synthesis of the precursors necessary for DNA replication. Eight of these genes (30.77%) are involved in DNA replication and related processes such as DNA replication initiation, DNA helicase activity or DNA-dependent DNA replication. These proportions contrast with the expected percentage of 22.5% of genes in a random list to be involved in nucleic acid metabolism and only 1.87% in DNA replication (FIGURE 22 and TABLE 7). Also, seven genes (26.92%) carry out functions related to cell cycle process and

regulation. It has to be taken into account that, as could be expected, there is a considerable overlap between the genes included in each of these groups (TABLE 7). For example, genes like *mus209* or *CycE* belong to the three functional categories mentioned above. Finally, there are other groups of genes showing a significant enrichment in our list: a total of 10 genes (38.46%) are involved in the response to stimulus, and more specifically in the response to stress. These groups are mainly determined by the presence in the list of differentially expressed genes of genes implicated in the response to heat or in protein folding, such as several heat shock proteins. Although most of the genes included in these last two functional groups are different from the ones contained in the previous categories, there are still some genes involved in DNA replication and cell cycle (such as *mus209* or *Hsp83*) that are also considered to carry out functions related to the response to stress (TABLE 7). Only 15 genes in our list were not clustered in this analysis.

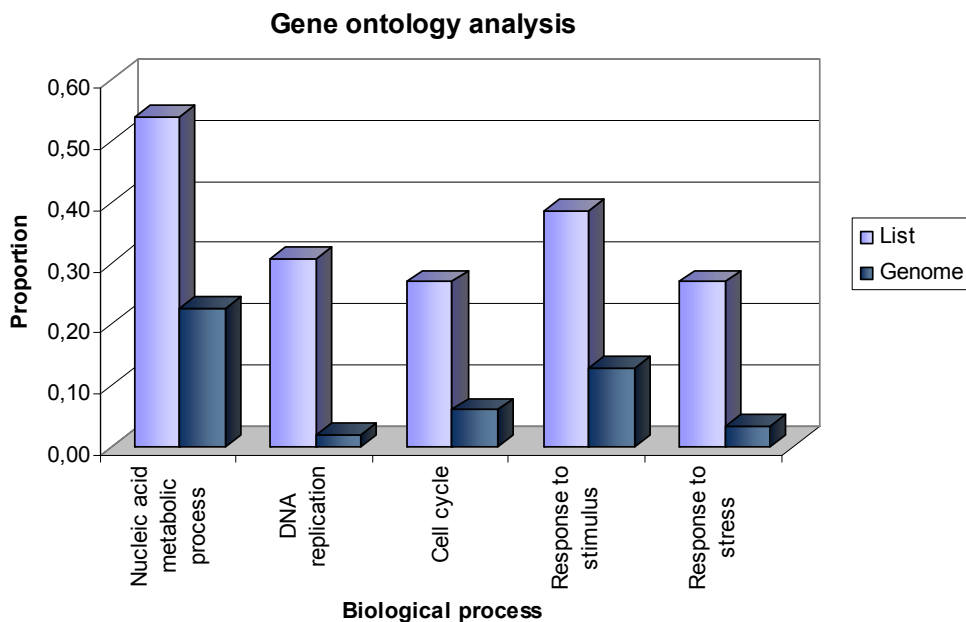






FIGURE 22 | Gene ontology analysis of the genes that are differentially expressed when *CG13617* is silenced. The proportion of genes classified in the different functional categories given below is represented both for the list of differentially expressed genes (light blue) and the whole set of genes belonging to that category in the *D. melanogaster* genome (dark blue). Percentages are calculated based on the number of genes with associated gene ontology categories, which is 26 for the list of differentially expressed genes obtained in this work and 8472 for the whole genome (see list total and genome total columns in TABLE 7). All the represented Gene Ontology terms refer to biological process categories that present statistically significant differences between the observed proportion in the list and what would be expected if we take into consideration the whole genome (the corresponding *P*-values are listed in TABLE 7).

TABLE 7 | Significantly enriched functional categories in the list of genes that are differentially expressed when *CG13617* is silenced. Lines shaded in blue correspond to gene ontology categories included within the previous category but that show significantly different proportions in the list of differentially expressed genes when compared to the whole genome. The GO number column displays the identification number of each Gene Ontology term in the Gene Ontology database . The list count and list total columns indicate the number of genes in the differentially expressed gene list included in a particular category, and the total number of genes in our list with an assigned biological process or molecular function category, respectively. The following column shows the corresponding percentage calculated based on the number of genes in the list with associated gene ontology terms to describe their function. The next three columns correspond to the same counts considering the whole *D. melanogaster* genome. A graphic comparison of the proportions found in the list and in the genome for each functional category is represented in FIGURE 22. Genes highlighted in red are genes included only in that particular category within the biological process classification.

Gene ontology term		GO number	List count	List total	List %	Genome hits	Genome total	Genome %	Fold enrichment	P-value	Genes
Biological process	Nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	GO:0006139	14	26	53.85	1912	8472	22.57	2.39	0.001230167	<i>CG5602, CycE, Mcm2, Mcm5, Mcm7, mus209, RnrL, RnrS, Art1, DebB, dnk, dUTPase, pont, Ts</i>
	 DNA replication	GO:0006260	8	26	30.77	159	8472	1.88	16.39	0.000000260	<i>CG5602, CycE, Mcm2, Mcm5, Mcm7, mus209, RnrL, RnrS</i>
	Cell cycle	GO:0007049	7	26	26.92	507	8472	5.98	4.50	0.002971396	<i>CG5602, CycE, Mcm2, mus209, Hsp83, pip, RanCap</i>
	Response to stimulus	GO:0050896	10	26	38.46	1084	8472	12.80	3.01	0.002669078	<i>CG5602, Hsp83, mus209, pont, Hsp70Ab, Hsp70Bbb, Hsp70Bc, Est-6, gp210, Ohp56a</i>
	 Response to stress	GO:0006950	7	26	26.92	279	8472	3.29	8.18	0.000126178	<i>CG5602, Hsp83, mus209, pont, Hsp70Ab, Hsp70Bbb, Hsp70Bc</i>
Molecular function	Nucleotide binding	GO:0000166	12	32	37.50	1187	11085	10.71	3.50	0.000225740	<i>CG5602, CG10514, dnk, Hsp70Ab, Hsp70Bbb, Hsp70Bc, Hsp83, Mcm2, Mcm5, Mcm7, pont, RnrL</i>
	 ATP binding	GO:0005524	12	32	37.50	796	11085	7.18	5.22	0.000005453	<i>CG5602, CG10514, dnk, Hsp70Ab, Hsp70Bbb, Hsp70Bc, Hsp83, Mcm2, Mcm5, Mcm7, pont, RnrL</i>

When we consider the molecular function (instead of the biological process) of the proteins included in the differentially expressed genes, we also find statistically significant gene ontology categories. In this case there is an excess of ATP-binding proteins ($P = 5.45 \cdot 10^{-6}$) and nucleotide-binding proteins ($P = 2.26 \cdot 10^{-4}$) (TABLE 7). This is due to 12 genes (37.50%) that encode proteins able to bind both substrates.

Some of these differentially expressed genes carry out important functions inside the cell. In fact, for some genes, certain mutant alleles (such as deletions that eliminate the gene completely) present lethal phenotypes, which suggest that their functions are essential for survival. For example, two proteins that are critical for the individual to be viable are the products of genes *mus209*, a cofactor that directly binds DNA polymerases δ and ϵ and affects their progress during the DNA strand synthesis (FIGURE 23), and *CycE*, which encodes a cyclin involved in the transition from G1 to S phase of the cell cycle. Some of the

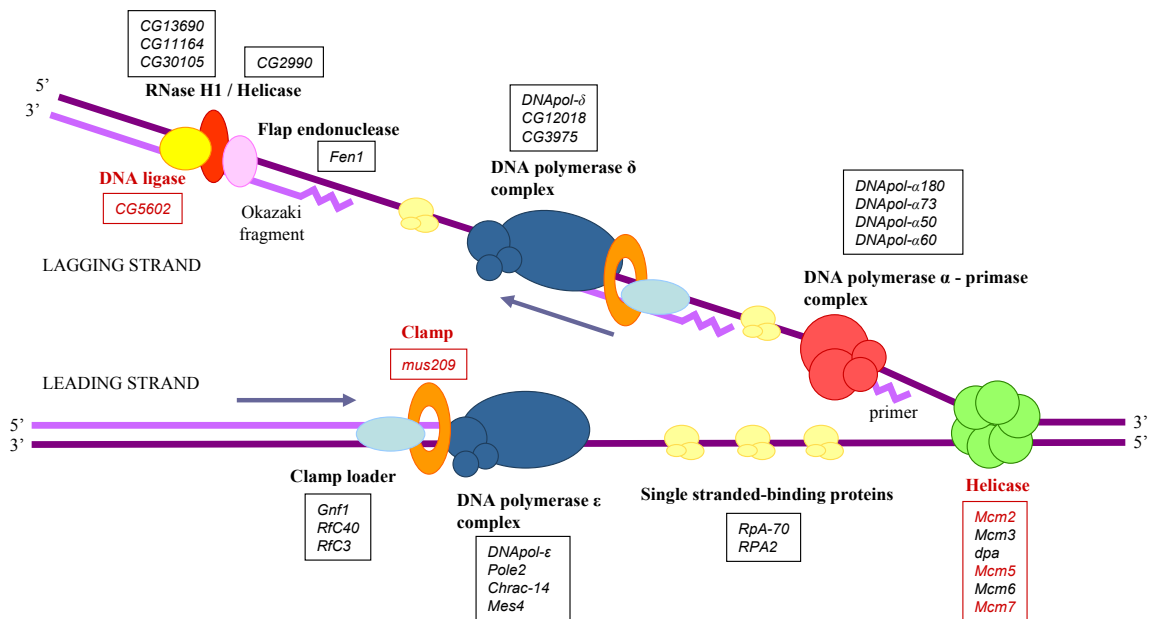


FIGURE 23 | Diagram showing the DNA replication complex in *D. melanogaster*. Enclosed in rectangles are the *D. melanogaster* genes that encode the proteins performing the distinct functions of the core DNA replication machinery specified in the figure according to the KEGG PATHWAY database . The gene names highlighted in red correspond to differentially expressed genes when first instar larvae with and without *CG13617* expression are compared using microarrays. An additional gene that is part of the clamp loader, *RfC40*, also shows a reduced expression if the fold-change cutoff is decreased to 1.5.

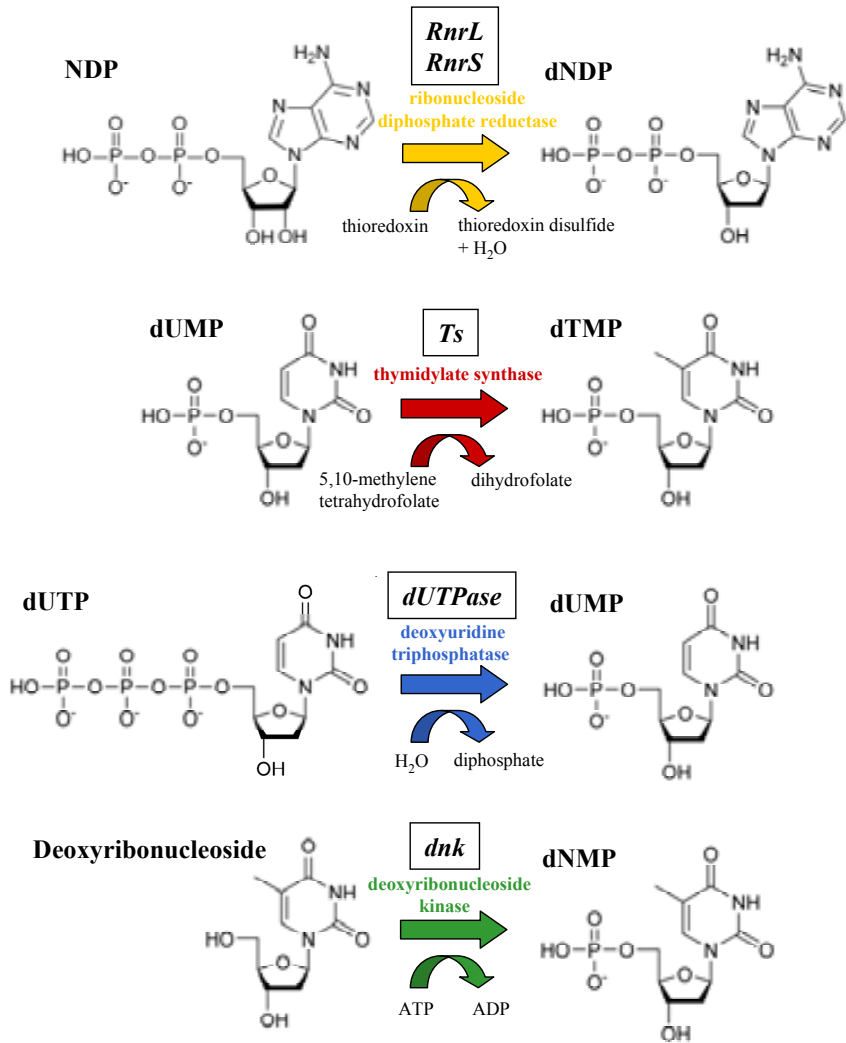


FIGURE 24 | Differentially expressed genes after *CG13617* silencing involved in nucleotide metabolism in *D. melanogaster*. The name of each gene is enclosed in a rectangle above the reaction catalyzed by the enzymes they encode. The complete enzyme names are shown in color below the gene name. NDP, ribonucleotide diphosphate; dNDP, deoxyribonucleotide diphosphate; dUMP, deoxyuridine monophosphate; dTMP, deoxythymine monophosphate; dUTP, deoxyuridine triphosphate; dNMP, deoxyribonucleotide monophosphate.

differentially expressed genes for which functional information is available are shown in FIGURES 23 and 24, which try to illustrate in more detail their specific functions. It is interesting that among the differentially expressed genes there are some proteins that physically interact inside the cell to perform a determined function. This is the case of *RnrL* and *RnrS* that encode respectively the large and small subunits of the ribonucleotide reductase, the enzyme required to synthesize the deoxyribonucleotides from the corresponding ribonucleotides (FIGURE 24), or the MCM proteins 2, 5 and 7 that form an hexamer together

with dpa and MCM 3 and 6 proteins that acts as the helicase that unwinds the two DNA strands to allow the DNA polymerase to copy the DNA molecule during replication (FIGURE 23). This suggests that these proteins that work together also share similar regulation mechanisms so that they can be expressed under the same circumstances and supports that the observed expression differences are caused by the experimental conditions and are not artifacts.

Another interesting group are four genes (*CG8087*, *CG13135*, *CG32198* and *CG14850*) that exhibit the greatest fold changes in their expression levels, with a ~4-14 fold reduction (depending on the gene and the analysis considered) in the samples where *CG13617* expression has been silenced. Even though the average amino acid identity between the proteins encoded by the four differentially expressed members is low (35.38%) and they appear to be distantly related, these genes have been described to belong to a protein family of 31 members in the *D. melanogaster* genome identified by the PANTHER (Protein ANALysis THrough Evolutionary Relationships) Classification System [🔗](#) as structural proteins involved in the maintenance of cellular structure (PTHR23246, New-glycine proteins). This putative family does not possess any known protein domains that could help in the determination of their molecular function inside the cell and their specific function remains unknown. In fact, in the FlyBase [🔗](#) and Gene Ontology [🔗](#) databases, these four genes do not have any gene ontology category assigned in their description, and therefore they could not be identified as a significant functional group in the previous computational analysis. With respect to their genomic localization, these genes are scattered through the whole genome, although two of the differentially expressed genes, *CG8087* and *CG14850*, are located in *D. melanogaster* in a 5.5 kb fragment of chromosome 3R that includes a group of four consecutive genes belonging to this family.

Finally, to ensure the validity of the microarray results, gene-expression changes were measured with an independent technique: real-time RT-PCR. Expression differences were considered to be validated when the direction of the change was the same in both methods. Gene-expression levels of eight genes involved in DNA replication and cell cycle (*mus209*, *Mcm2*, *Mcm5*, *Mcm7*, *RnrS*, *RnrL*, *CycE* y *Hsp83*) were measured in *D. melanogaster* using five independent samples of first instar larvae injected as controls and five injected with dsRNA

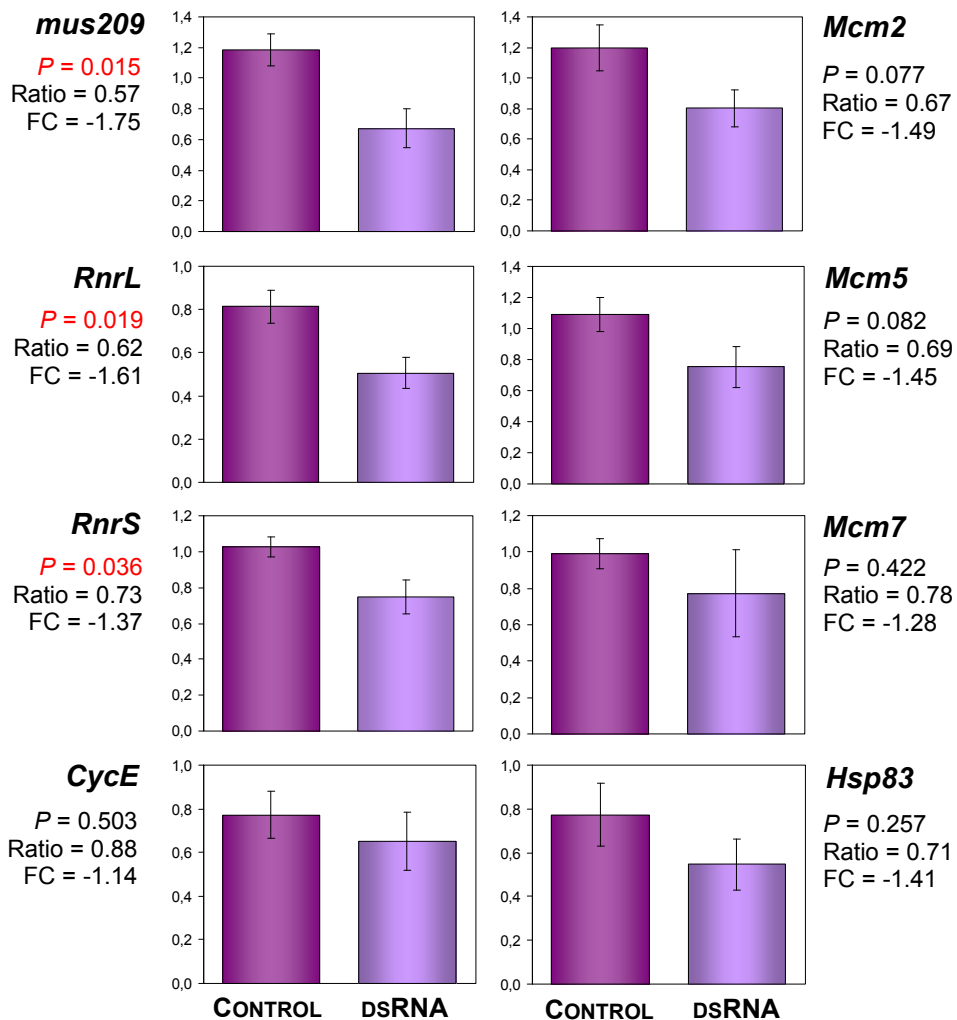


FIGURE 25 | Real-time RT-PCR results for the eight genes analyzed in *D. melanogaster*. For each gene (the name is indicated at the left or right of the graphs), the dark purple column corresponds to the average expression level measured in five different control samples and the light purple to the expression of five dsRNA-injected samples normalized to *Gapdh* expression. All expression values are shown with respect those of Control2 sample (one of the samples used in the microarray experiments), considered to be equal to one. The statistical probability to observe these differences according to the ANOVA analyses, the ratio of dsRNA/Control expression and the fold change values (FC) obtained in the real-time RT-PCR experiments are given below each gene name. The *P*-values of statistically significant differences between arrangements are highlighted in red. Error bars represent the standard error.

(for each group of samples, two were the same that had been hybridized to the microarrays and three were independent replicates). Differences in expression levels between CONTROL and DSRNA samples were found in almost all the analyzed genes. More importantly, these differences were always in the expected direction, that is, samples without *CG13617* expression show a reduction of the expression level of these genes with respect to samples

that express *CG13617* normally (FIGURE 25). The probability of observing a decreased expression in the samples where *CG13617* is silenced for eight out of eight genes tested is statistically significant ($P = 0.0039$) according to the sign test, and a *t*-test for paired comparisons also indicates that the expression levels of the analyzed genes are significantly lower in *CG13617*-depleted samples with respect to controls ($P = 1.93 \cdot 10^{-4}$) (SOKAL and ROHLF 2000). However, fold changes obtained by real-time RT-PCR were lower than those of microarrays with just a 12-43% reduction in expression levels in DSRNA samples, which corresponds to CONTROL samples having on average a level of expression between 1.14 and 1.75 times higher than that of samples where *CG13617* is silenced. Actually, only three genes (*mus209*, *RnrL* and *RnrS*) show statistically significant differences ($P < 0.05$) between silenced and control samples, with two more genes (*Mcm2* and *Mcm5*) presenting marginally significant differences ($P < 0.1$).

3.2.3 Molecular consequences of *CG13617* silencing in *D. buzzatii* 2j lines

Given that several genes experiment changes in their expression levels (either directly or indirectly) when *CG13617* expression is eliminated in *D. melanogaster*, next we tested whether these same genes also show a reduced expression level in *D. buzzatii* 2j embryos, where *CG13617* expression is naturally reduced five times with respect to 2st embryos due to the presence of an antisense RNA that overlaps the whole gene (PUIG *et al.* 2004). Therefore, we compared by real-time RT-PCR in *D. buzzatii* 2st and 2j samples the expression levels of some of the differentially expressed genes found in the *D. melanogaster* microarray analysis. In particular, we decided to study in *D. buzzatii* 10 genes involved in DNA replication (*mus209*, *Mcm2*, *Mcm5*, *Mcm7*, *RnrL*, *RnrS*, *Ts*, *CycE*, *RanGap* and *Hsp83*) and one gene (*CG8087*) that belongs to the putative protein family that presents the highest fold changes in *D. melanogaster*.

First, based on the homologous gene sequences in *D. mojavensis* and *D. virilis*, we designed primers to amplify and sequence fragments of the selected genes in *D. buzzatii* (more details are given in the Materials and Methods section). A list of the sequenced segments of each gene together with information about the coding regions of these genes in *D. melanogaster* and *D. mojavensis* is shown in TABLE 8. The gene fragments were amplified using as a template

either DNA or cDNA (TABLE 8), depending on the distance separating the two interspecific primers and the length of the introns. Our goal was to amplify fragments of ~800 bp, which can be sequenced completely with two reactions (one from each end). When cDNA was used as a template, DNA was always included in the PCR as a positive control (even though the expected product size is larger). All genes were sequenced in the st-1 line using genomic DNA or embryos cDNA as template. Line j-19 genomic DNA or embryos cDNA was also used as a control to verify that the same PCR product is obtained in more than one *D. buzzatii* line. The 10 DNA replication genes could be amplified and sequenced in *D. buzzatii* without any problems. *CG8087*, however, was not easy to identify in the *D. mojavensis* and *D. virilis* genomes due to variation in the copy number of genes of the same family at that particular chromosomal location. Given the observed variability in the number of gene copies and the fact that this gene belongs to a gene family dispersed across the whole genome, this specific region was located in the different genomes searching for the flanking non-repeated

TABLE 8 | Genes partially sequenced in *D. buzzatii* prior to the expression analyses by real-time RT-PCR. For the *D. buzzatii* (Dbuz) sequences, the DNA or mRNA columns indicate the template used to amplify the gene by PCR and the length of the resulting sequenced fragment. The parts of the gene that were included in the amplified fragment are listed as well (Ex, exon; In, intron). For *D. melanogaster* (Dmel, where all these genes were initially characterized) and *D. mojavensis* (Dmoj, the sequenced species phylogenetically closest to *D. buzzatii* and where the interspecific primers were designed), the length of the coding region of each gene is shown in the mRNA column (5' and 3' UTRs are not included even though data are available for some of the *D. melanogaster* genes). The DNA column indicates the total length of the gene including introns. If more than one alternative transcript is produced, the different coding regions are indicated in the mRNA column, as well as in the total number of exons for each gene. All genes were sequenced in *D. buzzatii* line st-1 except for *CG8087*, which could only be amplified in line j-13. All lengths are expressed in base pairs.

Gene	Dbuz sequences			Dmel coding region			Dmoj coding region		
	DNA	mRNA	Gene segment	DNA	mRNA	Total exons	DNA	mRNA	Total exons
<i>mus209</i>	718		Ex1-In1-Ex2	843	783	2	843	783	2
<i>Mcm2</i>		774	Ex3-Ex4	2721	2664	2	2838	2655	4
<i>Mcm5</i>		750	Ex3-Ex4	2377	2202	4	2402	2205	4
<i>Mcm7</i>		926	Ex4-Ex5	2427	2163	5	2405	2163	5
<i>RnrL</i>	1003		Ex4 (partial)	2780	2439	4	15633	2457	4
<i>RnrS</i>		703	Ex2-Ex3	1838	1182	3	2473	1182	3
<i>Ts</i>	729		Ex1 (partial)	966	966	1	933	933	1
<i>CycE</i>		765	Ex2-Ex3	11911	2130/1809	5/4*	2907	1944	4
<i>RanGap</i>		555	Ex3-Ex4-Ex5	2053	1791	5*	2332	1794	5
<i>Hsp83</i>	821		Ex1 (partial)	2154	2154	1*	2154	2154	1
<i>CG8087</i>	365		Ex1 (partial)	429	429	1	444	444	1

* Non-coding 5' exons not included.

orthologous genes (*Cys* and *CG14854*), and not based on where the putative *CG8087* was initially annotated in these species. In the *D. mojavensis* genome only one copy of a gene similar to *CG8087* was found to be present between genes *Cys* and *CG14854*, dmoj_GLEANR_10573 (GI10665), which was identified as the orthologous gene of *D. melanogaster* *CG8087*. In *D. virilis* three copies were detected in the syntenic region (although in this species a gene located further downstream had been initially annotated as *CG8087*). Of those three copies, the gene immediately upstream of *CG14854* was the first hit obtained in the *D. virilis* genome in BLASTP searches using *D. melanogaster* and *D. mojavensis* proteins as query, so dvir_GLEANR_9512 (GJ24218) was identified as *CG8087* ortholog in this species based on its position in the genome and this similarity. Interspecific primers were designed in *D. mojavensis* putative *CG8087* gene sequence, but a PCR product of the expected size could only be amplified from *D. buzzatii* line j-13 genomic DNA and not using any of the other lines' DNA as templates for the PCR reaction. This j-13 fragment was sequenced and it yielded as best hits in BLAST searches the expected homologous genes in both *D. mojavensis* and *D. virilis* genomes.

Once the *D. buzzatii* sequences were available, specific primers for real-time RT-PCR were designed for each gene. Real-time RT-PCR experiments were performed comparing four *2st* (st-1, st-11, st-13, and st-14) and four *2j* (j-2, j-9, j-13, and j-19) lines with different geographical origins (TABLE 4, see Materials and Methods). For each line, two samples were included: embryos 0-20 h old and first instar larvae. Embryos were studied because this is the developmental stage where the antisense RNA and *CG13617* silencing were described, while first instar larvae were analyzed because this was the stage where the differences in gene expression were initially detected in the *D. melanogaster* experiments using microarrays. Expression levels were finally assessed in 10 out of the 11 genes partially sequenced in *D. buzzatii*, because the specific primers designed on *CG8087* sequence failed to amplify any transcript in the real-time RT-PCR tests performed, and measurements of gene expression could not be finally carried out for this particular gene.

Differences in the average expression levels between *2st* and *2j* lines ranging between 1.33- and 1.92-fold were found in 9 out of 10 analyzed genes in the embryo samples (FIGURE 26a), but none of these genes showed differences in the first instar larvae samples (FIGURE

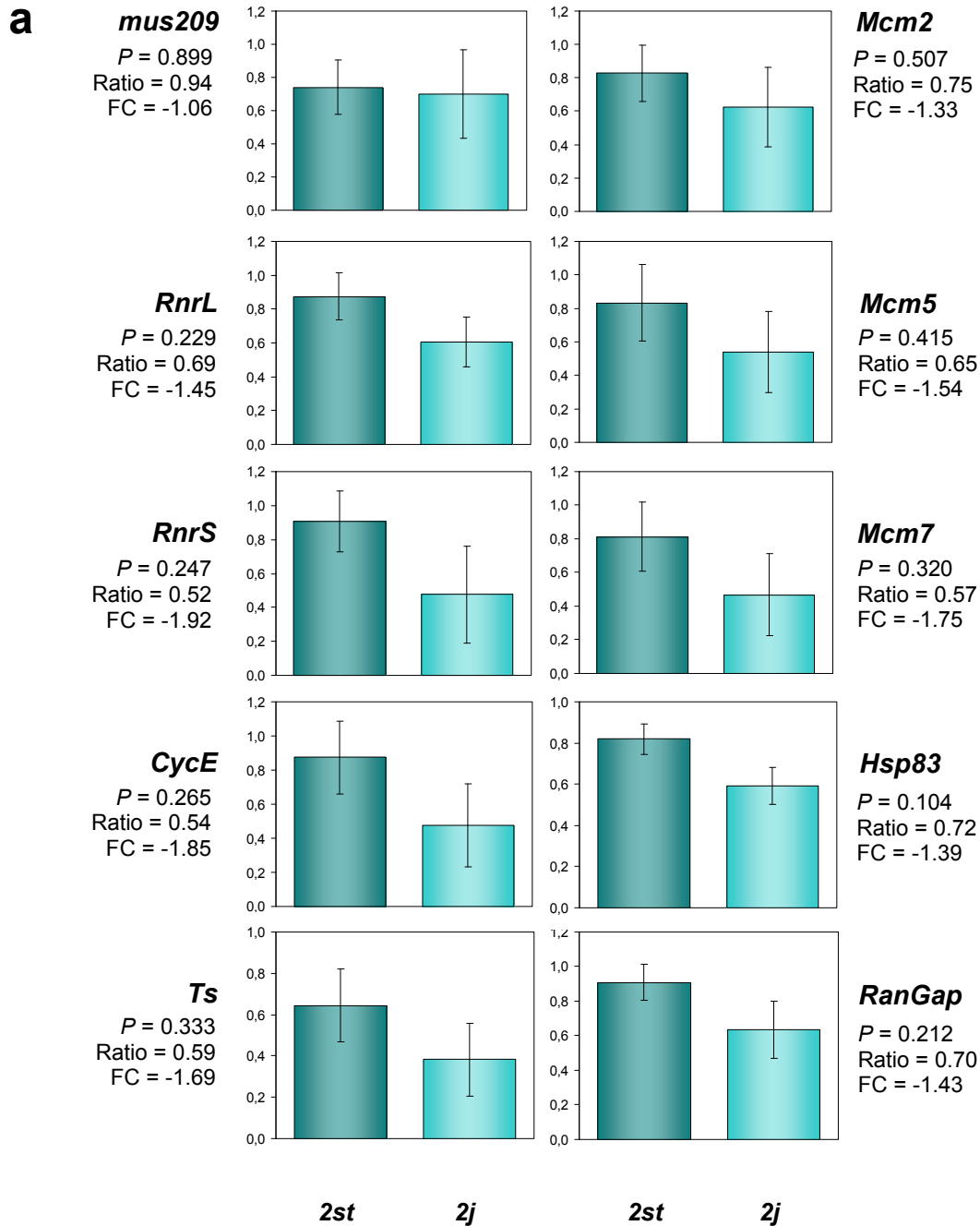
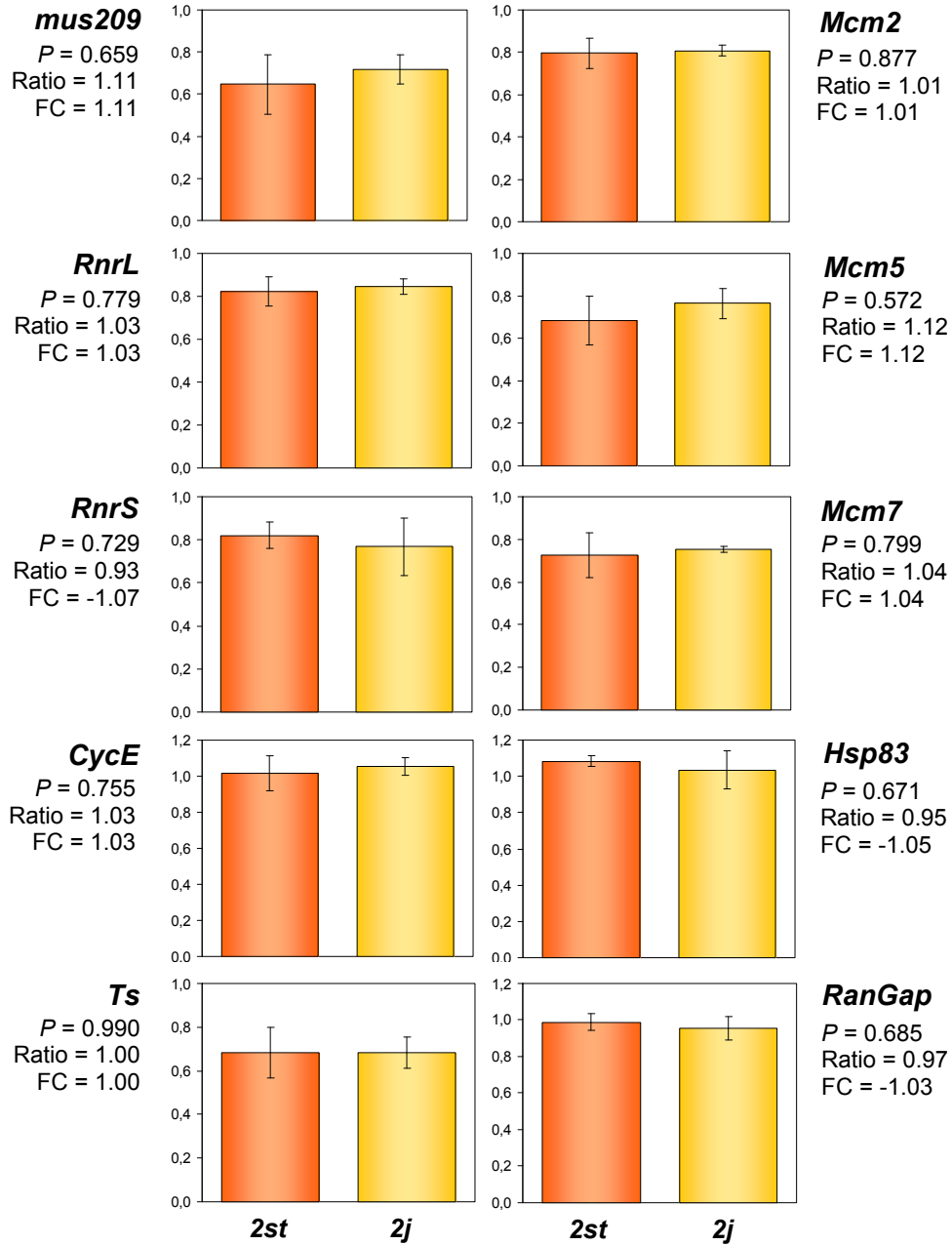
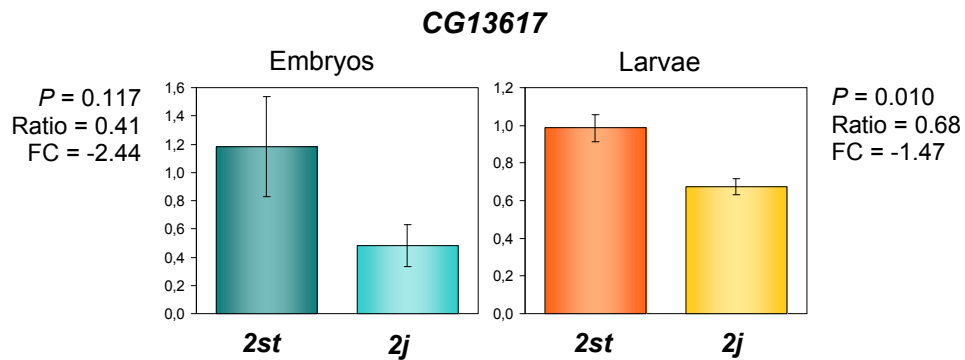


FIGURE 26 | Real-time RT-PCR results for the ten genes analyzed in *D. buzzatii*. (a) Average expression levels for each gene in embryos. Dark blue corresponds to the expression in *2st* lines and light blue to *2j* lines. (b) Expression levels for the same genes in first instar larvae. Orange indicates expression in *2st* lines and yellow in *2j* lines. (c) Expression of gene *CG13617* in the same *2st* and *2j* lines. Color code is the same as in the other parts of the figure. All expression values are normalized with respect to *Gapdh* to control for differences in cDNA concentration and to line *st-1* expression values, considered to be equal to one in all the cases. For each gene, the statistical probability to observe these differences is shown along with the *2j/2st* expression ratio and fold change (FC) expressed in the same terms used in the microarray results. Error bars represent the standard error.

b**c**

26b) where all $2j/2st$ expression ratios are very close to 1 (in some cases expression level is even slightly higher in $2j$ lines). Interestingly, the detected expression differences were always in the same direction and correspond to a reduction of the expression level in $2j$ embryos with respect to $2st$ ones, which agrees with our previous observation that the genes involved in DNA replication decreased their expression level when *CG13617* was silenced in *D. melanogaster*. Although the fold changes between $2st$ and $2j$ embryos are not especially high and the observed expression changes are not statistically significant in any case, there is a clear tendency for these genes to present a lower expression level in embryos of $2j$ lines. In fact, a sign test (SOKAL and ROHLF 2000) reveals that the probability of observing by chance a decreased expression in $2j$ lines in 9 out of 10 genes tested is only 0.0107, and according to a t -test for paired comparisons (SOKAL and ROHLF 2000) the differences between $2st$ and $2j$ expression values are significantly different from 0 in embryos ($P = 2.29 \cdot 10^{-5}$) but not in larvae ($P = 0.4373$). In part, the lack of statistical significance appears to be caused by a high variation in the expression levels inside each arrangement, which is largely due to the presence of a single $2st$ line with a low expression level in all the analyzed genes, combined with one $2j$ line that systematically exhibits expression levels much higher than the other lines carrying the inversion.

As additional confirmation, we also examined *CG13617* expression levels in embryos and first instar larvae (FIGURE 26c) because neither the *D. buzzatii* lines used in this part of the work nor the developmental stages analyzed are exactly the same than those used when the silencing of this gene was first detected. Therefore, it is not surprising that the results are slightly different. In this case, *CG13617* expression presents again a clear reduction in $2j$ embryos, but the expression level is only 2.5 times lower in these samples with respect to $2st$ ones (*vs.* the 5-fold reduction found previously in 0-12 h old embryos). The difference is not statistically significant due to line st-13, which shows an unusually low *CG13617* expression level for a $2st$ line (if this line is excluded from the analysis, expression level becomes 3 times higher in $2st$ lines). In first instar larvae, $2j$ lines still present a decreased *CG13617* expression compared to $2st$ lines, but in this case the reduction is smaller, with the average expression level being 1.47 times higher in $2st$ larvae.

3.3 Evolution and function of gene *CG13617*: comparative sequence analysis

In order to obtain information about the evolution and function of gene *CG13617*, we carried out an exhaustive comparative sequence analysis both inside the genus *Drosophila* and in other organisms. The analysis of *CG13617* sequence in the different available *Drosophila* genomes includes a comparison of the exonic structure of *CG13617* coding region among the different species, the identification of several shared protein motifs within the *CG13617* protein, as well as studies about the level of conservation of the distinct parts of the protein. In addition, similarity searches were performed to detect in other types of organisms proteins similar to *CG13617* which may have been studied and could shed some light on *CG13617* function. Finally, the non-coding sequences around gene *CG13617* were also analyzed in a subset of the *Drosophila* genomes to try to find conserved elements that could correspond to promoter or regulatory regions controlling the expression of this gene.

3.3.1 *CG13617* genomic structure in *Drosophila* species

Gene *CG13617* has been identified in all 12 sequenced *Drosophila* species, as well as in *D. martensis* and *D. buzzatii* (TABLE 9), both closely related species belonging to the *repleta* group inside the *Drosophila* subgenus. This gene is located in a conserved syntenic block between genes *nAcR β -96A* and *Pp1 α -96A* in all the sequenced genomes (except in the inverted *2j* chromosomes in *D. buzzatii*), although the distances between these two genes and *CG13617* vary greatly among different species. However, even though the gene can be easily identified through homology searches in the different genomes, comparison of *CG13617* gene annotations in all these species reveals that gene structure does not appear to be completely conserved among them (FIGURE 27). At least two events of intron gain or loss have occurred since the divergence of these 14 *Drosophila* species from their common ancestor. Gene *CG13617* has five exons and four introns in all the *Sophophora* subgenus species, except for *D. pseudoobscura* and *D. persimilis* that seem to have lost intron 3 and, as a consequence, former exons 3 and 4 have become one (FIGURE 27 and TABLE 10). Conversely, in the *Drosophila* subgenus, all species show a gene structure made up of four exons separated by three introns.

TABLE 9 | *CG13617* orthologous genes in *Drosophila* species. E-value was 0 for all of them in BLASTP similarity searches performed using *D. buzzatii* *CG13617* protein as a query. The total length of the gene corresponds to the sum of the coding regions and intronic sequences. UTRs are not included because they have not been determined for all the species. The length in amino acids (aa) and molecular weight in daltons (Da) of the proteins predicted to be encoded by these genes are indicated as well.

Gene name	GLEANR identification	Total (bp)	Coding region (bp)	Protein (aa)	Molecular weight (Da)
<i>Dbuz</i> \CG13617	-	2404	2205	734	83019.72
<i>Dmar</i> \CG13617	-	2408	2205	734	83161.98
<i>Dmoj</i> \GI10083	dmoj_GLEANR_10040	2409	2184	727	82441.34
<i>Dvir</i> \GJ23818	dvir_GLEANR_9145	2429	2253	750	84883.96
<i>Dgri</i> \GH18825	dgri_GLEANR_3167	2563	2268	755	85629.73
<i>Dwil</i> \GK22342	dwil_GLEANR_5559	2477	2241	746	85316.47
<i>Dpse</i> \GA12410	dpse_GLEANR_3576	2423	2238	745	84348.89
<i>Dper</i> \GL21868	dper_GLEANR_3943	2423	2238	745	84065.59
<i>Dana</i> \GF20753	dana_GLEANR_3998	2441	2217	738	83872.50
<i>Dere</i> \GG11297	dere_GLEANR_11407	2444	2205	734	83602.43
<i>Dyak</i> \GE23492	dyak_GLEANR_7274	2439	2211	736	83676.51
<i>Dmel</i> \CG13617	-	2443	2214	737	83822.40
<i>Dsim</i> \GD21105	dsim_GLEANR_4867	2448	2214	737	83948.61
<i>Dsec</i> \GM26604	dsec_GLEANR_9471	2448	2214	737	83890.53

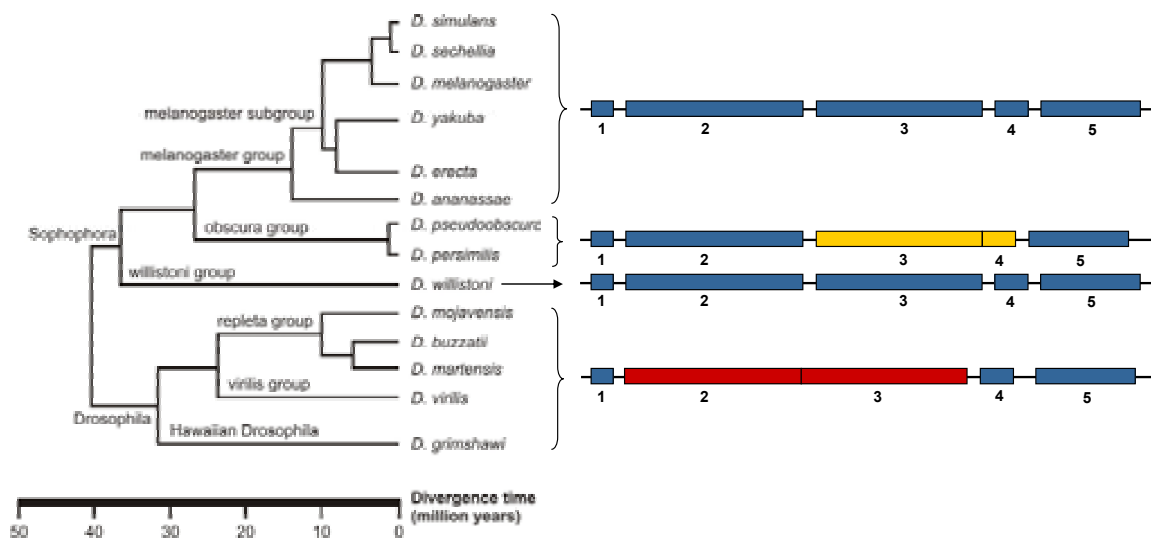


FIGURE 27 | *CG13617* gene structure in different *Drosophila* species. A phylogenetic tree of the 12 sequenced *Drosophila* genomes together with *D. buzzatii* and *D. martensis* is shown on the left. On the right a diagram of the corresponding gene exonic structures is depicted for each group of species. Each rectangle represents an exon, which are numbered below following *D. melanogaster* exon order. Those exons involved in intron loss or gain are highlighted in yellow and red.

In these species, exons 2 and 3 appear fused in one large exon (FIGURE 27 and TABLE 10). This means that intron 2 was either gained in the *Sophophora* subgenus or lost in *Drosophila*, depending on which was the original structure of the gene.


Nucleotide identities for *CG13617* among different species have not been calculated because *Drosophila* sequences are too divergent at nucleotide level to be properly aligned, even considering only the coding regions. In fact, although all these sequences belong to the same genus, the two subgenera started to diverge from one another 40 million years ago, and this great phylogenetic distance makes protein alignment a much more reliable tool to draw firm conclusions.

TABLE 10 | Structure of gene *CG13617* in the different *Drosophila* species. The length of each intron and exon is expressed in base pairs. In those cases where introns have been lost or gained, a larger cell in the table spanning the absent intron as well as the flanking exons indicates the structural change. For these species, the numeration of introns and exons is different than the one in the first column. Cells shaded in yellow correspond to the coding sequences. UTRs were determined experimentally only in *D. buzzatii* (PUIG *et al.* 2004). In the rest of the species they have been predicted by multiple sequence alignment. The length of the putative UTRs has been included only in those species where conservation was high enough to make a reliable prediction.

	Subgenus <i>Sophophora</i>									Subgenus <i>Drosophila</i>				
	Dsim	Dsec	Dmel	Dyak	Dere	Dana	Dpse	Dper	Dwil	Dmoj	Dbuz	Dmar	Dvir	Dgri
5' UTR	-	-	-	-	-	-	-	-	-	105	118	118	136	115
Exon 1	94	94	94	94	94	94	94	94	103	94	94	94	94	94
Intron 1	59	58	58	53	64	58	57	57	57	60	60	60	60	64
Exon 2	785	785	785	785	785	794	779	779	815	1495	1513	1516	1546	1567
Intron 2	57	58	58	59	59	52	64	64	62					
Exon 3	731	731	731	728	731	728	918	918	716					
Intron 3	60	60	55	55	55	62			63	56	56	57	60	63
Exon 4	157	157	157	157	157	157	918	918	154	157	157	157	166	166
Intron 4	58	58	58	61	61	52			64	64	54	109	83	86
Exon 5	447	447	447	447	438	444	447	447	453	438	441	438	447	441
3' UTR	-	-	-	-	-	-	-	-	-	14	14	14	-	-
Total	2448	2448	2443	2439	2444	2441	2423	2423	2477	2409	2404	2408	2429	2563

3.3.2 Sequence analysis of the CG13617 protein in *Drosophila* species and other organisms

To get more insight into the function and evolution of the CG13617 gene we have done an exhaustive computational analysis of the encoded protein in different organisms. The increased number of available sequences has greatly enhanced the analysis carried out initially in PUIG *et al.* (2004). Gene *CG13617* encodes a 734 amino acid (aa) protein in *D. buzzatii*, although the length of the protein varies in *Drosophila* species between 727 aa in *D. mojavensis* and 755 aa in *D. grimshawi* (TABLE 9). All the available *Drosophila* protein sequences were aligned (FIGURE 28) and their overall pairwise identity and similarity are shown in TABLE 11. The protein identity values (identical amino acids shared by two different sequences) fluctuate between the 57.4% identity between *D. grimshawi* and *D. willistoni* and the 98.6% identity shown by the two *D. buzzatii* sequences (*2st* and *2j*), with the 97.7% identity between *D. simulans* and *D. sechellia* being the highest value between sequences belonging to two different species. Similarity values include those amino acid changes that preserve the physico-chemical properties of the original residue, and they range between the 73.4% similarity between *D. grimshawi* and *D. willistoni* and the 99.5% between the two *D. buzzatii* chromosomal arrangements, with the comparison between *D. simulans* and *D. sechellia* showing the highest similarity value between different species with a 98.6%. However, it has to be taken into account that the presence of functional domains and other motifs in the protein sequence leads to variation in the conservation level of different parts of the sequence, which requires a more careful examination.


A total of five different putative functional motifs could be predicted in the *D. buzzatii* CG13617 protein using different bioinformatic approaches. The main functional domain is a C2H2-type zinc finger (INTERPRO  database accession number IPR007087) found in positions 149-170 of the alignment (FIGURE 28). This domain can be detected in all the *Drosophila* CG13617 protein sequences and presents a high level of conservation: 19 out of 22 amino acids (86.36% identity) are identical in the 14 analyzed species, and the remaining three changes are conservative (substitution of an amino acid by another one with similar properties). Two of these changes are each specific for a different species and the third is shared by six species distributed across both subgenera. The four key cysteine (C) and

histidine (H) residues that hold the zinc atom in place are strictly conserved in all the different species.

Other potential functional domains are three coiled-coil regions. Coiled coils are supercoiled structures encoded by a seven-residue repeat denoted $[abcdefg]_n$ that typically has hydrophobic residues at positions *a* and *d* and polar/charged residues at *e* and *g*. Coiled coil regions are thought to be involved in the interaction between proteins. In particular, the three coiled-coil regions of CG13617 are located at positions 100-140, 200-230 and 380-400 of the alignment (FIGURE 28). Positions are approximate because slightly different ends have been predicted for these primary structure motifs in each analyzed protein sequence. A fourth coiled-coil region (previously described in PUIG *et al.* 2004) in positions 305-330 of the alignment has also been detected in *D. willistoni*, *D. virilis*, *D. mojavensis*, *D. buzzatii* and possibly *D. martensis* (the prediction is not so reliable for this last species but the presence of the coiled coil in closely related species together with the conservation of the sequence, suggest that the protein probably adopts this structural conformation too) but not in the rest of species. Finally, the *D. mojavensis* sequence could include a specific fifth coiled-coil region in positions 260-290 of the alignment. The level of amino acid conservation of these motifs varies considerably among coiled coil segments. While the first two coiled coils show a high level of conservation, with 59.37% (19/32) and 83.33% (25/30) amino acid identity among all species, respectively, in the third coiled coil region this identity is reduced to a 34.61%, with only 9 identical amino acids out of 26 residues. These identities have been calculated using only the sequence identified as coiled coil shared by all the species. It has to be taken into account that, according to the software used in the analysis, the third coiled coil is not as well supported as the first two in some species, but the fact that a coiled coil prediction could be obtained (at least when certain parameters were used) in the exact same region where the majority of the proteins get a reliable prediction suggests that this could be a real motif in all the species.

TABLE 11 | CG13617 protein identity and similarity matrix in *Drosophila* species. Proportion of identical amino acids (white) and similar amino acids (blue) found among the different proteins. The highest and lower values for both identity and similarity are highlighted in yellow and red, respectively. Calculations are based on pairwise alignments and were made with the MATGAT software using the BLOSUM62 matrix.

	Dbuz <i>2st</i>	Dbuz <i>2j</i>	Dmar	Dmoj	Dvir	Dgri	Dwil	Dper	Dpse	Dana	Dere	Dyak	Dmel	Dsec	Dsim
Dbuz <i>2st</i>		98.6	93.9	87.0	74.3	71.3	58.6	60.6	60.4	61.6	60.2	61.8	59.7	60.5	59.8
Dbuz <i>2j</i>	99.5		93.3	87.2	73.9	71.4	58.5	60.1	60.0	62.0	60.7	62.2	60.2	60.9	60.9
Dmar	97.0	96.5		86.9	74.2	71.7	58.9	61.3	61.1	62.0	60.9	62.3	60.5	60.6	60.1
Dmoj	92.6	92.6	92.5		75.5	70.5	58.3	60.0	59.7	61.1	60.9	61.7	59.9	60.3	60.2
Dvir	84.7	84.5	84.3	83.9		73.2	58.8	60.6	61.3	64.0	61.7	61.5	61.2	61.4	61.4
Dgri	82.8	82.5	82.4	81.7	85.0		57.4	60.6	60.5	63.0	60.6	62.0	60.4	60.6	60.7
Dwil	75.1	75.3	74.4	75.1	73.9	73.4		60.9	60.7	62.4	62.6	63.0	62.4	61.3	61.7
Dper	75.2	75.3	74.8	74.4	74.0	74.4	75.5		97.6	68.3	68.5	69.4	68.1	68.1	68.2
Dpse	75.3	75.3	74.8	73.4	74.8	74.4	75.6	98.3		68.8	69.1	69.9	68.4	68.4	68.5
Dana	77.9	78.2	77.6	76.4	76.3	76.6	77.3	81.5	81.7		77.7	77.7	77.9	76.8	77.4
Dere	75.9	76.3	74.9	76.0	75.1	75.0	78.3	80.3	80.9	87.3		93.6	91.6	90.6	90.6
Dyak	77.4	77.9	76.4	76.6	75.5	76.2	78.6	80.7	80.8	87.9	96.3		92.9	92.5	92.7
Dmel	75.2	75.6	75.0	75.2	74.1	75.8	78.6	80.5	80.4	87.7	94.7	95.4		95.8	96.1
Dsec	76.4	76.7	75.2	76.9	75.5	75.6	77.9	80.9	80.8	87.4	94.4	95.4	96.7		97.7
Dsim	75.3	76.4	74.9	76.4	75.1	75.8	77.9	80.8	80.8	88.1	95.0	96.1	97.3	98.6	

FIGURE 28 | Alignment and conservation plot of CG13617 proteins of 14 different *Drosophila* species. The alignment was performed using MUSCLE  with some minor modifications according to the T-COFFEE alignment. This view was obtained with JALVIEW program (BLOSUM62 matrix, Conservation Threshold = 25). Amino acid residues are colored with different shades of blue depending on their level of conservation (the more intense, the more conservation). Below the alignment, a yellow-brown bar graphic indicates the level of conservation for each position in the alignment. The colored bars above the alignment mark the possible functional domains identified in these proteins. Red bars indicate the coiled-coil regions in the *D. buzzatii* sequence (as specified in the text, coiled-coil region predictions do not start or end in the exact same residue for all the sequences and, in fact, only the first two coiled coils and the last one are found in all the analyzed species). The dark green bar signals the position of the C2H2-type zinc finger and the purple bar near the C-terminal end corresponds to the putative NLS. The NES is represented with a light green bar overlapping the fourth coiled-coil region, and the positions of the two *D. buzzatii* PEST sequences are depicted with pink bars. Two small orange bars indicate some extra residues conserved in species not belonging to the *Drosophila* genus (see below). The *D. buzzatii 2j* sequence is not included in the alignment due to the high similarity to the *2st* sequence.

10 20 30 40 50 60 70 80 90 100 110 120 130

Dbaa1-734 1 MGYYKN -- NFPQVMREAGFKLRQYRDGP LDMROMGSYETERI LREONLEVVDDALQHLSEAPLGTMLETHI LDDGI AKYF IMSQYA IYLLCCRTYLDSEVDELREAEHSQOEIAKLRKLS LSESNNEWDLHKKI TOIE 137
Dbaa1-734 1 MGYYKN -- NFPQVMREAGFKLRQYRDGP LDMROMGSYETERI LREONLEVVDDALQHLSEAPLGTMLETHI LDDGI AKYF IMSQYA IYLLCCRTYLDSEVDELREAEHSQOEIAKLRKLS LSESNNEWDLHKKI TOIE 137
Dbaa1-727 1 MGYYKN -- NFPQVMREAGFKLRQYRDGP LDMROMGSYETERI LREONLEVVDDALQHLSEAPLGTMLETHI LDDGI AKYF IMSQYA IYLLCCRTYLDSEVDELREAEHSQOEIAKLRKLS LSESNNEWDLHKKI TOIE 137
Dbaa1-750 1 MGYYKN -- NFPQVMREAGFKLRQYRDGP LDMROMGSYETERI LREONLEVVDDALQHLSEAPLGTMLETHI LDDGI AKYF IMSQYA IYLLCCRTYLDSEVDELREAEHSQOEIAKLRKLS LSESNNEWDLHKKI TOIE 137
Dbaa1-755 1 MGYYKN -- NFPQVMREAGFKLRQYRDGP LDMROMGSYETERI LREONLEVVDDALQHLSEAPLGTMLETHI LDDGI AKYF IMSQYA IYLLCCRTYLDSEVDELREAEHSQOEIAKLRKLS LSESNNEWDLHKKI TOIE 137
Dbaa1-746 1 MGYYKN -- NFPQVMREAGFKLRQYRDGP LDMROMGSYETERI LREONLEVVDDALQHLSEAPLGTMLETHI LDDGI AKYF IMSQYA IYLLCCRTYLDSEVDELREAEHSQOEIAKLRKLS LSESNNEWDLHKKI TOIE 137
Dbaa1-745 1 MGYYKS -- NYPQVMREAGFKLRQYRDGP LDMRLMGSYETERI LREONFELVDKALQHLSEAPLGTVLETHI LDDGI AKYF IMSQYA IYLLCCRTYLDSEVDELREAEHSQOEIAKLRKLS LSESNNEWDLHKKI TOIE 137
Dbaa1-738 1 MGYYKS -- NYPQVMREAGFKLRQYRDGP LDMRLMGSYETERI LREONFELVDKALQHLSEAPLGTVLETHI LDDGI AKYF IMSQYA IYLLCCRTYLDSEVDELREAEHSQOEIAKLRKLS LSESNNEWDLHKKI TOIE 137
Dbaa1-734 1 MGYYKS -- NYPQVMREAGFKLRQYRDGP LDMRLMGSYETERI LREONFELVDKALQHLSEAPLGTVLETHI LDDGI AKYF IMSQYA IYLLCCRTYLDSEVDELREAEHSQOEIAKLRKLS LSESNNEWDLHKKI TOIE 137
Dbaa1-736 1 MGYYKS -- NYPQVMREAGFKLRQYRDGP LDMRLMGSYETERI LREONFELVDKALQHLSEAPLGTVLETHI LDDGI AKYF IMSQYA IYLLCCRTYLDSEVDELREAEHSQOEIAKLRKLS LSESNNEWDLHKKI TOIE 137
Dbaa1-737 1 MGYYKS -- NYPQVMREAGFKLRQYRDGP LDMRLMGSYETERI LREONFELVDKALQHLSEAPLGTVLETHI LDDGI AKYF IMSQYA IYLLCCRTYLDSEVDELREAEHSQOEIAKLRKLS LSESNNEWDLHKKI TOIE 137
Dbaa1-737 1 MGYYKS -- NYPQVMREAGFKLRQYRDGP LDMRLMGSYETERI LREONFELVDKALQHLSEAPLGTVLETHI LDDGI AKYF IMSQYA IYLLCCRTYLDSEVDELREAEHSQOEIAKLRKLS LSESNNEWDLHKKI TOIE 137
Dbaa1-737 1 MGYYKS -- NYPQVMREAGFKLRQYRDGP LDMRLMGSYETERI LREONFELVDKALQHLSEAPLGTVLETHI LDDGI AKYF IMSQYA IYLLCCRTYLDSEVDELREAEHSQOEIAKLRKLS LSESNNEWDLHKKI TOIE 137



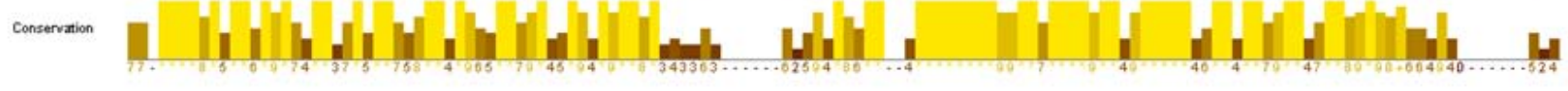
150 160 170 180 190 200 210 220 230 240 250 260 270

Dbaa1-734 138 TIREVVF PCHLCTKNF ISNEALNVHI GRKHRMGTPSS I VAGGA --- TVRDKENDMOL INT IKMELEIKOLKERLNAAERN I KERSGG SRS - HLQLDDQTTST - MROVGIQSNLADYKEKDDVSS EATES E262
Dbaa1-734 138 TIREVVF PCHLCTKNF ISNEALNVHI GRKHRMGTPSS I VAGGA --- TVRDKENDMOL INT IKMELEIKOLKERLNAAERN I KERSGG SRP - LLQLDDQTTST TMMNRVGIQSNLADYKEKDEI SSEATES E264
Dbaa1-727 138 TIREVVF PCHLCTKNF ISNEALNVHI GRKHRMGTPSS I VAGGA --- TVRDKENDMOL INT IKMELEIKOLKERLNAAERN I KERSGG QL - - - - - QDQTTST TMMNRVGIQSNLGEYKEKDEIVSMEAES E259
Dbaa1-750 138 TIREVVF PCHLCTKNF ISNEALNVHI GRKHRMGTPSS I VAGGA --- TVRDKENDMOL INT IKMELEIKOLKERLNAAERN I KERSGG SKLPASPRDQATATMMNRVGIQSNLANEYKEDKDEL SNEALNS E265
Dbaa1-755 138 AIREVVF PCHLCTKNF ISNEALNVHI GRKHRMGTPSS I VAGGA --- TVRDKENDMOL INT IKMELEIKOLKERLNAAERN I KERSGG SRRHVSPRDQATATMMNRVGIQSNLANEYKEDKDEL SNEALNS E266
Dbaa1-746 141 TIREVVF PCHLCTKNF ISNEALNVHI GRKHRMGTPSS I VAGGA --- TVRDKENDMOL INT IKMELEIKOLKERLNAAERN I KERSGG SRRHVSPRDQATATMMNRVGIQSNLANEYKEDKDEL SNEALNS E266
Dbaa1-745 138 TIREVVF PCHLCTKNF ISNEALNVHI GRKHRMGTPSS I VAGGA --- TVRDKENDMOL INT IKMELEIKOLKERLNAAERN I KERSGG SKR - P SPRODD - - - - - RVNGI QSNLTDKEKDDLS ETOGES E257
Dbaa1-745 138 TIREVVF PCHLCTKNF ISNEALNVHI GRKHRMGTPSS I VAGGA --- TVRDKENDMOL INT IKMELEIKOLKERLNAAERN I KERSGG SKR - P SPRODD - - - - - RVNGI QSNLTDKEKDDLS ETOGES E257
Dbaa1-738 138 TIREVVF PCHLCTKNF ISNEALNVHI GRKHRMGTPSS I VAGGA --- TVRDKENDMOL INT IKMELEIKOLKERLNAAERN I KERSGG SKR - V SPRODD - - - - - RVNGI QSNLAEPEKEDDES GEARDES E262
Dbaa1-734 138 AIREVVF PCHLCTKNF ISNEALNVHI GRKHRMGTPSS I VAGGA --- TVRDKENDMOL INT IKMELEIKOLKERLNAAERN I KERSGG SKR - V SPRODD - - - - - RVNGI QSNLAEPEKEDDES GEARDES E259
Dbaa1-736 138 AIREVVF PCHLCTKNF ISNEALNVHI GRKHRMGTPSS I VAGGA --- TVRDKENDMOL INT IKMELEIKOLKERLNAAERN I KERSGG SKR - V SPRODD - - - - - RVNGI QSNLAEPEKEDDES GEARDES E259
Dbaa1-737 138 AIREVVF PCHLCTKNF ISNEALNVHI GRKHRMGTPSS I VAGGA --- TVRDKENDMOL INT IKMELEIKOLKERLNAAERN I KERSGG SKR - V SPRODD - - - - - RVNGI QSNLAEPEKEDDES GEARDES E259
Dbaa1-737 138 AIREVVF PCHLCTKNF ISNEALNVHI GRKHRMGTPSS I VAGGA --- TVRDKENDMOL INT IKMELEIKOLKERLNAAERN I KERSGG SKR - V SPRODD - - - - - RVNGI QSNLAEPEKEDDES GEARDES E259
Dbaa1-737 138 AIREVVF PCHLCTKNF ISNEALNVHI GRKHRMGTPSS I VAGGA --- TVRDKENDMOL INT IKMELEIKOLKERLNAAERN I KERSGG SKR - V SPRODD - - - - - RVNGI QSNLAEPEKEDDES GEARDES E259



290 300 310 320 330 340 350 360 370 380 390 400 410

Dbaa1-734 263 AT - ERKEQLHGLAERLSNF EAWQADLKOSNEOF IRG INOKLEDL IALEOTKRNAESK - - GAAPTAAEERVATP - - CLEDLERI LTEKVAEIGKVSANKLEEVVHOLETOYKEKLEALERDLRRLS LRKQ - - - - - PAA391
Dbaa1-734 265 AT - ERKEQLHGLAERLSNF EAWQADLKONNEOF IRG INOKLEDL IALEOTKRNAESK - - SATPAAEERVATP - - CLEDLERI LTEKVAEIGKVSANKLEEVVHOLETOYKEKLEALERDLRRLS LRKQ - - - - - PVT392
Dbaa1-727 265 AT - ERKEQLHGLAERLSNF EAWQADLKOSNEOF IRG INOKLEDL IALEOTKRNAESK - - VTAPAAEERVATP - - CLEDLERI LTEKVAEIGKVSANKLEEVVHOLETOYKEKLEALERDLRRLS LRKQ - - - - - PVA387
Dbaa1-750 266 AG - ERKEQLHGLAERLSNF EAWQADLKOSNEOF IRD INOKLTDLSRAMEGQKLEAEAKVTVTPAAEERVATP - - SLEDLERI LTOKVAEIGKVSANKLEEVVHOLETOYKEKLEALERDLRRLS LRKQ - - - - - AIV396
Dbaa1-755 267 AT - ERKEQLHGLAERLSNF EAWQADLKOSNEOF IOD INOKLADLSHALEOSKOASG - - - - - VTPAAEERVATP - - CLEDLERI LSEKVAEIGKVSANKLEEVVHOLETOYKEKLEALERDLRRLS LRKQ - - - - - SLEK395
Dbaa1-746 272 TS PERKERLHGLAERLSNF EAWQADLKOSNEOF IOD INOKLADLSHALEOSKOASG - - - - - AASPTSEERLQTP - - RLEDLERI LTEKVAEIGKVSANKLEEVVHOLETOYKEKLEALERDLRRLS LRKQ - - - - - NEP400
Dbaa1-745 258 AS - ERKEQLHGLAERLSNF EAWQADLKOSNEOF IMD INOKLEGLSHALEOQTKTGTONT - - - - - EPTPLDGRMPTPTPCLEDLERI LTEKVAEIGKVSANKLEEVVHOLETOYKEKLEALERDLRRLS LRKQ - - - - - PKP386
Dbaa1-745 258 AS - ERKEQLHGLAERLSNF EAWQADLKOSNEOF IMD INOKLEGLSHALEOQTKTGTONT - - - - - EPTPLDGRMPTPTPCLEDLERI LTEKVAEIGKVSANKLEEVVHOLETOYKEKLEALERDLRRLS LRKQ - - - - - PKP386
Dbaa1-738 263 AT - ERKEQLHGLAERLSNF EAWQADLKOSNEOF IOD INOKLEGLSHALEOQTKTGTONT - - - - - PPLEDRVATP - - NLEDLERI LTEKVAEIGKVSANKLEEVVHOLETOYKEKLEALERDLRRLS LRKQ - - - - - PEP387
Dbaa1-734 260 AS - ERKEQLHGLAERLSNF EAWQADLKOSNEOF IOD INOKLEGLSHALEOQTKTGTONT - - - - - PPLEDRVATP - - CLEDLERI LSEKVAEIGKVSANKLEEVVHOLETOYKEKLEALERDLRRLS LRKQ - - - - - PEP384
Dbaa1-736 260 AS - ERKEQLHGLAERLSNF EAWQADLKOSNEOF IOD INOKLEGLSHALEOQTKTGTONT - - - - - PPLEDRVATP - - CLEDLERI LSEKVAEIGKVSANKLEEVVHOLETOYKEKLEALERDLRRLS LRKQ - - - - - PEP384
Dbaa1-737 260 AS - ERKEQLHGLAERLSNF EAWQADLKOSNEOF IOD INOKLEGLSHALEOQTKTGTONT - - - - - PPLEDRVATP - - CLEDLERI LSEKVAEIGKVSANKLEEVVHOLETOYKEKLEALERDLRRLS LRKQ - - - - - PEP384
Dbaa1-737 260 AS - ERKEQLHGLAERLSNF EAWQADLKOSNEOF IOD INOKLEGLSHALEOQTKTGTONT - - - - - PPLEDRVATP - - CLEDLERI LSEKVAEIGKVSANKLEEVVHOLETOYKEKLEALERDLRRLS LRKQ - - - - - PEP384
Dbaa1-737 260 AS - ERKEQLHGLAERLSNF EAWQADLKOSNEOF IOD INOKLEGLSHALEOQTKTGTONT - - - - - PPLEDRVATP - - CLEDLERI LSEKVAEIGKVSANKLEEVVHOLETOYKEKLEALERDLRRLS LRKQ - - - - - PEP384



430 440 450 460 470 480 490 500 510 520 530 540 550

Dmao1-734 392 EVQQAAS ---- TSK I - PKALPKKEE - NMDRIKRLVETEF LKAKRDDOTYS IEPPFP - - - - - PPEEHVTH - - - - - VQEQPSONSSGSHPTTYTKPAPAPAPARDEVKPKATTS EPEQSEATD ISESLS - - - - - 502
Dmao1-734 393 EVQQAAS ---- TSK I - PKALPKKEE - NMERIKRLVETEF LKAKRDDOTYS IEPPFP - - - - - PPEEHVTH - - - - - VQELPSONSSGSHPTTYTKPAPAPAPARDELKPRTTTSEPEQPEATD ISESLS - - - - - 503
Dmao1-727 388 EVQQAAS ---- ASK I - PKPLPKKEE - SMERIKRLVETEF LTKRDDOTYS IEKPPP - - - - - QPEQQAQ - - - - - KQEQDASGSSSSHPPTTYTKQAPASAPARNEPKKAAINOPEQPEATD ISESLS - - - - - 496
Dvir1-750 397 V - QPPAN - - - - - PSK I PKPKPLPKKEE - SMERIKRLVETEF LNAKRDDOTYS IEPPAQSAQOQG - QOKHAVTH - - - - - APEQPSGSSGSHPTTYTK - - - - - PTAEPFRNEPEPKPSTSTOTETOETD ISESLA - - - - - 513
Dvir1-755 401 KQQPQPS - - - - - ATTSK I - PKPQAKKEL - TMERIKKQVETEF LQAKRDDOTYS IEELPPAQOQEPK - QOKHAVTH - - - - - VEELVSGSSSSHPPTTYTKP - PAPEP IVNEAKSG - - - - - ANQMAQES TVS ELS DEE - - - - - 520
Dwil1-746 401 ESKLEPVALVTNVSK I - PKP IVNRRENS LEH I KNOVEDEF VEKKRDDOTYS IEELPKPL - - - - - VTQVQVQVH - - - - - EAKOP IVESLTANOSNKEI - - - - - QOTS DSTO ISESLS HEDD - - - - - KA507
Dped1-745 387 - - QTVAIFHGSOTSK I - PKPVS RKEETNVER I RKOVES EF LKS KADDOTYS IEG - PREKS LAPP P P P P PVEVQVE - - - - - INVQPSGSSGSHPTTYTKP - DA I EPA - - - - - LSKPEARETTD IDESLS HEDAG IVHDD 508
Dped1-745 387 - - QTVAIFHGSOTSK I - PKPVS RKEETNVER I RKOVES EF LKS KADDOTYS IEG - PREKS LAL P P P P PVEVQVE - - - - - INVQPSGSSGSHPTTYTKP - DA I EPA - - - - - LSKPEARETTD IDESLS HEDAG IVHDD 508
Dana1-738 388 V - Q I LFN - - - - - ASK I - PKP INRKEETSMDR I RKOVES EF VKEKHDDOTYS IEAPREIPKI - - - - - QKTQVQVHEKNEKEQPSAGSSGSHPTTYTKP - PAFSA - - - - - SNKPET I ETTOVS ELS QEEV - - - - - VG 501
Dana1-734 385 L - QTVPA - - - - - ASK I - PKPVALKEETN I DR I RKOVES EF LKKKHDDOTYS IEAPRKS PVK P P P LVTQVQV - - - - - EKEQPSAGSSGSHPTTYTKS - PR - EPV - - - - - PTKLETREATD ISESLS QEETE - - - - - NE 499
Dyale1-736 385 V - QTVPA - - - - - ASK I - PKPVALKEETN I DR I RKOVES EF LKKKHDDOTYS IEAPRKS PVK P P P LVTQVQV - - - - - EKELPSAGSSGSHPTTYTKS - PR - EPV - - - - - SSKPEAREATD ISESLS QEETE - - - - - NE 499
Dmao1-737 385 V - QTVPV - - - - - ASK I - PKPVVRKEETN I DR I RKOVES EF LKOKHDDOTYS IEAPRKS EKP P P LVTQVQV - - - - - EKEQPSAGSSGSHPTTYTKS - PR - EPA - - - - - PNKQETKEATQVS ELS QEETE - - - - - NE 499
Dmao1-737 385 V - QTVPV - - - - - VSK I - PKPVVRKEETN I DR I RKOVES EF LKOKHDDOTYS IEAPRKS SEN P P P LVTQVQV - - - - - EKEQSSAGSSGSHPTTYTKS - PR - EPF - - - - - PSKPETKETTVS ELS QEETE - - - - - NE 499
Dmao1-737 385 V - QTVPV - - - - - ASK I - PKPVVRKEETN I DR I RKOVENEF LKOKHDDOTYS IEAPRKS EKP P P LVTQVQV - - - - - EKEQSSAGSSGSHPTTYTKS - PR - EPI - - - - - PSKPETKETTVS ELS QEETE - - - - - NE 499

Conservation



570 580 590 600 610 620 630 640 650 660 670 680 690

Dmao1-734 503 GEES I S - DEGS E V L T S E P E - - - - - RQV F M S P K I K S S L K T K L - - - - - P P K P L T R K D A R K L I N O R L S P H G F N M K S K T I S N T T A K R V S A E L A Q O R A R L K L D Y P N F Y T T R N R I R K F V E K L C S A K M P E R A Q I L L K N K T P L O P M E V P K S R N 636
Dmao1-734 504 GEES I S - DEGS E V L T S E P E - - - - - RQV F M S P K I K S A L K T K L - - - - - P P K P L T R K D A R K L I N O R L S P H G F N M K S K T I S N T T A K R V S A E L A Q O R A R L K L D Y P N F Y T T R N R I R K F V E K L C S A K M P D H A O N L L K N K T P L O P M E V P K S R N 637
Dmao1-727 497 GEES I S - DEGS E V L T S E P E - - - - - RQV F M S P K I K S V L K N R L - - - - - P P K P L T R K D A R K L I N O R L N P H G F N M K S K A I S N T S A K R V S A E L A Q O R A R L K L D Y P N F Y T T R N R I R K F V E K L C S V K M P E R A Q I L L K N K T P L O P M E V P K R R N 630
Dvir1-750 514 EEEESV S - DEGS E V L T S E P E - - - - - R R V F M S P S K T T K K A R V P P K P P L T R K D A R K L V N L K L N P H G F N M K T S L S N T S L K R V S A E L A Q H R N R L K L D Y P H F Y T T R N R I R I F V E K L C S A K M P E R A Q V L L K T K T P L O P M V P K R R T 650
Dvir1-755 521 EEEES I S - DQGS E L L T S E P E - - - - - RQV F K S P K I K V T P K I K S Q K P P O P L T R K D A R K L V N L K L N P H G F N M K S K S I S N T Y L K R A S E L A Q N R N K L K L E H P H F Y A T R N R I R K F V D K L C S A K M P E R A E E L L M N K T P L O P M E V P K R R S 657
Dwil1-746 508 EQGSP S - EDS E V L T S D R E S F E E I L K K P S P I I N P A F K M - - - - - S P P K A M T R K D M K L V N R L S H G F D M S K G I S H T S M K L I N M E L A E N R N K L K L O Y P N F Y A T R N R I R K F V E K L C S A K L P O H A E V L K H K T P L O P I E V P K R R A 644
Dvir1-745 509 TEQTL S - DQGS E L T S E A D L P R T V - E K A A S - S R P T A L S L R S P O K P L T R K D A R K L V I R K M S S H G Y D V K S K T I S H T T M K R I N G E L T E O R N K L K L O Y P O F Y A T R N R I R K F V E K L C S T K L P O R A E N L L K H R T P L O P V E A P K R G S 645
Dped1-745 509 TEQTL S - DQGS E L T S E A D L P R T V - E K A A S - S R P T A L S L R S P O K P L T R K D A R K L V I R K M T S H G Y D V K S K T I S H K T M K R I N G E L T E O R N K L K L O Y P O F Y A T R N R I R K F V E K L C S T K L P O R A E N L L K H R T P L O P I E A P K R G S 646
Dana1-738 602 EEQS L S - E D G S E I S S S G S E P H R E S - N R P S T A T K P P V R I T R P P K P L T R K D A R K L V N R K L N P H G F D L S K G I S H T S L K R V N S E L A E O R N K M L H Y P O F Y A T R N R I R K F V E K L C S A R F S E R A O V I L K H K T P L K P I E A P K R G V 639
Dana1-734 500 EES L P E E E G T E A S T S E S E A P R E D - P K P K T - I K P S G R I I K S P O K P L T R K D A R K M N R K L M P H G F D M S K G I S H T S L K R V N S E L A E R N K L K L O Y P H F Y A T R N R I R K F V E K L C S A K F S E R A E M L L K H K S P L K P M E V P G R G I 637
Dyale1-736 500 EERS L T - E E G T E V S A S E S E A P R E D - S K P K T - I K P S G R I I K S P O K P L T R K D A R K M N R K L M P H G F D M S K G I S H T S L K R V N S E L T E H R N K L K L O Y P H F Y A T R N R I R K F V E K L C S A K F S E R A E M L L K H K S P L K P M E V P G K R I 636
Dmao1-737 500 EERS L T E E E G T D V P T S G S E A P R E D - P T P K T - I K P S G R I I K S P O K P L T R K D A R K M N R K L M P H G F D M S K G I S H N S L K R V N S E L T E H R N K L K L O Y P H F Y A T R N R I R K F V E K L C S A K F S E R A E M L L K H K S P L K P M E V P G K G I 637
Dmao1-737 500 EERS L T E E E G T D V P T S E S E A P R E D - P K P K T - I K P S G R I I K S P O K P I T R K D A R K M N R K L M P H G F D M S K G I S N T S L N R V N S E L T E H R N K L K L O Y P H F Y A T R N R I R K F V E K L C S A K F S E R A E M L L K H K S P L K P M E V P G K G I 637
Dmao1-737 500 EERS L T E E E G T D V P S G S E S P R E D - P K P K T - I K P S G R I I K S P O K P I T R K D A R K M N R K L M P H G F D M S K G I S H T S L K R V N S E L T E H R N K L K L O Y P H F Y A T R N R I R K F V E K L C S A K F S E R A E M L L K H K S P L K P M E V P G K G I 637

Conservation



710 720 730 740 750 760 770 780 790 800 810

Dmao1-734 637 L S L T T D D D D E L N E G S D V T S - - - - - A S Q G V E E D E E I E K A S T S S K - - - - - R Q Q - K N F K A Q L E Q L L A K P A A H V V P K - - - - - P K L V Q I Q Q A K P M P L P R K R V M F N T E G S R K N - - - - - N E D S A E 734
Dmao1-734 638 L L L T T D D D D E L N E G S D V T S - - - - - A S Q G M D G D E E I E N A S S S - - - - - K R Q Q N F K A Q L E Q L L A K P A A H V V P K - - - - - P K L V Q I Q Q A R P V P L P R K R V M F N T E G S R K S - - - - - N E D S A K 734
Dmao1-727 631 P L L T T D D D D E L N E G S D V S S - - - - - A S Q - E M D N E K V D N A S T S S - - - - - K A Q D P N F K A Q L E Q L L A K P A A H V V P K - - - - - P N M V L Q D A K P V P L P R K R V M F N T G S S R R S - - - - - S E D N V E 727
Dvir1-750 651 G L P A T T D D D D G N D A S E R T S - - - - - V S Q E E E E V D E A S T T S S K P O - - - - - R Q E Q H L N F K A Q L E Q M L A K P V A Q V I A K - - - - - P T L G Q T Q A K P M P L P R K R V M F N T L G S G K S - - - - - I D T N D E 750
Dvir1-755 658 A V L G A T D D D D D N E A S E H T S L E E A G E D E G T E E K K Q S T S S K A L R Q Q H G N F K A Q L E Q M L A R P V A Q V S K - - - - - P N M G - I M Q A R P V P L P R K R V M F N T L G S G - - - - - 755
Dwil1-746 645 K N O R A V T T D E D N L E K - - - - - Q P D E K D D S A S S S R S T S P P - - - - - R R Q V S H D F K A H L D O I L V K P V A A V A S K S S L A S G N A N P S P N V R P V P L P R K R V M F N T L S S G K S F - - - - - N E H D E N 746
Dped1-745 646 L R S A P M O E P D T A A H S S Q E - - - - - E E D R S G S E S E S R G S S P Q - - - - - R Q V D E N F K A R L E E I L V K P A A I A A P R - - - - - E A S K S S L S R P V P L P R K R V M F N T M V G G K G I N D S E D L K 745
Dped1-745 646 L R S T S M O E P D A A A H S S Q E - - - - - E E D R S E S E S E S R G S S P Q - - - - - R Q V D E N F K A R L E E I L V K P A A I A A P R - - - - - E A S K S S L S R P V P L P R K R V M F N T M V G G K G I N D S E D L K 745
Dana1-738 640 L K A G T S E K S E E D L - - - - - S O H T S - - - - - O D E E E G S D E S O S R S S P R - - - - - O R S I S R D F K A R L E E I L V K P A A M T I - - - - - G A S K P A S I S K P V P L P R K R V M F N T L N S G K S F N E S D E D Y K 738
Dana1-734 638 V R - - - - - S E K S V E D I A S S Q D E E - - - - - Q A D E Q T D S E Q Q S S S P R - - - - - R P V S R D F K A R L E E I L V K P A A T V - - - - - G A S K S S L S R P V P L P R K R V M F N T T E D G K S F M D R D D N L K 734
Dyale1-736 637 A R S A T S E K S E E D V A S S Q D E E - - - - - Q T D E Q T D S E Q Q S S S P R - - - - - K L V S R D F K A R L E E I L V K P A A T V - - - - - G A S K S S L S R P V P L P R K R V M F N T T E D G K S F N D S D D N L K 736
Dmao1-737 638 P R S A I S E K S E E D I A S S Q G E E - - - - - Q T D E Q T D S E Q Q T R S S P Q - - - - - R L V S R D F K A R L E E I L V K P A A T I - - - - - G A S K S S L S R P V P L P R K R V M F N T T E D G K S F N D S D D N L K 737
Dmao1-737 638 P M S A I S E K S E E D I A S S Q A E E - - - - - Q E D E Q T D S E Q Q S R S P Q - - - - - R L V S R D F K A R L E E I L I K P A A T I - - - - - G A S K S S L S R P V P L P R K R V M F N T T E G G S F N D S D D N L K 737
Dmao1-737 638 P M S A I S E K S E E D I A S S Q E E D - - - - - Q E D E Q T D S E Q Q S R S P Q - - - - - R L V S R D F K A R L E E I L V K P A A T I - - - - - G A S K S S L S R P V P L P R K R V M F N T T E G G S F N D S D D N P K 737

Conservation



Two putative PEST sequences in positions 466-492 and 531-582 of the alignment were also identified in CG13617 protein (FIGURE 28). PEST domains are polypeptide sequences enriched in proline (P), glutamic acid (E), serine (S) and threonine (T) whose function is to target proteins for rapid degradation. They consist of hydrophilic stretches of at least 12 residues in length that do not contain positively charged amino acids (RECHSTEINER and ROGERS 1996). Multiple PEST signals can be found in a single protein but their positions are variable within the polypeptide chain. PEST sequences are found in metabolic enzymes, transcription factors, protein kinases and phosphatases, as well as components of signal pathways and cyclins (RECHSTEINER and ROGERS 1996), all of them proteins that are short-lived and that exhibit fast changes in concentration. In these situations proteolysis has been shown to be a widespread regulatory mechanism and PEST signals appear to be widely distributed among these proteins. Most of the PEST sequences are conditional signals that have to be activated to promote proteolytic degradation (a conformational change can expose the signal). Calpain proteases have been shown to catalyze some cases of PEST-dependent degradation (SANDOVAL *et al.* 2006) but caspase cleavage followed by proteasomal degradation has also been reported for proteins containing PEST motifs (BELIZARIO *et al.* 2008, LUKOV and GOODELL 2010).

Finally, another feature that could be identified through bioinformatic analysis in the CG13617 protein sequences is a putative nuclear localization signal (NLS) located close to the C-terminal end of the protein, in positions 783-789 of the alignment (FIGURE 28). NLSs are generally short peptides that contain a high proportion of positively charged amino acids that target a protein to be transported into the nucleus (JANS *et al.* 2000). Although different classes have been described, the putative NLS detected in the CG13617 protein (PLPRKRV) is similar to the well-characterized NLS of the SV40 large T antigen (with the sequence PKKKRKV) (KALDERON *et al.* 1984). This short 7-residue sequence has a high proportion of basic positively charged amino acids (lysine, K, and arginine, R) and it is completely identical in the 14 analyzed *Drosophila* species, with no amino acid changes in any of them. This high conservation level is especially striking since the putative NLS is located in the C-terminal portion of the protein, an otherwise poorly-conserved and very divergent region (FIGURE 28). A nuclear export signal (NES) could also be identified in *D. buzzatii*, *D. martensis* and *D. mojavensis* CG13617 protein sequences in positions 398-407 of the alignment, overlapping the

fourth *D. buzzatii* coiled coil region (FIGURE 28). A NES is a short amino acid sequence of 5-6 hydrophobic residues that is recognized and bound by exportines and that targets a protein for export from the cell nucleus to the cytoplasm through the nuclear pore complex (LA COUR *et al.* 2004). The NES found in CG13617 adjusts perfectly to the consensus sequence LX₁₃LX_{2,3}LXL, in *D. buzzatii*, *D. martensis* and *D. mojavensis*. In the remaining species the last leucine residue has been replaced by different amino acids, but the other three leucines are conserved in all the proteins except for *D. willistoni* and *D. grimshawi*, where the third leucine has been substituted by an isoleucine.

Apart from these relatively conserved domains and motifs, the rest of the protein shows varying degrees of conservation between species. A graphic representation of conservation values calculated all along the protein using the AL2CO software (PEI and GRISHIN 2001) can be observed in FIGURE 29. As can be seen in the graph, conservation tends to be higher in the N-terminal region of the protein than in the C-terminal region. For example, the first part of the protein (from the initial methionine to the beginning of the first coiled-coil region) does not seem to include any functional domain but presents nevertheless a 72.92% identity (70 conserved aminoacids out of 96 aligned positions) considering the 14 sequences, so it is probably an essential part for the proper folding and functioning of the protein. In contrast, of the last 406 aligned positions (positions 411-816 of the alignment), which correspond to the region comprising from the end of the last coiled coil to the C-terminal end of the protein, only 80 (19.70%) exhibit identical amino acids in the 14 sequences and they include 111 positions (27.34%) presenting gaps at least in one of the sequences, while only 3.12% of gaps (3/96) were observed in the first segment of the protein mentioned above.

Finally, given that all *Drosophila* CG13617 genes have been annotated based on their sequence homology with the predicted gene in *D. melanogaster* and that nothing is known about the gene product encoded by this gene in any of these species, we performed similarity searches to try to find CG13617 orthologues in species outside the genus *Drosophila* that could provide more information about the possible function of this protein. Two different searches were performed using *D. buzzatii* 2st CG13617 protein sequence as query: BLASTP (protein query *vs.* protein database) to identify any homologous proteins in other species (TABLE 12) and TBLASTN (protein query *vs.* translated nucleotide database) to detect any unannotated

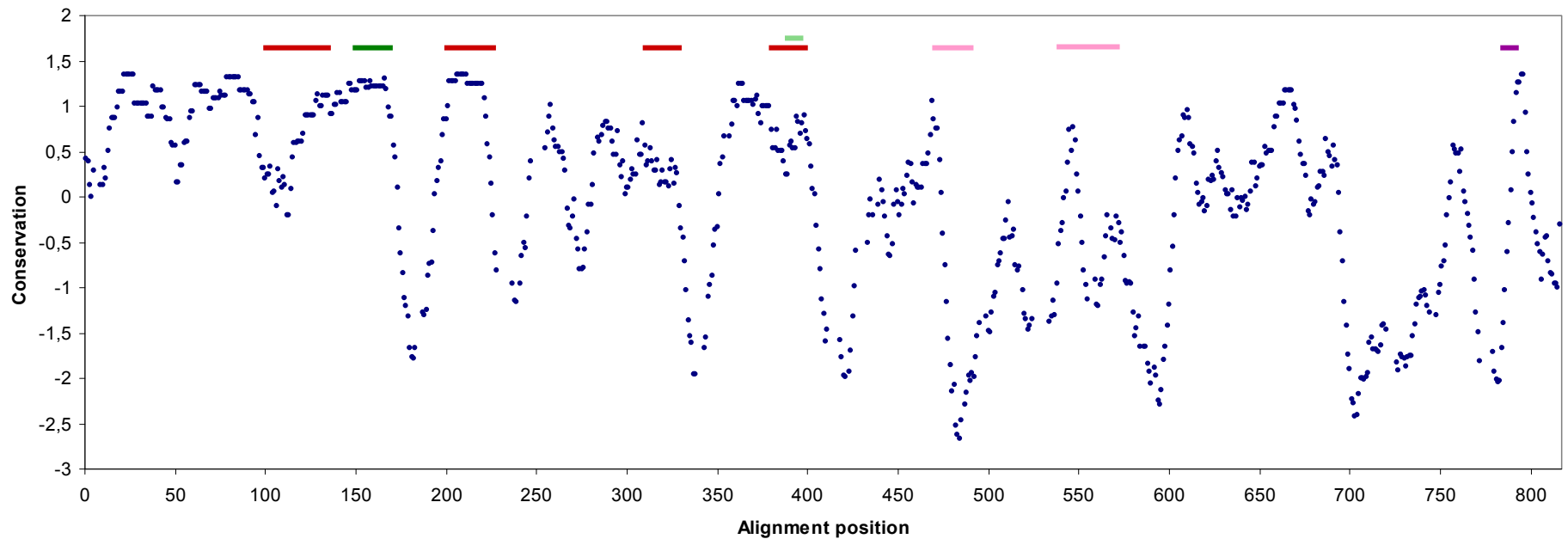



FIGURE 29 | Sliding window graph of CG13617 protein sequence conservation in the 14 *Drosophila* species. AL2CO program  was used to calculate conservation values for 10-residue windows along the CG13617 protein (see Materials and Methods for details). The bars above the graphic denote the location of the functional domains using the same color code than in FIGURE 28: red, coiled-coil regions; dark green, C2H2-type zinc finger; pink, PEST sequences; purple, NLS, and light green, NES.


orthologous sequences (TABLE 13). These two searches yielded very similar results and led to the identification of CG13617 homologous sequences in several organisms, ranging from insects to mammals (TABLES 12 and 13). Only two insect species (the mosquito *Aedes aegypti* and the beetle *Tribolium castaneum*) appear in the TBLASTN results but not in the BLASTP list, which probably indicates an erroneous or poor annotation of the corresponding regions in the genomes of these two species and indicates that further analysis would be required to identify the correct ORF. For the majority of species, the region of homology obtained in the similarity searches corresponds basically to the initial part of the protein (the zinc finger domain and surrounding sequences), but not to the whole peptide. The only exception is the mosquito species, where the aligned segment could be extended much farther away, sometimes including the whole mosquito protein.

The comparison of the homologous protein sequences in different organisms allowed us to determine the amino acid positions that are critical for CG13617 function. First, all the predicted homologous proteins appear to include a C2H2-type zinc finger with the two cysteine residues and the two histidines perfectly conserved in all the species. The only protein without a zinc finger is that of *Apis mellifera* (TABLE 12) which is also extremely short (242 aa). However, the absence of this domain is probably due to the fact that the genome sequence where the protein was annotated has several gaps that do not allow us to determine the correct and complete ORF. There are also some groups of a few amino acids that show a very high level of conservation among the different homologous proteins, although they are not part of any of the known functional motifs. Even in proteins as distant from *Drosophila* as those from mammals these residues seem to be unchanged and retain the same relative positions. This is the case of the residues DWR in positions 30-33 of the alignment (FIGURE 28) or residues YLL in positions 89-91. Also, it is noticeable that many of the CG13617 orthologous proteins found in these similarity searches (mainly those not belonging to insects) have been tentatively named DAZ-interacting protein-like, which means they are similar in sequence (and hence probably function) to the DAZ interacting protein (*DZIP1*), a protein partially characterized in humans (MOORE *et al.* 2004) and that has been more extensively studied in zebrafish, where it has been identified as a new member of the Hedgehog signaling pathway (SEKIMIZU *et al.* 2004, WOLFF *et al.* 2004). This similarity with human DZIP1 and

TABLE 12 | CG13617 protein orthologues in species other than *Drosophila* identified using BLASTP. E-values and pairwise alignments were computed with BLASTP (protein query *vs.* protein database) using *D. buzzatii* 2st protein sequence as query. The proteins for which further information exists about their structure and function are highlighted in yellow.

Gene name	Protein name	Species	Species common name	GeneID	Protein Refseq	Length (aa)	Zinc finger	Score	E-value
AgaP_AGAP001165	AGAP001165-PA	<i>Anopheles gambiae</i>	African malaria mosquito	4577365	XP_001238569.2	607	YES	220	4.00E-55
Cpip_CPIJ011569	hypothetical protein	<i>Culex quinquefasciatus</i>	southern house mosquito	6044733	XP_001862196.1	758	YES	214	3.00E-53
BRAFLDRAFT_119689	hypothetical protein	<i>Branchiostoma floridae</i>	Florida lancelet (Amphioxus)	7224854	XP_002208802.1	1564	YES	80.5	7.00E-13
NEMVEDRAFT_v1g247454	hypothetical protein	<i>Nematostella vectensis</i>	starlet sea anemone	5503868	XP_001624813.1	436	YES	77.0	7.00E-12
LOC724732	similar to CG13617-PA	<i>Apis mellifera</i>	honey bee	724732	XP_001120624.1	242*	NO	70.9	4.00E-10
LOC575991	hypothetical protein	<i>Strongylocentrotus purpuratus</i>	purple sea urchin	575991	XP_799867.2	920	YES	70.5	7.00E-10
LOC100185506	similar to DAZ interacting protein 1-like	<i>Ciona intestinalis</i>	tunicate	100185506	XP_002129615.1	740	YES	68.2	3.00E-09
LOC100024195	similar to DAZ interacting protein 1-like	<i>Monodelphis domestica</i>	gray short-tailed opossum	100024195	XP_001375526.1	786	YES	64.3	5.00E-08
zgc:123017	DAZ interacting protein 1-like	<i>Danio rerio</i>	zebrafish	553313	NP_001032304.1	756	YES	63.5	8.00E-08
BRAFLDRAFT_147670	hypothetical protein	<i>Branchiostoma floridae</i>	Florida lancelet (Amphioxus)	7227261	XP_002218210.1	279	YES	63.5	9.00E-08
Dzip1l	DAZ interacting protein 1-like	<i>Mus musculus</i>	mouse	72507	NP_082534.2	774	YES	63.2	1.00E-07
Dzip1l	DAZ interacting protein 1-like	<i>Rattus norvegicus</i>	rat	315952	NP_001014117.1	776	YES	62.8	1.00E-07
LOC611021	similar to DAZ interacting protein 1-like	<i>Canis familiaris</i>	dog	611021	XP_853725.1	766	YES	62.8	1.00E-07
DZIP1L	DAZ interacting protein 1-like	<i>Bos taurus</i>	cow	512800	XP_590382.3	767	YES	61.6	3.00E-07
LOC716551	similar to DAZ interacting protein 1-like	<i>Macaca mulatta</i>	Rhesus monkey	716551	XP_001114691.1	766	YES	61.6	3.00E-07
dzip1l	DAZ interacting protein 1-like	<i>Xenopus tropicalis</i>	western clawed frog	780329	NP_001072868.1	467	YES	60.8	5.00E-07
LOC100197270	similar to predicted protein	<i>Hydra magnipapillata</i>	hydrozoan	100197270	XP_002170852.1	255	YES	60.8	5.00E-07
DZIP1L	DAZ interacting protein 1-like	<i>Homo sapiens</i>	man	199221	NP_775814.1	767	YES	60.8	5.00E-07
DZIP1L	DAZ interacting protein 1-like	<i>Equus caballus</i>	horse	100066729	XP_001496970.1	768	YES	60.8	5.00E-07
DZIP1L	DAZ interacting protein 1-like	<i>Pan troglodytes</i>	chimpanzee	460723	XP_516774.2	767	YES	60.8	5.00E-07
LOC100208186	similar to predicted protein	<i>Hydra magnipapillata</i>	hydrozoan	100208186	XP_002170687.1	326	YES	60.5	7.00E-07
dzip1	DAZ interacting protein 1	<i>Xenopus laevis</i>	African clawed frog	446844	NP_001087009.1	817	YES	58.9	2.00E-06
dzip1	DAZ interacting protein 1	<i>Xenopus tropicalis</i>	western clawed frog	733791	NP_001073034.1	858	YES	58.2	3.00E-06
DZIP1	DAZ interacting protein 1	<i>Gallus gallus</i>	chicken	418789	XP_416984.2	825	YES	50.8	5.00E-04
DZIP1	DAZ interacting protein 1	<i>Canis familiaris</i>	dog	476964	XP_534164.2	1130	YES	50.1	8.00E-04
DZIP1	DAZ interacting protein 1	<i>Macaca mulatta</i>	Rhesus monkey	695529	XP_001085577.1	912	YES	49.7	0.001
LOC100225719	similar to DAZ interacting protein 1	<i>Taeniopygia guttata</i>	zebra finch	100225719	XP_002196853.1	944	YES	48.9	0.002
DZIP1	DAZ interacting protein 1	<i>Pan troglodytes</i>	chimpanzee	452627	XP_001138000.1	856	YES	48.9	0.002
DZIP1	DAZ interacting protein 1	<i>Homo sapiens</i>	man	22873	NP_945319.1	867	YES	48.5	0.003
iguana (dzip1)	DAZ interacting protein 1	<i>Danio rerio</i>	zebrafish	402875	Q7T019	898	YES	48.5	0.003


* This protein sequence is probably incomplete due to the presence of gaps in the genomic sequence.

TABLE 13 | Additional CG13617 protein orthologues in species other than *Drosophila* obtained using TBLASTN. A TBLASTN  search (protein query vs. translated nucleotide database) was performed using *D. buzzatii* 2st protein sequence as query in order to detect any possible unannotated orthologous genes. E-values and pairwise alignments were computed with TBLASTN. In this case the search identified DNA sequences containing regions that present homology to CG13617 protein once they are translated into protein in one of the six possible reading frames. Coordinates allow the precise location of these sequences within the genomic contigs of the corresponding species.

GenBank accession	Species	Species common name	Genomic contig	Coordinates	Zinc finger	Score	E-value
AAGE02010559.1	<i>Aedes aegypti</i>	yellow fever mosquito	cont1.10559	30380-29649	YES	173	1.00E-40
NW_001092832.1	<i>Tribolium castaneum</i>	red flour beetle	linkage group 5 genomic contig	51133-51516 51629-51799	YES	49.3	2.00E-10

zebrafish Iguana proteins can provide valuable information about the cellular function of CG13617.

3.3.3 Identification of regulatory sequences in gene *CG13617*

As a way to gain insight on the regulation of gene *CG13617*, we examined at the nucleotide level the sequence upstream of *CG13617* coding region in the different *Drosophila* species to try to find conserved non-coding sequences (CNSs) that could correspond to potential regulatory elements. In order to do this, the intergenic region between *nAcR β -96A* and *CG13617* (including the first two exons of each gene) of the 14 available *Drosophila* species was aligned and analyzed using VISTA  (see Materials and Methods for details). This region includes *nAcR β -96A* and *CG13617* 5' UTRs, which are regions that may be conserved in different species but do not correspond to regulatory elements. *CG13617* 5' UTR was experimentally determined in this work to be 118 bp long in *D. buzzatii*. On the other hand, the location of *nAcR β -96A* TSS was established by comparison with *D. melanogaster* annotation, which suggests that *nAcR β -96A* 5' UTR could be 139 bp long in *D. buzzatii*, even though this prediction has not been validated for any of the analyzed species. However, no CNSs shared by all the available sequences were found in this region.

The analysis was then restricted to the five *Drosophila* subgenus species (FIGURE 27), which share a higher sequence identity with *D. buzzatii*. *D. martensis* sequence is too similar to *D. buzzatii* to distinguish conserved sequences that could have a potential functional role in gene regulation, but the comparison of *D. buzzatii* with *D. mojavensis* revealed three clearly defined CNSs, as can be seen in FIGURE 30. CNS1 is located next to *nAcR β -96A* coding region and it corresponds to this gene 5' UTR and an additional 105-bp region that could include regulatory elements (FIGURE 31). CNS2 is ~150 bp long and is located approximately 500 bp away from the TSSs of both genes (FIGURE 30), so if this sequence has a functional role it is not clear which gene it could be affecting. However, CNS2 could not be identified in either *D. virilis* or *D. grimshawi* (in any sequence orientation). Therefore, it might just be a segment of the sequence that by chance has diverged less than the surrounding sequences and

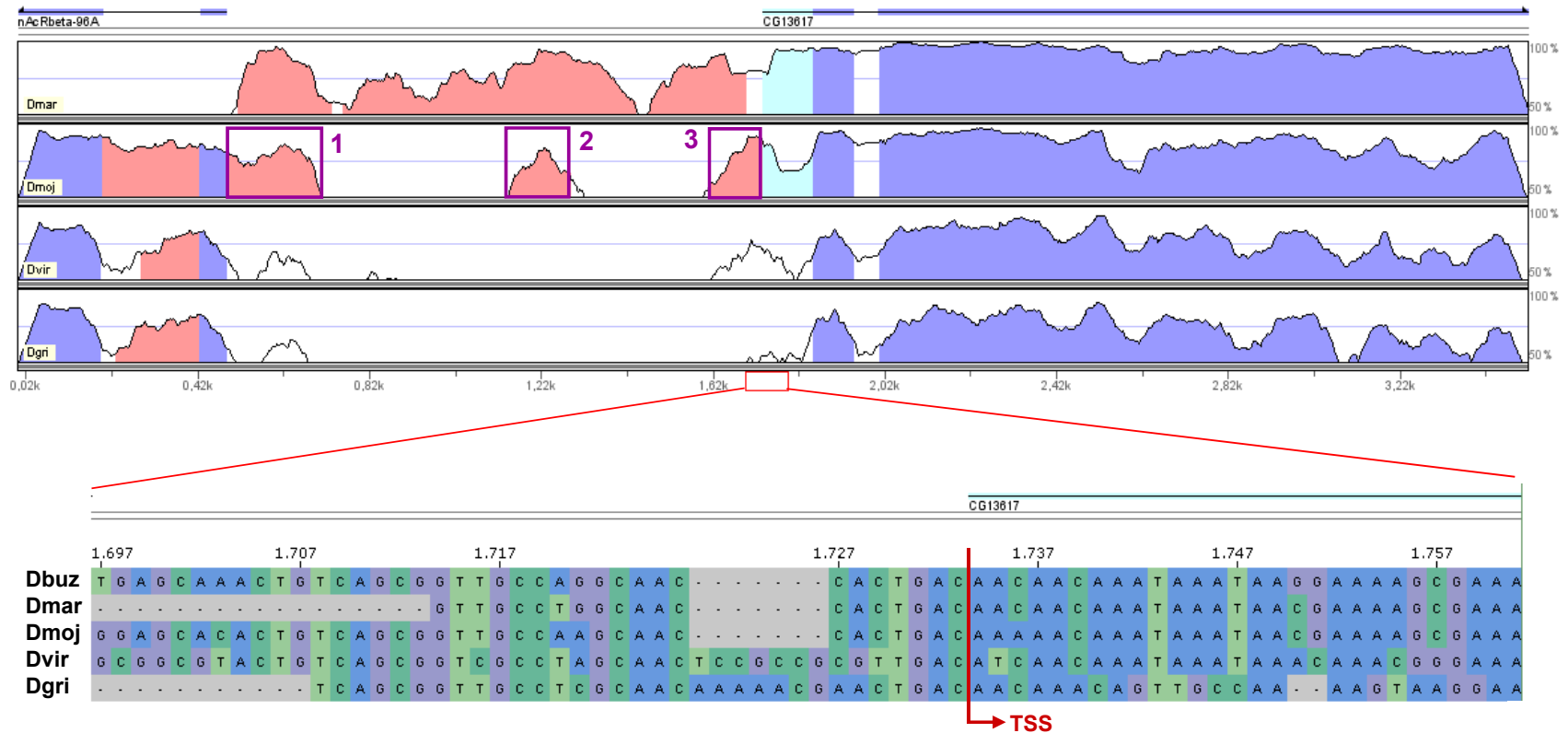





FIGURE 30 | Analysis of the sequences upstream gene *CG13617* using VISTA. The graph shows the level of sequence conservation of the five *Drosophila* subgenus species using *D. buzzatii* as reference (represented on top as a thin colored line) in the sequence comprised between the first two exons of genes *nAcRβ-96A* and *CG13617*. Each line represents the comparison of the species indicated on the left hand corner with the *D. buzzatii* sequence. The blue-shaded areas of the graph correspond to the coding sequences annotated in *D. buzzatii*, green-shaded areas to UTRs and pink-shaded zones to CNSs. The criteria used to identify these CNSs are given in the Materials and Methods section. The three significant CNSs identified in the *D. buzzatii*-*D. mojavensis* comparison are included in purple rectangles and numbered from 1 to 3. Below the graphs, a red box marks the part of the sequence represented in more detail in the multiple sequence alignment, which shows the identification of *CG13617* TSS (marked by a red arrow) in the different species based on *D. buzzatii* annotation. The coding region of gene *nAcRβ-96A* in *D. martensis* is not included because this sequence is not available for this species. *nAcRβ-96A* 5' UTR has not been annotated in the *D. buzzatii* reference sequence because the location of the TSS of this gene is based only on the comparison with the *D. melanogaster* annotation and has not been validated.


and it simply reflects the closer phylogenetic relationship between *D. mojavensis* and *D. buzzatii*. Finally, CNS3 is found immediately upstream of *CG13617* TSS and it is the most conserved CNS between these two species, with an 80.6% identity (TABLE 14). This CNS does not include the *CG13617* 5' UTR because it has been experimentally characterized and was annotated separately in the *D. buzzatii* reference sequence (FIGURE 30). MCPROMOTER  software identifies in the region corresponding to CNS3 a putative core promoter with a DNA-replication related element (DRE) able to be bound by TRF2 (TATA-box-binding protein related factor 2) (HOCHHEIMER *et al.* 2002). However, a more detailed sequence analysis failed to identify such element (consensus sequence TATCGATA) (HIROSE *et al.* 1993) in the sequences upstream of *CG13617* TSS, questioning the reliability of this prediction.

Although, according to the VISTA plot, the CNS3 segment does not meet the criteria to be considered a CNS either in *D. virilis* or *D. grimshawi* (FIGURE 30), the corresponding sequences of these two species can be reasonably well aligned with *D. mojavensis* and *D. buzzatii* sequences (FIGURE 31). This multiple alignment reveals a series of conserved fragments 5-10 bp long located upstream of *CG13617* TSS (within positions -1 and -124) that could correspond to regulatory elements or TFBSs that are part of the promoter controlling the expression of this gene (FIGURE 31). Curiously, some of these highly conserved short

TABLE 14 | CNSs upstream of gene *CG13617* identified by VISTA  in the alignment of *D. buzzatii* and *D. mojavensis* sequences. The coordinates of each CNS are given with respect to *D. buzzatii* st-1 line sequence in GenBank database and to *D. mojavensis* scaffold in the CAF1 genome assembly. Lengths are given in base pairs. Beginning and end positions are shown using the TSS (position +1) of the corresponding gene as a reference (*nAcRβ-96A* for CNS1 and *CG13617* for the rest). *nAcRβ-96A* 5' UTR is not annotated separately (and is therefore included in CNS1) because the precise location of the transcription start has not been experimentally validated in any of the analyzed species. Identities between *D. buzzatii* and *D. mojavensis* sequences correspond to the percentages calculated by VISTA.

CNS	Dbuz					Dmoj			Identity
	Sequence	Coordinates	Lenght	Beg.	End	Sequence	Coordinates	Lenght	
1	AY551073	584 - 820	237	+132	-105	scaffold_6540	20728731 - 20728973	243	76.1
2	AY551073	1222 - 1379	158	-606	-449	scaffold_6540	20729107 - 20729255	149	73.1
3	AY551073	1704 - 1827	124	-124	-1	scaffold_6540	20729755 - 20729866	112	80.6
<i>CG13617</i> 5' UTR	AY551073	1828 - 1945	118	+1	+118	scaffold_6540	20729867 - 20729971	105	71.5

stretches of nucleotides include the motif CAAC (or its reverse sequence GTTG). A search for TFBSs with the MATCH  software revealed that positions -20 to -8 with respect to *CG13617* TSS could correspond to a Rfx binding site. This sequence adjusts perfectly to the human Rfx binding consensus (5' GTNRCC/N-N_{0,3}-RGYAAC 3') (EMERY *et al.* 1996), and, interestingly, this transcription factor is known to participate in the regulation of genes involved in DNA replication (LIU *et al.* 1999, OTSUKI *et al.* 2004). In addition, the alignment of these sequences allowed the identification of the putative *CG13617* TSS in the other *Drosophila* subgenus species since it seems to be located in a sequence highly conserved in all of them (positions -1 to -4 are TGAC in the five analyzed species) (FIGURES 30 and 31).

Finally, sequences downstream of gene *CG13617* were also analyzed using VISTA . These sequences present a higher level of conservation and significantly-conserved non-coding regions practically covering the whole intergenic region between *CG13617* and *Pp1α-96A* can be detected also in more distant species like *D. virilis* and *D. grimshawi*. However, again none of these sequences are conserved in any of the species belonging to the *Sophophora* subgenus (results not shown).

DISCUSSION

*'If the giving of information is to be the cure of your inquisitiveness, I shall spend all the rest of my days in answering you. What more do you want to know?'
'The names of all the stars, and of all living things, and the whole history of (...) earth (...). Of course! What less? But I am not in a hurry tonight.'*

– J.R.R. TOLKIEN, *The lord of the rings. The two towers.* (1955)

In this work we have investigated the molecular mechanisms responsible for the adaptive significance of inversions. More specifically, we have tested for position effects caused by inversion breakpoints, i.e. changes in the expression of nearby genes, by studying a particular case: the polymorphic inversion *2j* in *D. buzzatii*. This is a widespread inversion that increments adult body size at the same time that it slows down development (BETRÁN *et al.* 1998). The detailed characterization of the breakpoints of inversion *2j* (CÁCERES *et al.* 2001) allowed the precise identification of a putative ORF located only a few base pairs away from the proximal breakpoint, which provided a unique opportunity to determine if either the inversion breakpoints or the TE insertions at the junctions, had modified the expression of the gene in any manner that could have contributed to the evolutionary success of this inversion.

4.1 Position effect of inversion *2j* on *CG13617* gene expression

CG13617 was first described in the genome of *D. melanogaster* as a potential ORF located between genes *nAcRβ-96A* and *Pp1α-96A*, previously thought to be the closest genes to inversion *2j* proximal breakpoint in *D. buzzatii*. In this work, we have characterized the *CG13617* ORF in the latter species and demonstrated that it is a fully functional gene expressed through the entire life cycle in embryos, larvae, pupae and adults (FIGURE 3 of PUIG *et al.* 2004, see Results). *D. buzzatii* *CG13617* is transcribed into a 2.3-kb mRNA that encodes a 734-aa protein and it is situated adjacent to the proximal breakpoint of inversion *2j*, immediately outside the inverted segment, in region D according to CÁCERES *et al.* (1999, 2001). The accurate annotation of *CG13617* coding region showed that the stop codon of this gene is located only 12 bp away from the breakpoint, and the experimental characterization of

CG13617 mRNA revealed that actually its 3' UTR spans the breakpoint and that there are at least two different bases in the 3' end of the mRNA between *2st* and *2j* arrangements.

In order to check for gene expression differences between inverted and non-inverted chromosomes, *CG13617* expression levels were compared between *2st* and *2j* homozygous lines in four different developmental stages: embryos, larvae, pupae and adults. A 5-fold (80%) reduction in *CG13617* expression level was found in 0-12 h old embryos from *2j* lines with regard to *2st* ones (PUIG *et al.* 2004), but a mixture of larvae of different ages did not show any significant difference between both arrangements, as neither did pupae nor adult stages (FIGURE 3 of PUIG *et al.* 2004). When, during the course of this work, *CG13617* gene expression was measured again in embryos 0-20 h old (that include older embryos than the first set of embryonic samples) from a new selection of *D. buzzatii* lines (some of which had not been analyzed when *CG13617* silencing was first discovered), *CG13617* mean expression

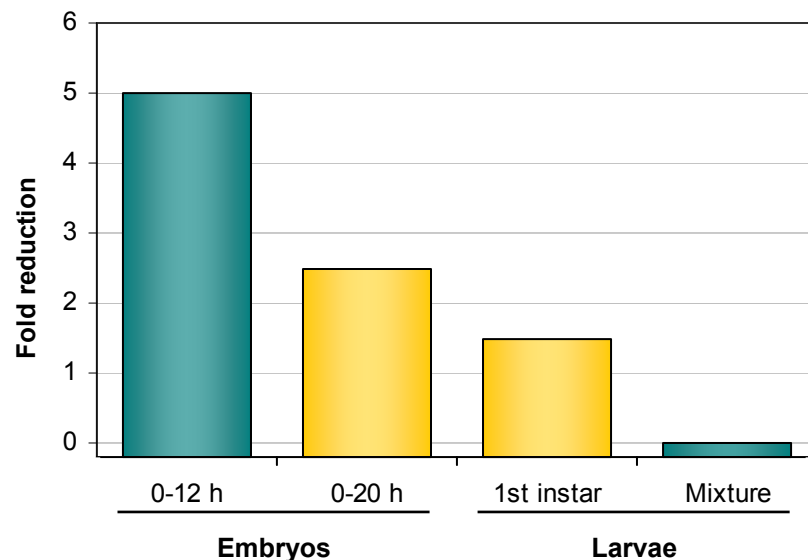


FIGURE 32 | Fold reduction of *CG13617* expression level in different developmental stages of *D. buzzatii* *2j* lines. Below each column are indicated the ages of the embryos and the types of larvae where this expression data were obtained from. In the mixture of larvae of different ages (first to third instar) no differences were found between *2st* and *2j* lines and thus fold reduction is equal to zero. Blue and yellow colors indicate respectively the samples used for the initial detection of *CG13617* silencing (PUIG *et al.* 2004) and for the real-time RT-PCR experiments performed to measure the expression level of the differentially expressed genes. The bar for each sample starts slightly below 0 to allow visualization of the mentioned case where no expression changes were detected.

was ~2.5 times higher in *2st* embryos, which represents a reduction of 60% in *2j* individuals (FIGURE 26c). In these same lines, first instar larvae also presented different expression levels, but the mean values were only ~1.5 times higher in *2st* lines (or 30% decreased in *2j* lines) when compared to lines carrying the inversion (FIGURE 26c). Therefore, even though there is a clear reduction in *CG13617* expression level that affects *2j* embryos, the magnitude of the expression difference between *2st* and *2j* lines depends on the age of the individuals included in the analyzed samples (FIGURE 32) and this difference declines as age increases between embryonic and larval samples, where gene expression is restored to equivalent levels in both chromosomal arrangements. However, the fact that unfortunately we could not study the same lines in each part of the work can also influence the mean *CG13617* expression values obtained for each arrangement.

4.1.1 Possible causes of *CG13617* silencing

Several causes have been explored as possible explanations for this decrease in *CG13617* expression in *2j* embryos. One possibility is that the insertion of TEs right after the stop codon of the gene could have modified the 3' UTR of *CG13617* mRNA or the functional sequences downstream of the coding region, impairing transcription termination and/or mRNA processing (PROUDFOOT *et al.* 2002). UTRs are known to play crucial roles in the post-transcriptional regulation of gene expression, including modulation of the transport of mRNAs out of the nucleus and of translation efficiency, their subcellular localization and their stability (MIGNONE *et al.* 2002), so any change in these sequences can have functional consequences for the affected transcript. Particularly, nucleotide patterns or motifs located in UTRs can form secondary structures or interact with specific RNA-binding proteins and complementary non-coding RNAs, all of which can play key regulatory roles. The importance of UTRs in regulating gene expression is emphasized by the finding that mutations that alter the UTRs can lead to serious diseases (CONNE *et al.* 2000). For a specific gene, the length and sequence of the 3' UTR will be defined by the processing of the mRNA 3' end.

Several processing events are needed to produce a mature mRNA ready to be exported out of the nucleus and translated. These comprise the acquisition of a 7-methyl-

guanylate (m⁷G) cap structure at the 5' end, the splicing of introns, and the generation of a 3' end, usually modified by the addition of a stretch of 100-250 adenine residues (the polyA tail). Although each of these reactions is a biochemically distinct process, they are all interconnected and influence one another's specificity and efficiency. Furthermore, all these events occur while transcription is taking place, so transcription itself can also affect or be affected by mRNA processing. Specifically, the mRNA 3' ends of protein-coding genes are generated by coupled cleavage and polyadenylation. These two processes take place as RNA polymerase II (Pol II) proceeds through the transcription of the pre-mRNA, and are both dependent on the presence of the C-terminal domain (CTD) of Pol II, which is positioned outside the overall globular tridimensional structure of this enzyme and interacts with components involved in these RNA processing mechanisms directly activating the reactions. The site of cleavage in most pre-mRNAs lies between the highly conserved AAUAAA hexamer (polyA signal) and a downstream sequence element (DSE), which is a U- or GU-rich motif. Cleavage occurs predominantly at a CA dinucleotide situated ~10–30 nt downstream from the polyA signal (PROUDFOOT *et al.* 2002, PROUDFOOT 2004). Several factors are involved in this processing reaction (see FIGURE 33 for additional details). In addition to the 3' end processing of the mRNA, the Pol II complex must be halted and released to allow its recycling to transcription initiation at the promoter. This stage is critical for successful gene expression because it releases the mature transcripts from the transcription site and prevents read-through transcription that may perturb the proper expression of downstream genes. Even though this process is not completely characterized, the current model states that there are two independent events that are required for transcription termination: the 3' end processing of the mRNA and the cotranscriptional cleavage of certain transcribed sequences downstream the polyA signal (PROUDFOOT *et al.* 2002). Both sequence elements located downstream of the polyA site and specific factors are needed for termination. Sequence elements include transcriptional pause sites able to cause a transient pause to the Pol II progression that enhances polyA signal recognition, as well as sequence tracts that promote the heterogeneous cleavage of the nascent transcript (DYE and PROUDFOOT 2001). Also, several transcription factors have been found to interact with certain components of the 3' end processing machinery, which suggests a connection between the initiation and termination stages of transcription. These factors could be involved in setting up an appropriate chromatin structure to facilitate transcriptional elongation and termination (PROUDFOOT 2004).

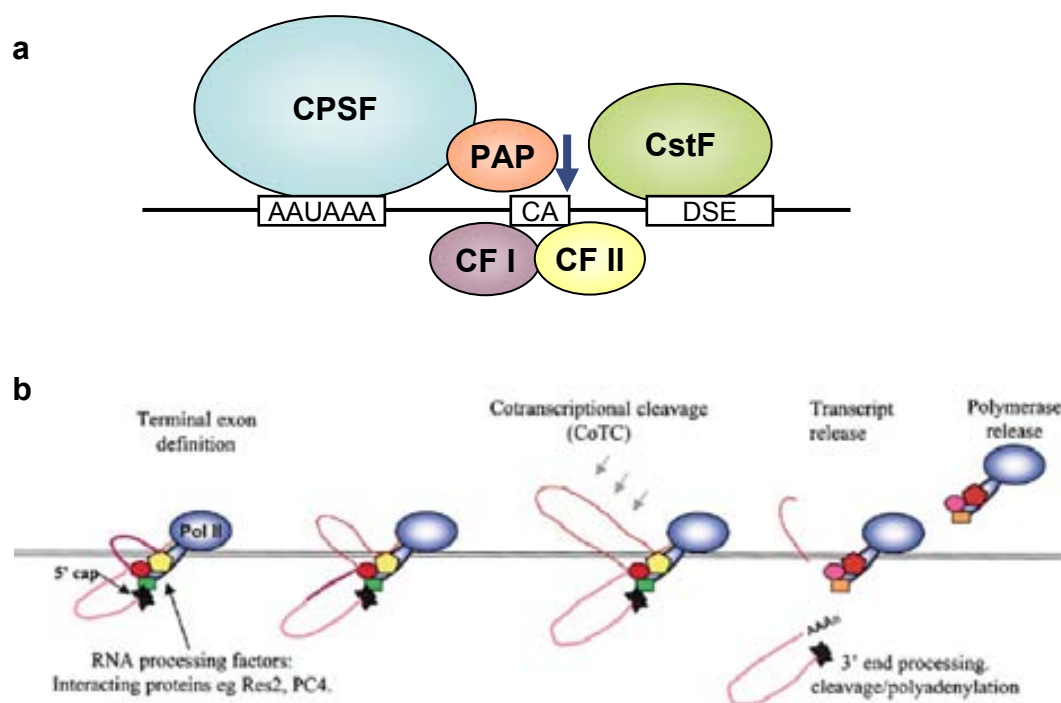


FIGURE 33 | Factors and sequence elements required for mRNA 3' end processing and transcription termination. (a) Protein complexes involved in mRNA 3' processing. CPSF (cleavage and polyadenylation specificity factor) and CstF (cleavage stimulatory factor) protein complexes interact with the polyA signal and the DSE respectively in the pre-mRNA molecule. Cleavage factors I and II (CF I and II) are also essential to direct cleavage. PAP (polyA polymerase) is also required for the cleavage reaction, and together with CPSF performs polyA addition. Finally, PABP II (polyA binding protein) binds the emerging polyA tail and enhances the processivity of PAP. The exact molecular nature of the endoribonuclease activity that performs the cleavage of the pre-mRNA remains unknown, but it is probably located within a subunit of CPSF. Some of these proteins interact with Pol II CTD, which has not been represented here but is also required to complete these reactions. (b) Current model of the processes that lead to transcription termination and the release of both Pol II and the mature transcript. Panel a is based on Figure 1 from PROUDFOOT (2004) and panel b is extracted from PROUDFOOT *et al.* (2002).

In *D. buz̄z̄atii* *CG13617* gene, both processes (3' end processing and transcription termination) could be compromised by the TE insertions at the breakpoint. Two putative polyA signals identical to the consensus sequence AAUAAA were found at the end of *CG13617* in the sequenced *2st* line: one overlapping the stop codon of the gene and another one located 15 bp downstream of the stop codon (FIGURE 34). In *2j* chromosomes, large TE insertions are found just 12 bp downstream of the stop codon and the contiguous single-copy sequence has been moved to a distant region of the chromosome by the inversion *2j*. As a consequence, the second putative polyA site has been removed in lines carrying the inversion.

The identification of the 3' end of *CG13617* mRNA in one *2st* and one *2j* line revealed that the 3' UTR of the transcript produced by this gene is extremely short, extending only 17 bp after the coding region (the stop codon is considered part of the 3' UTR) in the analyzed *2st* line and 19 bp in the *2j* line (FIGURE 34). Therefore, *CG13617* mRNA ends in non-inverted chromosomes 2 bp after the point of insertion of the TEs (i.e., 2 bp corresponding to the single-copy sequence inside the inverted segment are actually transcribed in *2st* lines) whereas it includes 4 bp of the *Galileo* TE inserted at the breakpoint junction in inverted chromosomes. This suggests that the first polyA signal is used for the processing of these transcripts and indicates that, in fact, the TE insertions took place inside *CG13617* 3' UTR. Thus, the alteration of *CG13617* 3' UTR occurred already when TEs first inserted (originally in a *2st* chromosome where inversion *2j* would later arise by ectopic recombination between two oppositely oriented *Galileo* copies located at both breakpoint regions) and should not depend on the orientation of the inverted segment. Furthermore, any putative downstream sequence elements required for transcription termination would also have been replaced in *2j* chromosomes by sequences located inside the TEs.

When compared to other genes, *CG13617* mRNA possesses a really short 3' UTR since the average length for 3' UTRs in invertebrates is 444.5 bp (MIGNONE *et al.* 2002). We also searched for longer transcripts that might be generated using the second polyA signal in the *2st* lines, or alternative sequences provided by the inserted TEs in the *2j* lines for their 3' end processing, in both arrangements by RT-PCR using specific primers for each line located further downstream of the gene. No evidence that these transcripts exist in embryos was found (results not shown), but longer transcripts could be amplified and sequenced in adult individuals both for *2st* and *2j* lines. In *st-1* line, this alternative 3' UTR is 45-bp long, extending 21 bp downstream of the second polyA signal, which suggests that this site is probably the one being used for the 3' processing of this transcript (FIGURE 34). In *j-1* line, the longer mRNA has a 49-bp long 3' UTR that includes 34 bp of the TE *Galileo* inserted at the breakpoint junction (FIGURE 34), but it remains unclear which are the sequence elements involved in the processing of this longer transcript. So, the second polyA signal, absent in *2j* chromosomes, is likely used in the *2st* arrangement to produce a longer mRNA, but only in the adult stage, and not in embryos. Also, *2j* chromosomes manage to transcribe an mRNA of similar length (although different 3' UTR sequence) to that found in the *2st* line in the adult

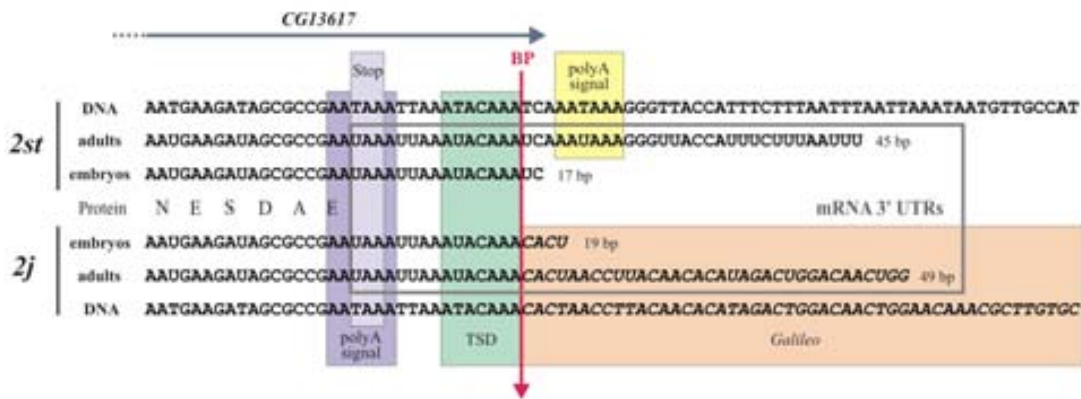


FIGURE 34 | Nucleotide sequences of the 3' UTRs of *CG13617* mRNAs in lines *st-1* and *j-1*. The stop codon, the two polyA signals, the sequence duplicated by the TE insertion in *2j* chromosomes (TSD) and the *Galileo* element are included in colored boxes. The vertical arrow indicates the *2j* inversion proximal breakpoint (BP). 3' UTR sequences are enclosed in a grey rectangle and their lengths are indicated. Sequences present only in lines with the inversion are shown in italics. The developmental stage where these transcripts were isolated is shown at the left of each sequence. Figure modified from PUIG *et al.* 2004.

stage using some unknown cryptic signals that must be located within the *Galileo* element TIR for the 3' processing and transcription termination. Since the reduced *CG13617* expression has been detected in *2j* embryos, but not in any other developmental stage, and the longer mRNA has been detected only in adults, the absence of the second polyA signal in *2j* lines does not seem to be related to the expression change. The functional polyA signal for this gene in embryos is therefore the one overlapping the stop codon, unaltered in both chromosomal arrangements, which leaves the two final nucleotides of the mRNA as the only difference between the transcripts produced by *2st* and *2j* chromosomes during this developmental stage. Since 15 out of the 17 nucleotides of the *2st* 3' UTR are identical in the *2j* mRNA, the change introduced in the 3' UTR end by the presence of the TEs in *2j* chromosomes is probably too small to affect the 3' processing of the transcript, its stability or any possible regulatory binding sites the 3' UTR might contain (MIGNONE *et al.* 2002). It does not seem likely either that the addition of only two nucleotides and the change of two more nucleotide positions in a 3' UTR as short as this one could have any functional impact on *CG13617* expression. However, the functional connection between transcription initiation and termination still makes it possible that a less efficient termination process caused by the

change in the sequences immediately downstream *CG13617* polyA signal might translate into a reduced transcription rate for this gene (PROUDFOOT *et al.* 2002).

Another alternative hypothesis to explain *CG13617* expression change could be a silencing effect of the large repetitive DNA blocks inserted at the breakpoint, similar to that exerted by heterochromatin in adjacent sequences (GIRTON and JOHANSEN 2008). Repetitive DNA blocks at the proximal breakpoint are 3.2-6.3 kb long in the different analyzed *2j* chromosomes (CÁCERES *et al.* 2001) and they are made up of several TE insertions that occurred inside the original *Galileo* copy in a short period of time. These breakpoint junction sequences have become hotspots for multiple structural changes besides the mentioned TE insertions, such as deletions, duplications and additional microinversions (CÁCERES *et al.* 2001). This complex structure reminds that of heterochromatic regions (PIMPINELLI *et al.* 1995), which are known to induce the silencing of euchromatic genes when these are relocated close to them (either by chromosomal rearrangements or *P* element transformation), causing a phenomenon known as position-effect variegation (PEV) in *Drosophila* (GIRTON and JOHANSEN 2008). PEV is the result of the inactivation of a gene in some cells caused by its abnormal localization near or into heterochromatin. In a study of heterochromatin formation on *D. melanogaster* chromosome 4, SUN *et al.* (2004) identified the *1360* TE as an initiator of heterochromatin formation, since the variegated expression caused by the partial silencing of a *white* transgene that was introduced in several positions along the chromosome was correlated with proximity to transposon *1360* fragments (FIGURE 6). Deletions that lead to a loss of *1360* elements can result in a switch from a variegating to a normal red-eye phenotype, while those that position the transgene close to a *1360* copy result in the opposite expression change. Their study also suggests that heterochromatin assembly spreads for ~10 kb. Interestingly, *Galileo* belongs to the *P* superfamily of DNA transposons and is related to *1360* (MARZO *et al.* 2008), so this TE may present some features that facilitate heterochromatin formation. Most of the *1360* copies analyzed in *D. melanogaster* chromosome 4 retain their TIRs and known TSSs and range in size from 453 to 1113 bp (SUN *et al.* 2004). Thus, short stretches of repetitive elements seem to be sufficient to induce heterochromatin formation and its expansion into the adjacent sequences, and probably the TE insertions at inversion *2j* breakpoints, which contain several *Galileo* copies, are large enough to be able to behave like a heterochromatic site in this respect. However, if heterochromatinization of the breakpoint

insertions was the cause of *CG13617* silencing, we would expect *Pp1 α -96A* expression to be affected as well, since the initial methionine of this gene is located only 651 bp away from the distal breakpoint insertions in *2j* chromosomes (its transcription initiates even closer to the repetitive DNA, FIGURE 10) and the silencing effect caused by heterochromatin has been shown to extend through several genes (LEWIS 1950, WEILER and WAKIMOTO 1995). *Pp1 α -96A*, located inside the inverted segment, is closer to the distal breakpoint insertions in *2j* chromosomes, where TEs form slightly smaller insertions (1.2-3.4 kb long) than those at the proximal breakpoint (3.2-6.3 kb long), but in this gene the long repetitive DNA blocks are found next to the promoter and could easily affect gene regulation through several mechanisms besides heterochromatinization. Nevertheless, the expression level of *Pp1 α -96A* in embryos seems to be equivalent for *2st* and *2j* lines (FIGURE 4 of PUIG *et al.* 2004). In addition, the expression of *Rox8*, the gene located outside the inversion at the distal inversion breakpoint in region A, had been previously analyzed by Northern blot (CÁCERES *et al.* 1999) and no detectable differences were observed either. Finally, the silencing effect of heterochromatin would probably reduce gene expression in all developmental stages where *CG13617* is transcribed, and not only in embryos, as we have observed (FIGURE 3 of PUIG *et al.* 2004). These evidences suggest that not all repetitive elements can serve as initiators of heterochromatin formation.

A third possible cause for *CG13617* down-regulation could be the separation of the gene from a 3' enhancer required to achieve the proper level of gene expression, or to express this gene in a particular tissue. The presence of downstream enhancers has been demonstrated in several cases. For example, in humans, an enhancer located 1.4 kb downstream of the *c-Myc* proto-oncogene stop codon that is able to bind β -catenin provides the principal mechanism of regulation of the expression of this gene (YOCHUM *et al.* 2008) and the deletion of a 800-bp enhancer located 200 kb downstream of gene *SHOX* (short-stature homeobox-containing gene), involved in skeletal development, is sufficient to cause Léri-Weill dyschondrosteosis (a form of dwarfism) in individuals with an intact coding region of the gene (FUKAMI *et al.* 2006). In rats, a downstream enhancer located at position +4.4 kb from the TSS of gene *B29/Ig- β* and conserved in humans, is able to increase several-fold the level of transcription when compared to the activity of the promoter alone (KOMATSU *et al.* 2002). In *Drosophila*, *ato* transcription in proneural clusters is regulated by distinct tissue-specific *cis*-regulatory

sequences that lie within the 5.8 kb downstream of its ORF (TANAKA-MATAKATSU and DU 2008). Also, in the bithorax complex of *Drosophila*, the IAB5 enhancer is able to interact specifically with the *Abdominal-B* (*Abd-B*) promoter, even though the enhancer is located 55 kb 3' of the promoter and they are separated by at least two insulators, and directs the expression of this gene in abdominal segment 5 (AKBARI *et al.* 2008).

If a downstream enhancer was needed for the normal expression of *CG13617* in *D. buzsatii* it would be located inside the inverted segment and it would have been moved to a very distant chromosomal location by inversion *2j* (which comprises approximately one third of *D. buzsatii* chromosome 2). This change in chromosomal position would mean that the enhancer could no longer regulate *CG13617* gene expression, which, as a result would be transcribed at a lower rate or would not be expressed in a certain tissue or group of cells. Also, this hypothesis has additional implications, because the putative enhancer could be affecting the expression of other genes in its new location (if distance and orientation favored its action). The removal of such an enhancer is a possibility that has not been ruled out. However, according to our results it should be an enhancer regulating transcription rate, because *CG13617* expression pattern in embryos seems to be the same for *2st* and *2j* lines (SUPPORTING FIGURE 7 of PUIG *et al.* 2004). In order to identify putative regulatory elements downstream *CG13617* coding region, we searched for conserved non-coding regions between the stop codon of *CG13617* and the initial methionine of *Pp1 α -96A* (672 bp in *D. buzsatii*). Although the level of conservation of this intergenic region is higher than that of the region between genes *nAcR β -96A* and *CG13617*, no significant conservation was detected in multiple sequence alignments beyond the *Drosophila* subgenus group of species (results not shown). Given the proximity of the genes, even in the case that conserved sequences with potential functional roles were detected it would be impossible to distinguish between *CG13617* downstream regulatory sequences and *Pp1 α -96A* promoter elements. There is also the possibility that the putative 3' enhancer is located further away, but the identification of such a distant regulatory element would require more complex functional studies.

Yet another explanation for the decreased level of expression in *2j* lines could be the presence in these lines of a change in the promoter or the regulatory elements located upstream of *CG13617* coding region. It is possible that by chance inversion *2j* arose in a

chromosome presenting a decreased *CG13617* expression caused by a pre-existing promoter variant. In this case, the nucleotide changes responsible for the reduced expression of this gene would be linked to inversion *2j* due to its proximity to one of the breakpoints (which prevents recombination in the adjacent sequences), but the breakpoint itself would not be the primary cause of the observed change in *CG13617* expression level. To investigate this possibility, first structural variation in the *CG13617* gene region was studied in several *2st* and *2j* lines by restriction mapping of PCR products (SUPPORTING TABLE 4 of PUIG *et al.* 2004), although no differences shared by all the lines with the same arrangement that distinguish them from the alternative chromosomal organization could be detected. In fact, the only structural variation was located in the non-coding sequences upstream *CG13617* coding region within the lines carrying *2j* inversion. Sequenced line *j-1* differs from *st-1* line by the presence of a 166-bp tandem duplication and a 26-bp insertion located 1010 bp and 408 bp upstream of the *CG13617* start codon, respectively, but of the other analyzed *2j* lines only *j-19* has this same structural changes. A 900-bp insertion found in *jz³-4* in the intergenic region upstream of this gene and a few restriction site polymorphisms are the only other detected variants, neither of them shared by all inverted chromosomes. Second, we also attempted to identify *CG13617* promoter or regulatory regions by analyzing the DNA sequences upstream of the TSS in five *Drosophila* species. The 5' UTR was experimentally determined to be 118 bp long in *D. buzzatii* and a multiple alignment with the four other available four *Drosophila* subgenus sequences allowed us to define the location of the putative TSS in these species based on sequence conservation. The comparative analysis of the intergenic sequence separating the 5' end of *CG13617* and the 5' end of gene *nAcRβ-96A* revealed the presence of three conserved non-coding sequences (CNSs) in these species (FIGURE 31 and TABLE 14). CNS1 corresponds to *nAcRβ-96A* 5' UTR and promoter, CNS2 is a 158-bp region located approximately 500 bp away from the TSS of both flanking genes in *D. buzzatii*, and CNS3 is a 124-bp segment immediately upstream of *CG13617* TSS that includes several short stretches of 5-10 bp conserved in the different studied species (FIGURE 31) that may correspond to regulatory elements involved in controlling *CG13617* expression. No recognizable TATA box has been found in this analysis, but many expressed genes do not seem to possess this box in their core promoters (GERSHENZON *et al.* 2006). However, a putative promoter with DNA replication related elements bound by TRF2 (TATA-box-binding protein related factor 2) (HOCHHEIMER *et al.* 2002) can be computationally predicted in the CNS3 sequence located immediately

upstream of *CG13617* TSS, suggesting that these sequences can indeed represent the core promoter controlling the expression of this gene. It is surprising that five of these short conserved elements comprise the sequence motif CAAC (or its reverse sequence GTTG), which indicates that they might be specific sequences recognized by a transcription factor able to bind them (i.e. TFBS). In fact, positions -20 to -8 with respect to *CG13617* TSS (position +1), which include two CAAC motifs in opposite orientation, have been identified bioinformatically as a putative binding site for transcription factor Rfx, also known to bind the promoter of DNA replication gene *mus209* in *Drosophila* (OTSUKI *et al.* 2004) and its homologous gene *PCNA* in humans (LIU *et al.* 1999). Remarkably, a putative Rfx binding site is also found in *CG13617* upstream sequences in *D. melanogaster* (150 bp upstream the initial methionine) and other closely related species (results not shown), although it can not be identified using the same method in all of the 12 sequenced genomes. A multiple alignment of the 12 *Drosophila* genomes also fails to detect any of these motifs as conserved elements because they are small and slight variations in some species are sufficient to cause a misalignment in sequences as divergent as these ones that does not allow their recognition as conserved elements. However, the presence of a binding site for the same transcription factor in species as phylogenetically distant as *D. melanogaster* and *D. buzzatii* suggests that this can be a functional sequence involved in the regulation of *CG13617* gene expression through the action of the Rfx transcription factor.

When the sequences of both *D. buzzatii* chromosomal arrangements were compared, only two nucleotide changes between *2st* and *2j* sequenced lines were found in the CNS identified in the 5' region of *CG13617* (FIGURE 31), and one of this nucleotide changes seems to have occurred in the *st-1* line (results not shown). Besides, none of these changes is found in the short conserved sequences that may be candidates to act as regulatory elements of this gene. Therefore, even though variation exists in the region upstream *CG13617* coding region, none of the detected changes can be correlated with the expression difference. Thus, although the hypothesis that a promoter variant linked to inversion *2j* is causing the reduced expression level does not seem likely, it has to be taken into account that the precise regulatory elements of this gene (promoters, enhancers, etc.) have not been identified yet and they can be very distant from the coding region of the gene. So, it still remains possible that variation in one of

these elements could play a role in causing a lower expression of *CG13617* if it was present in all *2j* chromosomes but not in *2st* ones.

4.1.2 *CG13617* antisense transcript

On the other hand, our results indicate that the most likely silencing mechanism for *CG13617* involves an antisense RNA overlapping the whole coding region of this gene discovered in *2j* embryos. This antisense transcript was detected due to the repeated amplification of a DNA-sized band in *2j* lines (together with the expected spliced band) in RT-PCR experiments with primers located in different exons. After characterizing this RNA and analyzing its expression, several evidences point to the antisense transcript to be responsible for *CG13617* decreased expression in *2j* embryos. First, the reduction of *CG13617* expression occurs only in *2j* embryos, which have the highest amounts of antisense RNA. No differences in the expression level of this gene were observed in other developmental stages, where the antisense RNA was found only in very small quantities, probably insufficient to cause gene silencing. Second, the antisense RNA expression level is negatively correlated with *CG13617* mRNA expression, presenting a 5-fold increase in *2j* embryos when compared to *2st* embryos that is accompanied by the 5-fold decrease detected for *CG13617* transcript. Third, the antisense RNA expression level has an intermediate value in *2st/2j* heterozygotes whereas *CG13617* mRNA expression level is similar to that in *2j* lines, which suggests a *trans* effect able to affect the two copies of the gene even though in heterokaryotypes the antisense molecule is transcribed only from one of them (the *2j* chromosome). Finally, the higher expression of the antisense RNA is the only feature shared by all *2j* lines that distinguishes them from *2st* lines, apart from the TEs inserted at the breakpoint junction and the *2j* inversion itself.

The molecular characterization of the antisense RNA revealed that it is transcribed from the TEs inserted at the breakpoint junction in *2j* chromosomes, and more specifically from a *GalileoK* copy present in all the analyzed *2j* lines. This antisense transcript has an estimated length of ~3 kb and includes the complete coding region of *CG13617*, as well as all of its introns (FIGURE 1 of PUIG *et al.* 2004). The antisense RNA is probably a non-coding transcript since its sequence contains only small ORFs (53-143 aa) without homology to any

known protein. It also seems to be unspliced (i.e., does not have any introns of its own) and polyadenylated (according to KATAYAMA *et al.* (2005) many highly expressed polyA⁻ RNAs have been identified among the antisense and non-coding RNAs in the mouse FANTOM3 database).

Antisense transcripts able to down-regulate gene expression have been reported in an increasing number of genes in many different species ranging from yeast to plants and animals (LAPIDOT and PILPEL 2006). In humans, it has been reported that 22% of transcriptional units form sense-antisense pairs (CHEN *et al.* 2004), in mouse almost 29% of all mapped transcriptional units overlap with a cDNA in the opposite strand (KATAYAMA *et al.* 2005), while annotations of the *D. melanogaster* genome identified sense-antisense pairs in 15-17% of the genes (MISRA *et al.* 2002, SUN *et al.* 2006). There are also well-characterized examples of antisense transcripts with regulatory function. For example, in *Neurospora crassa* regulation of the circadian clock, which is critical for the correct temporal expression of genes, is controlled partly by two antisense RNAs in the gene *frq*, whose cyclic expression is a prerequisite for rhythmicity (KRAMER *et al.* 2003). These antisense *frq* RNAs are ~5 and 5.5 kb long and non-coding and in the dark, their expression levels cycle with the opposite phase to sense transcripts (FIGURE 35a). Sense and antisense transcripts are inducible by light, but when the antisense RNAs promoter is eliminated and their expression abolished, several alterations in the circadian clock take place, revealing that these antisense *frq* transcripts have an essential role in setting the phase of the circadian clock, even though the exact mechanism by which they exert their effects has not been determined yet. Also, in the snail *Lymnaea stagnalis* a nitric oxide synthase (NOS) pseudogene that includes a region of significant antisense homology to a neuronal NOS (nNOS)-encoding mRNA, is expressed in the central nervous system. Both transcripts are co-expressed in certain neurons and form stable dsRNA duplexes *in vivo*. As a consequence, the antisense region of the pseudogene transcript prevents the translation of nNOS protein from the nNOS-encoding mRNA, showing that expressed pseudogenes can act as regulatory elements and that a natural antisense RNA can mediate the translational control of nNOS expression in *Lymnaea* (KORNEEV *et al.* 1999).

Antisense RNAs have also been involved in genomic imprinting, a phenomenon where one of the two parental alleles of an autosomal gene is silenced epigenetically by a *cis-*

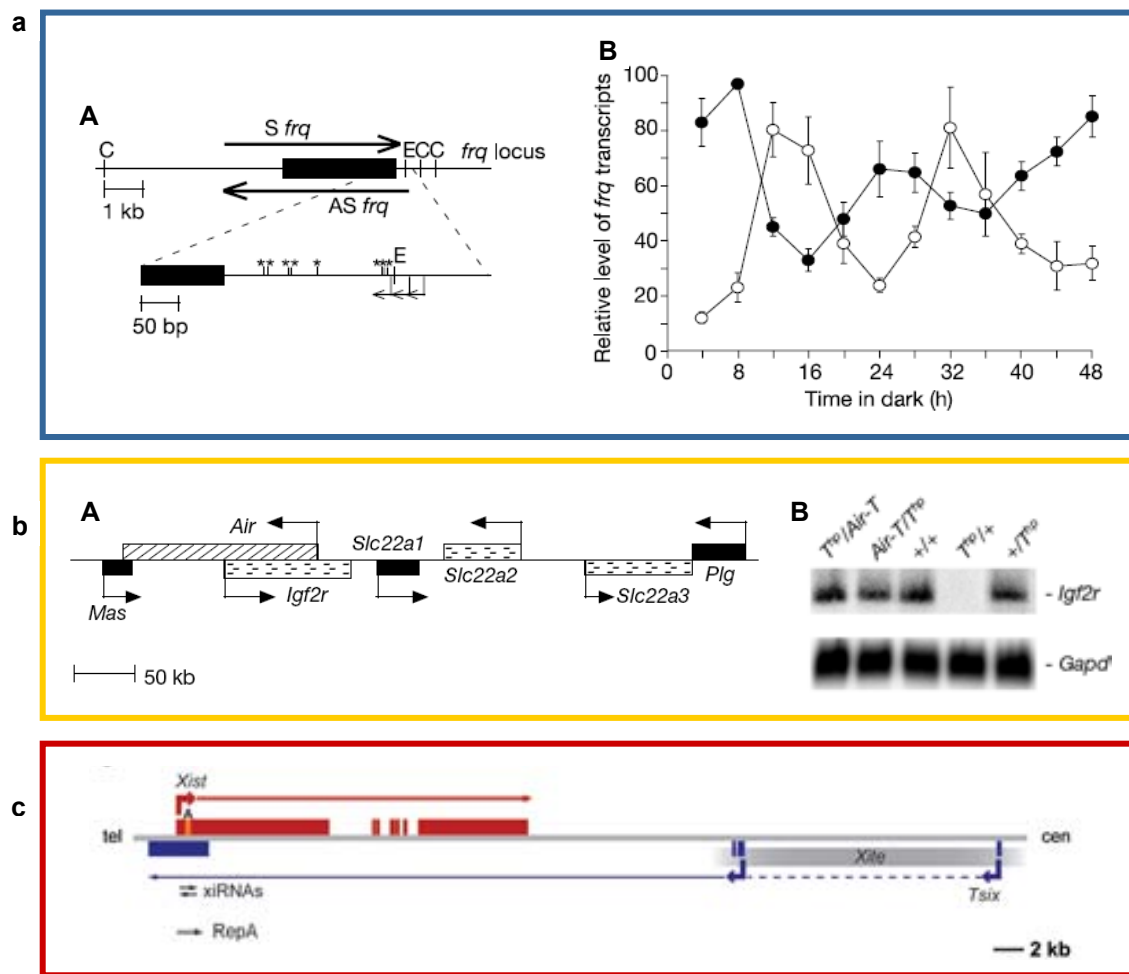


FIGURE 35 | Examples of natural antisense RNAs with regulatory functions. (a) *frq* locus in *Neurospora crassa* (KRAMER *et al.* 2003). In panel A, black boxes indicate *frq* ORF and thick black arrows represent the sense (S) and antisense (AS) *frq* transcripts. Details of the region downstream of the *frq* ORF are shown below: the asterisks correspond to polyadenylation sites of the sense *frq* mRNA, and small arrows to the transcription start points of the antisense *frq* transcripts. E and C mark restriction sites. Panel B shows the cycling of *frq* transcript levels in the dark. Open circles correspond to the sense *frq* mRNA levels and black circles to the antisense *frq* RNA. **(b)** Map spanning the 400-kb imprinted cluster regulated by the *Air* transcript (panel A). Paternally expressed *Air* is represented as a hatched box, maternally expressed genes *Igf2r*, *Slc22a2* and *Slc22a3* as dotted boxes and non-imprinted genes as black boxes (SLEUTELS *et al.* 2002). In panel B, *Igf2r* expression was analyzed by RNase protection assays in individuals carrying several combinations of a truncated copy of *Air* (*Air-T*), a deletion of the whole imprinted cluster (*T^{tr}*) and a wild-type *Air* allele (+). For each genotype the maternal allele is indicated first. When the maternal allele is deleted (*T^{tr}*) no *Igf2r* expression is detected (lane 4), which indicates that the paternal allele is normally repressed. When a truncated *Air-T* is present in the absence of a maternal allele (lane 1), *Igf2r* is expressed from the paternal allele, which suggests that *Air* RNA is essential for the genomic imprinting of this gene. The same results were obtained for genes *Slc22a2* and *Slc22a3* (SLEUTELS *et al.* 2002). **(c)** Representation of the RNAs produced in the *Xist/Tsix* locus. *Xist* is shown in red and the antisense RNA *Tsix* in blue. Other non-coding RNAs have been described in this locus that could be implicated in the regulation of X inactivation. *RepA* RNA and xiRNAs (depicted as small arrows) have been characterized as RNA species overlapping the repeat A element (in yellow) and *Xite* transcripts (grey zone) originate from a region between the next gene *Tsix* (not depicted) and the *Tsix* major promoter (LEEB *et al.* 2009). Figures extracted in each case from the corresponding cited articles.

acting mechanism. In mouse chromosome 17, the paternal expression of *Air* RNA is correlated with repression of the paternal alleles of three protein-coding genes (*Igf2r*, *Slc22a2* and *Slc22a3*), so that these imprinted genes are exclusively maternally expressed. The *Air* transcript is a 108-kb long, non-coding and polyadenylated RNA that is apparently unspliced and overlaps the *Igf2r* promoter (FIGURE 35b). A truncated *Air* allele maintains imprinted expression and methylation of the *Air* promoter, but shows a complete loss of silencing of genes *Igf2r*, *Slc22a2* and *Slc22a3* on the paternal chromosome (FIGURE 35b), which suggests an active role of this RNA in the genomic imprinting of this region (SLEUTELS *et al.* 2002). The silencing of the two genes that do not overlap with the *Air* transcript probably occurs by the spreading of a silent chromatin state induced by the bidirectionally transcribed sequences. Another example of functional antisense RNAs is found during X chromosome inactivation in female mice (LEEB *et al.* 2009). *Xist* is a 17-kb non-coding RNA transcribed from the *XIC* (X inactivation center) *locus* that is required for the initiation of X inactivation (although not for the maintenance of this silencing). It is expressed exclusively from the X copy that will be inactivated, where it accumulates coating the whole chromosome and, by recruiting several proteins, induces chromatin changes that silence transcriptionally this chromosome. *Tsix* is an antisense transcript that overlaps *Xist* transcription unit and that is expressed in the active X chromosome (FIGURE 35c). It acts as a repressor of X inactivation *in cis*. However, it remains unknown if *Tsix* contributes to the repression of *Xist* by acting as a functional RNA or whether transcription over the *Xist* promoter facilitates its repression through epigenetic changes. Finally, antisense RNAs have also been implicated in human disease (TUFARELLI *et al.* 2003), TE silencing (SIJEN and PLASTERK 2003, CHUNG *et al.* 2008) and defense against viruses (OBBARD *et al.* 2009).

Several mechanisms have been described to explain the effect of an antisense transcript on the expression of the corresponding sense mRNA, also known as antisense regulation (LAPIDOT and PILPEL 2006, MUNROE and ZHU 2006). Some of these mechanisms require the formation of dsRNA molecules by the specific pairing of sense-antisense pairs (FIGURE 36a). For example, the formation of dsRNA could block the binding of proteins to the target sense transcript, a process known as RNA masking (MUNROE and ZHU 2006). The presence of a complementary transcript could therefore prevent the binding of proteins involved in the splicing, export, stability or translation of the sense mRNA and difficult these

processes, thus affecting the final amount of available mRNA (HASTINGS *et al.* 1997). Regulation could also occur by the well-known RNA interference (RNAi) mechanism, where dsRNA molecules are cleaved into discrete 21-23 nt fragments known as small interfering RNAs (siRNAs) that act by targeting the destruction of homologous single-stranded RNAs or by repressing translation (see BOX 2). While this pathway is usually considered a post-transcriptional silencing mechanism (BORSANI *et al.* 2005, DUHRING *et al.* 2006), there is also evidence that some components of the RNAi pathway are involved in the creation of epigenetic alterations through DNA methylation (IMAMURA *et al.* 2004) or chromatin remodeling that could lead to gene silencing at the transcriptional level by RNA-dependent heterochromatin assembly (CARTHEW and SONTHEIMER 2009). This last mechanism of transcriptional regulation is especially important when antisense RNAs overlap the sense promoter (ANDERSEN and PANNING 2003, IMAMURA *et al.* 2004).

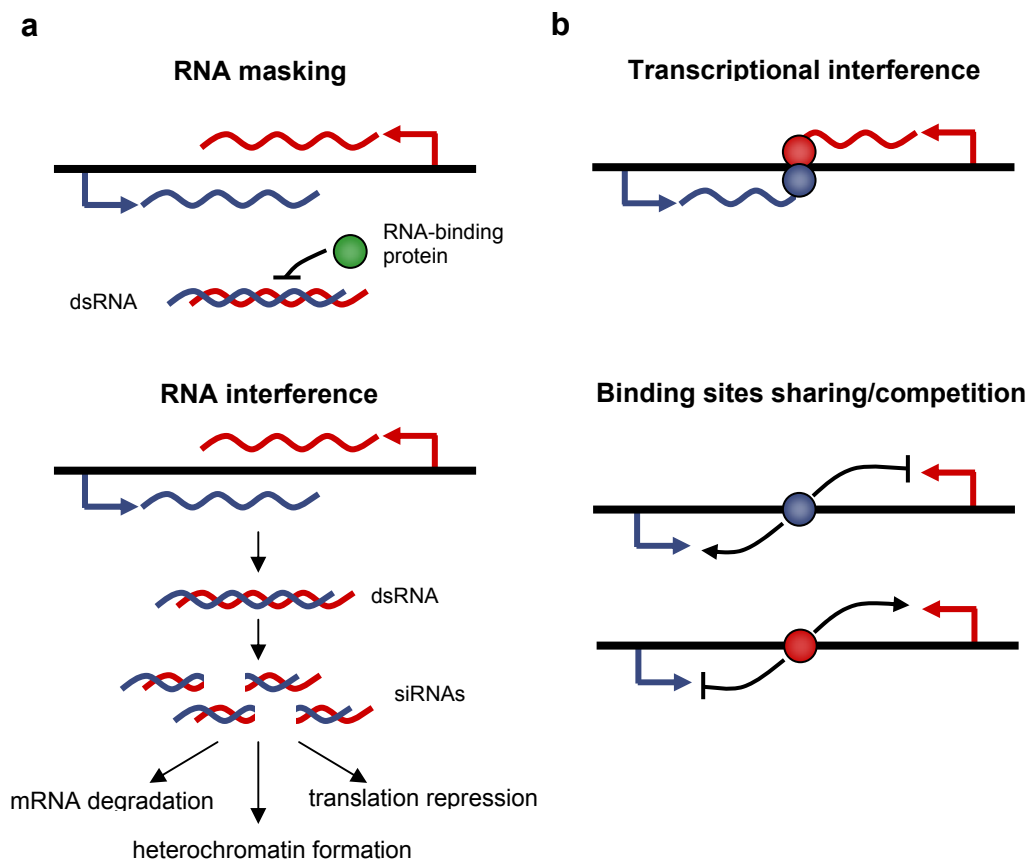


FIGURE 36 | Mechanisms of antisense regulation. (a) Mechanisms that involve sense-antisense pairing. (b) Mechanisms that do not require dsRNA formation. Blue arrows and lines correspond to sense transcripts and red arrows and lines represent antisense RNAs. Colored circles represent regulatory proteins. Figure redrawn from MUNROE and ZHU (2006).

However, there are also other regulatory mechanisms that do not involve dsRNA formation (FIGURE 36b). This is the case of transcriptional interference, where transcription from one DNA strand hinders transcription of the opposite strand, especially in convergent 3' overlapping sense-antisense pairs (PRESCOTT and PROUDFOOT 2002, MUNROE and ZHU 2006). In transcriptional interference, the progression of the machinery transcribing one strand of DNA can prevent the elongation of the transcript on the complementary strand, or, alternatively, interacting elongation complexes from the sense and antisense promoters can cause premature termination of one or both complexes (MAZO *et al.* 2007). Experimental evidence of sense or antisense transcriptional interference has been found in different genes and species (MAZO *et al.* 2007). This is the case of the *IME4* gene in *Sacharomyces cerevisiae*, whose expression is repressed by an antisense transcript only when transcription of both molecules occurs *in cis*, ruling out RNA interference mechanisms (HONGAY *et al.* 2006), or the untranslated RNAs regulating genes of the *Drosophila* bithorax complex of homeotic genes (PETRUK *et al.* 2006). Another antisense regulation mechanism that does not require dsRNA involves the sharing of binding sites and the competition for transcription factors. This mechanism is particularly important in gene pairs with juxtaposed promoters (PAULER *et al.* 2005), but it can also affect other kinds of sense-antisense pairs because it is known that regulatory elements can be located in introns as well as in the sequences downstream of a gene.

In the case of gene *CG13617*, the extensive overlap of the two transcripts (mRNA and antisense RNA) with ~2.3 kb of shared sequence, makes possible that the two complementary transcripts bind each other to form long stretches of perfectly matched dsRNA in *2j* embryos. Indeed, we have experimentally tested the formation of dsRNA in the case of the *CG13617* sense-antisense pair. After digestion of single-stranded RNA, fragments of the coding region of the gene present in both transcripts could still be amplified by RT-PCR in *2j* embryos, although not in *2st* ones, where the antisense RNA is found at very low levels and no dsRNA is formed that can protect *CG13617* mRNA from degradation (FIGURE 6 of PUIG *et al.* 2004). This indicates that the two complementary transcripts are able to bind each other but it is not a proof that this happens *in vivo*, which would require the two transcripts to be expressed simultaneously in the same cells. The detection of small fragments of 21-23 nt corresponding to siRNAs with *CG13617* sequence would indicate unequivocally that the silencing is being

triggered by the dsRNA through the RNAi pathway. Meanwhile, other mechanisms like transcriptional interference can not be completely ruled out in this case. In fact, more than one antisense regulatory mechanism could be affecting *CG13617* expression in *2j* embryos. For example, since both mRNA and antisense RNA are transcribed from the same *locus*, it is likely that some type of transcriptional interference plays a role in the detected reduced expression level. Transcriptional interference depends on the relative strength of the two promoters driving transcription of the target gene and the regulatory RNA. Thus, a strong promoter controlling antisense transcription alone could cause a significant reduction in the amount of sense transcript, if they were both being transcribed at the same time in the same cells. If the two promoters were equally efficient and enough quantities of both transcripts were produced in the same cells, then they could form long dsRNA molecules able to trigger other silencing mechanisms, such as the RNAi pathway, which could act by degrading *CG13617* mRNA or through the induction of heterochromatin formation. Actually, in inversion *2j*, the formation of heterochromatin could be facilitated by the large insertions at the breakpoints that involve several different TEs. Repetitive elements have been implicated as nucleation sites for heterochromatin formation via an RNAi mechanism in different organisms including *Drosophila* (PAL-BHADRA *et al.* 2004). Specifically, TE *1360*, which belongs to the *P* superfamily like *Galileo*, has been implicated as a target for sequence-specific heterochromatic silencing through an RNAi-dependent mechanism (HAYNES *et al.* 2006). Since both *1360* (ARAVIN *et al.* 2001) and *Galileo* (PUIG *et al.* 2004) are known to produce transcripts that could target dsRNA-mediated silencing to the region of chromatin where they are located, a transcriptional silencing mechanism based on dsRNA-dependent heterochromatin formation can not be completely ruled out either. Thus, even though the formation of dsRNA is probably not essential for an antisense RNA to be able to alter *CG13617* mRNA expression, what is really indispensable is that the two transcripts are transcribed simultaneously in time and space for any type of interference to take place. In the case of *CG13617*, the negatively correlated expression pattern of the sense mRNA and the antisense transcript is the strongest proof of regulatory activity, and therefore of sense and antisense expression happening in a coordinated manner. The *in situ* hybridization in embryos with a probe able to bind the antisense RNA did not yield any clear results, so the expression pattern of this antisense transcript remains to be characterized.

It has not escaped our attention that very low levels of an antisense RNA could be detected in the embryos of some *2st* lines when *CG13617* sense and antisense specific cDNAs were synthesized and used as templates in RT-PCR experiments (FIGURE 5 of PUIG *et al.* 2004). This could mean that an antisense RNA is part of the normal regulation of this gene in the *2st* arrangement. Such a transcript would require the presence of a promoter downstream of *CG13617* coding region in non-inverted chromosomes. Promoters in 3' UTRs of protein-coding genes have been recently detected in the human genome, and the resulting transcripts might regulate the adjacent downstream genes using a sense-antisense mechanism (CARNINCI *et al.* 2006). Evidence of downstream promoters driving antisense transcription was also provided by a recent study that mapped the binding sites of three human transcription factors (SP1, c-Myc and p53) to chromosomes 21 and 22 and found that 36% of the binding sites were located within or immediately 3' of well-characterized genes and are significantly correlated with non-coding RNAs (CAWLEY *et al.* 2004). If a downstream antisense promoter was present in the *2st* arrangement, the insertion of TE sequences in *2j* chromosomes could have replaced the original weak promoter (producing small, almost-undetectable quantities of antisense RNA) with a stronger one contained inside the TEs that increased the amount of the antisense transcript to a level able to reduce significantly *CG13617* mRNA expression. Alternatively, what we are detecting in *2st* lines could be unspecific transcription without a distinct biological function. We should take into consideration that up to 93% of DNA is thought to be transcribed in the ENCODE regions of the human genome but the biological relevance of most of these transcripts remains unclear (BIRNEY *et al.* 2007). In any case, the low amount of this RNA precluded any further analysis.

4.1.3 Transposable elements can induce transcription of adjacent sequences

Although the exact localization of the antisense RNA 5' end could not be determined, transcription appears to start within a 698-bp *GalileoK* element inserted inside the *GalileoG* element that originated inversion *2j* (CÁCERES *et al.* 2001). Transcripts initiated within TEs may seem unusual events, but increasing evidence reveals that they are quite frequent in many species. For example, in wheat, retrotransposon *Wis 2-1A* LTRs drive the synthesis of

new transcripts from adjacent sequences, including the antisense or sense strands of known genes (KASHKUSH *et al.* 2003). These sense and antisense RNAs increase or decrease, respectively, the expression levels of the corresponding genes. A recent high-throughput analysis of TSSs revealed that 6-30% of mouse and human RNA transcripts initiate within repetitive elements (retrotransposons, satellite DNA, or simple repeats) (FAULKNER *et al.* 2009). Retrotransposons located 5' of protein-coding genes can function as alternative promoters or express non-coding RNAs. Also, more than 25% of mouse and human genes contain retrotransposons in their 3' UTRs, and on average they show reduced expression levels when compared to retrotransposon-free transcripts (FAULKNER *et al.* 2009). These two examples demonstrate that retrotransposons (LTR-retrotransposons in plants and mainly LINEs and SINEs in mammals) are able to generate transcripts that can affect the expression of adjacent genes, either causing the transcription of the coding region from an alternative promoter (which may result in expression in different tissues or with a different timing) or by synthesizing regulatory RNAs.

However, *Galileo*, which seems to be driving the transcription of the antisense RNA affecting *CG13617* expression, is not a retrotransposon but a DNA transposon (MARZO *et al.* 2008). Another well-known example of repeated elements transcribing into adjacent sequences involving a DNA transposon is that of the *Stellate* (*Ste*) gene in *D. melanogaster*. In this case, the TE *1360* (also known as *Hoppel*) causes the synthesis of antisense transcripts in the bidirectionally transcribed Y-linked *Suppressor of Stellate* [*Su(Ste)*] repeats that are able to cause *Ste* silencing in testes of wild-type males, even though the X-linked *Ste* gene and *Su(Ste)* present only 90% nucleotide identity. Transgenic constructs revealed that a short 102-bp sequence of *Ste* that contains only 33 transcribed nucleotides is sufficient to confer *Su(Ste)*-dependent silencing of a *LacZ* reporter gene (ARAVIN *et al.* 2001). *Su(Ste)* is essential for male fertility and its deletion leads to abnormalities in spermatogenesis. Antisense *Su(Ste)* RNA has three different start sites within the *1360* sequence, with the longest transcript containing 441 bp of TE, and seems to be non-polyadenylated (ARAVIN *et al.* 2001). All the transcripts involved co-localize in the nuclei of spermatocytes, which indicates that they are expressed at the same time and are able to form dsRNA *in vivo*. The detection of short RNAs resulting from *Ste* and *Su(Ste)* dsRNA processing together with the fact that mutations of genes involved in the RNAi pathway eliminate *Ste* silencing, suggest that a post-transcriptional mechanism is regulating *Ste*

expression in the *D. melanogaster* germ line (ARAVIN *et al.* 2004). Moreover, the *P* element has also been reported to induce transcription of adjacent sequences in certain alleles of gene *nup154* due to the presence of an outward-reading promoter near its 3' end (KIGER *et al.* 1999). As mentioned above, the recent identification of *Galileo* transposase in *D. buzzatii* and other *Drosophila* species has allowed to classify *Galileo* as a TE related to *1360* and *P* elements that belongs to the *P* superfamily of DNA transposons (MARZO *et al.* 2008). Thus, it is surprising that the few known examples of regulatory transcripts initiated inside DNA transposons have all been found in *Drosophila* species and caused by related TEs. This could suggest that, in spite of their different structures and sequences, *Galileo* and *1360* might share some features that facilitate their ability to originate transcripts that extend into adjacent sequences.

Why are TEs capable of driving transcription of adjacent regions? Since they can provide promoter and *cis*-regulatory regions, TEs have been proved to be a rich source of regulatory elements that can be used by the host organisms to evolve regulatory mechanisms for the genes adjacent to them. In particular, TEs can contain promoters, enhancers, TFBSs, insulators, splice sites or polyadenylation signals within their sequences (MEDSTRAND *et al.* 2005, FESCHOTTE 2008). For example, 25% of experimentally characterized human promoters contain TE-derived sequences (JORDAN *et al.* 2003). This same study showed that ~8% of all proximal promoter regions (500 bp upstream of known TSSs) and 2.5% of known TFBSs are located within TEs. Besides, many promoters and polyadenylation signals in human and mouse genes are derived from primate-specific and rodent-specific TEs (VAN DE LAGEMAAT *et al.* 2003). These regulatory elements come from two possible origins. On the one hand, TEs provide raw material from which *cis*-regulatory elements can evolve *de novo* through point mutations (BRITTEN 1996a, MEDSTRAND *et al.* 2005). This is the case of certain human *Alu* sequences that are able to bind PAX6, a critical transcription factor involved in the development of the eye, pancreas and central nervous system (ZHOU *et al.* 2002). On the other hand, TEs already have pre-existing ready-to-use regulatory elements (promoters and TFBSs) to control their own expression that can be incorporated into the natural regulation of adjacent genes directly or after modifications of the surrounding environment. For example, *B2 SINE* elements carry an active RNA pol II promoter able to induce transcription (FERRIGNO *et al.* 2001) and the TFBSs that make possible that an LTR of endogenous

retrovirus *ERV-L* acts as a promoter for human gene *B3GALT5* were already present in the original consensus sequence for this kind of LTRs (DUNN *et al.* 2005). In the case of the retrotransposon-initiated transcripts detected in mice and humans mentioned above, the vast majority started in previously unidentified promoters (FAULKNER *et al.* 2009).

The *GalileoK* element downstream *CG13617* is a small 698-bp defective copy that does not have any coding capability. However, this copy could have retained the pre-existing promoter that *Galileo* must contain in order to control transcription of its own transposase gene, which is needed for transposition. To determine if this copy of *GalileoK* includes the region where the natural promoter of *Galileo* should be located (upstream of the transposase gene) and the relative orientation of this sequence with respect to gene *CG13617*, the *GalileoK* sequence was aligned with the full-length copy of *Galileo* isolated in *D. buzzatii* (MARZO *et al.* 2008). The *GalileoK* sequence aligns completely within the 1229-bp *Galileo* TIRs (results not shown) and does not seem to contain any of the internal sequences where the transposase gene and its promoter are located (in fact, the same alignment is obtained with the two possible orientations of the full-length *Galileo*). Therefore, no evidences could be found that this copy of *GalileoK* includes the original promoter of the TE or that it is responsible for the transcription of the antisense RNA.

Alternatively, a new promoter sequence could have evolved *de novo* in this particular copy through the introduction of one or a few point mutations, becoming able to recruit RNA polymerase and initiate transcription when a determined combination of transcription factors are present in the cell. Most *cis*-regulatory elements like TFBSs tend to be short and degenerate in sequence (WRAY *et al.* 2003), so it is not difficult to imagine that the high mutation rate that occurs in TEs (where most of the time there is no selective pressure to maintain the sequences) can originate such elements. In fact, putative promoter elements could be predicted in the TIRs of *GalileoK* using the NEURAL NETWORK PROMOTER PREDICTION tool (results not shown), although the small size and degenerate nature of regulatory sequences and TFBS, together with the lack of knowledge about many types of core promoter sequences makes them difficult to predict bioinformatically and these results are not always very reliable. For example, MCPROMOTER, a different software, does not predict any promoter in the *GalileoK* element where the antisense RNA is thought to initiate

(results not shown), even though it detects putative promoters in other TEs inserted at the proximal breakpoint junction of line $j-1$. However, these TEs are not found in all $2j$ chromosomes, but specific insertions that occurred in this particular line and that are probably not related to the antisense RNA production, a feature shared by all chromosomes carrying the inversion. In any case, a newly evolved promoter sequence could have been maintained by natural selection if the presence of the antisense in embryos turned out to be useful for the individual. TE insertions that have acquired a function are usually conserved in different species or present high frequencies inside one species. In the inversion $2j$, complex insertions made up of multiple nested TEs are found at the breakpoints. However, as mentioned above, the TE that provides the antisense transcript promoter seems to be present in all $2j$ chromosomes (CÁCERES *et al.* 2001) which might suggest that it acquired a function useful for the host that caused its increase in frequency in *D. buzzatii* populations.

Moreover, the promoter contained within the *GalileoK* copy seems to work mainly in embryos, the only developmental stage where it is able to generate the antisense transcript that causes the *CG13617* expression change with respect to *2st* individuals. FAULKNER *et al.* (2009) observed that the transcripts starting inside retrotransposons in the human and mouse genomes are frequently tissue-specific, with 35% of all retrotransposon-associated transcripts showing spatially or temporally restricted expression, in contrast to the 17% observed in transcripts initiated in non-repetitive elements. VAN DE LAGEMAAT *et al.* (2003) also found many cases where TE-derived promoters contribute to tissue-specific gene expression, like the placenta-specific promoter of the human *CYP19* gene or the erythroid-specific promoter of the carbonic anhydrase (*CA1*) gene, which are both found within LTRs. The presence of specific TFBSs within TE sequences can explain these expression patterns, which will be restricted to those tissues (or moments) where (or when) the transcription factor able to bind them is expressed. For example, in a *D. melanogaster* strain, the insecticide resistance gene *Cyp6g1* shows an increased expression restricted to tissues important for detoxification in larvae due to *cis*-regulatory sequences located within an *Accord* retrotransposon (CHUNG *et al.* 2007), and in humans several genes adjacent to *ERV* retroelement copies containing a p53 binding site are expressed in response to DNA damage (WANG *et al.* 2007). Besides, there are some TEs that are expressed only in certain tissues, so we could expect that if their regulatory elements were coopted by the host individual to express adjacent genes, at least in the initial

stages of the process, these genes would display the same tissue-specific activity. For example, in *D. melanogaster* *P* element is only able to transpose in germ cells due to a regulatory mechanism controlling the expression of the transposase gene (RIO 2002). Also, the tissue-specific expression of retroelements *roo*, strongly expressed during embryogenesis in certain restricted regions of the embryo (BRONNER *et al.* 1995), and *F*, transcribed in specific cells of the female and male germ lines and in various tissues during embryogenesis of *D. melanogaster* (KERBER *et al.* 1996), are both mediated by internal *cis*-acting elements contained inside the transposon. Therefore, the expression of the *CG13617* antisense RNA predominantly in embryos is not an uncommon phenomenon.

4.2 Consequences of *CG13617* silencing

After pinpointing the antisense RNA as the most likely cause of *CG13617* silencing in *D. buzzatii* *2j* embryos, we explored the consequences that the reduced level of expression of this gene might have for those individuals carrying inversion *2j*. Gene *CG13617* was silenced in *D. melanogaster* using RNA interference techniques in order to mimic what occurs naturally in *D. buzzatii* *2j* chromosomes. This technique allows the silencing of a single gene in a specific manner with the additional advantage that it uses the same mechanism operating in *D. buzzatii* *2j* embryos. At the same time, it provides a transient silencing effect that fades as development progresses, a situation again similar to what seems to be happening in *2j* individuals. By doing this, we constructed an artificial system in a totally different species, *D. melanogaster*, which allowed the use of microarrays to assess at a genomic level any gene expression changes that could be associated to the silencing of *CG13617*.

4.2.1 Expression changes associated to *CG13617* silencing

The comparison of the expression patterns from *CG13617*-depleted *D. melanogaster* first instar larvae and larvae expressing this gene normally using oligonucleotide microarrays revealed that only 41 genes showed expression changes, according to the chosen criteria. Surprisingly, all these genes exhibited a reduction in their expression levels when *CG13617*

was silenced, and none showed a consistent increase. This imbalance towards down-regulated genes strongly indicates the detection of a real effect because by chance we would expect to find both up- and down-regulated genes in similar proportions, even though the decreased expression levels of these genes could also be due to the unspecific action of the injected dsRNA (see below). In addition, a gene ontology analysis of the differentially expressed genes unveiled a significant overrepresentation of genes involved in DNA replication and cell cycle, with a total of 17 out of 26 genes with gene ontology annotation (65.4%) included in these functional categories. Some of these genes participate in nucleic acid metabolic processes (FIGURE 24) and some of them have roles in the replication fork (FIGURE 23), although there are also cell-cycle regulatory proteins like CycE or proteins implicated in nuclear transport like RanGap. In a systems genetics study where genome-wide transcript abundance was quantified in *D. melanogaster* wild-derived lines showing significant variation in six ecologically relevant complex traits, the variable transcripts could be grouped into modules of intercorrelated genes that were enriched for pathways, gene ontology categories, tissue-specific expression or TFBSs (AYROLES *et al.* 2009). Of the 41 differentially expressed genes detected in our experiments, 19 are included in the same transcriptional module, which corresponds to genes affecting basic cellular processes, a fact that further indicates that these genes are functionally related and expressed in similar conditions.

To confirm the microarray results, the expression levels of eight of these genes were measured again by real-time RT-PCR in a larger number of *D. melanogaster* samples, and differences in expression levels were validated for all of them. The five samples where *CG13617* had been silenced by RNAi showed a lower average expression level for the eight analyzed genes than the control samples expressing *CG13617* normally. It is noteworthy that despite hybridizing only two microarrays with each type of samples (with and without *CG13617* expression), the expression changes seem to persist when they are measured using a completely different method and additional independent samples are added to the initial two. However, we noticed that the magnitude of the observed differences is reduced in the real-time RT-PCR quantification with respect to the fold changes calculated based on the signal intensity detected in the microarrays (TABLE 6 and FIGURE 25). Although it is a common observation that the amount of change tends to be greater in real-time RT-PCR measurements than in hybridization-based methods due to the smaller detection range of microarrays

(CANALES *et al.* 2006, ARIKAWA *et al.* 2008), a similar decrease in the magnitude of gene-expression differences in real-time RT-PCR with respect to microarrays has also been observed in a larger comparison of expression levels in the human and non-human primate brain (M. CÁCERES, unpublished results). The addition of three new samples in the real-time RT-PCR experiments does not seem to be the cause of the observed differences between the two methods, since they exhibit intermediate expression values that do not differ significantly from those obtained for the two samples that had been previously hybridized in the microarrays. Still, there are several reasons that might explain the differences in the fold changes measured by the two independent techniques. First, we have to take into account that a fold change of 2 is commonly considered as the cutoff below which microarray and quantitative PCR data begin to lose correlation (WURMBACH *et al.* 2003, MOREY *et al.* 2006), and many of the analyzed genes present fold-changes of approximately this value. And second, an increasing distance between the location of PCR primers and microarray probes seems to decrease the correlation between the two methods (ETIENNE *et al.* 2004, ARIKAWA *et al.* 2008) and for many of the analyzed genes the primers used in real-time RT-PCR experiments are not located in the same sequences interrogated by the microarray probes.

In any case, for three of the analyzed genes (*mus209*, *RnrL* and *RnrS*) the differences detected by real-time RT-PCR between control and silenced samples are statistically significant ($P < 0.05$), and two more genes (*Mcm2* and *Mcm5*) show marginally significant ($P < 0.1$) differential expression. Therefore, five out of eight genes tested present substantial differences between those samples expressing *CG13617* normally and those where this gene has been silenced. The remaining three genes either show very small expression changes (like *CycE* with a fold change of only -1.14, which corresponds to a 12% reduction of the mean expression level in silenced samples with respect to controls) or exhibit considerable variation within each type of samples that does not allow the differences between both types to be considered significant. Despite this, the mean expression values are in all cases decreased in the samples where *CG13617* has been silenced.

Next, the average expression levels of ten of the differently expressed genes involved in DNA replication identified in the *D. melanogaster* microarrays, were also compared between *D. buzzatii* 2st and 2j embryos (FIGURE 26a). The results indicate that nine genes (90%) also

show expression differences between chromosomal arrangements in *D. buz̄z̄atii*. Besides, the expression changes are always in the expected direction: a reduction of the expression level in those lines carrying the inversion, which are the lines with a decreased *CG13617* expression in embryos. Taken separately for each gene, the differences between lines carrying the *2j* or *2st* arrangement are not statistically significant. However, considering the results obtained for the ten genes as a whole, a clear trend can be observed: several genes involved in DNA replication show lower expression levels in *2j* embryos when compared to *2st* ones. That nine out of ten genes have reduced expression levels in *D. buz̄z̄atii* as in *D. melanogaster* is a significant result according to the sign test and a *t*-test for paired comparisons.


Gene expression levels for these ten genes were also measured in first instar larvae from *2st* and *2j* lines but, in this case, no differences could be detected for any of them (FIGURE 26b). This agrees with the fact that *CG13617* expression difference between arrangements is lower in first instar larvae (~1.5-fold) than in embryos (~2.5-fold) (FIGURES 26c and 32). This declining difference with age probably reflects the normalization of *CG13617* expression level in *2j* individuals as development progresses and the antisense RNA is no longer transcribed (FIGURE 3 of PUIG *et al.* 2004). The decrease in the expression level of DNA replication genes specifically in embryos (the only developmental stage where *CG13617* expression is altered in *2j* lines) but not in later stages points to *CG13617* having a causal role in this reduction. Interestingly, the expression levels of DNA replication genes seem to level off between arrangements almost at the same time that *CG13617* levels do: in the larval stage, coinciding with the antisense RNA disappearance. Thus, it is possible that *CG13617* expression reduction in first instar larvae is not sufficient to generate a detectable effect in any of the other genes. Although the differentially expressed genes were initially detected by the microarray analysis not in embryos but in *D. melanogaster* first instar larvae, we have to take into account that experimental *CG13617* silencing caused by dsRNA injection persisted at least until third instar larvae, which still show an almost complete repression of *CG13617* expression (FIGURE 18). Hence, the situation in dsRNA-injected *D. melanogaster* larvae resembles more that found in *D. buz̄z̄atii* *2j* embryos (where *CG13617* shows a decreased expression) than that of *D. buz̄z̄atii* larvae (where *CG13617* expression is no longer affected by the antisense RNA).

It is remarkable that *mus209* is the one gene that does not seem to show any difference in its expression level between *2st* and *2j* lines in *D. buzzatii* embryos, since this is the gene with the highest fold change in *D. melanogaster* real-time RT-PCR experiments (-1.75) and also one of the better supported expression changes in the microarray results, appearing as a differentially expressed gene in the three independent analyses we performed (TABLE 6). In general, *D. buzzatii* real-time RT-PCR fold changes are larger than those detected by the same method in *D. melanogaster*. For example, *RnrS* presents a fold change of -1.37 in *D. melanogaster*, but -1.92 in *D. buzzatii*, or *CycE* with -1.14 and -1.85, respectively. This is surprising because the silencing induced by the exogenous dsRNA injection in *D. melanogaster* seems to be stronger (*CG13617* expression is barely detectable and this effect persists much longer) than that caused by the natural *CG13617* antisense RNA in *D. buzzatii* embryos.

4.2.1.1 Magnitude of the expression changes induced by *CG13617* silencing

For most of the differentially expressed genes, the expression changes detected in the two species are small according to both measuring techniques: oligonucleotide microarrays and real-time RT-PCR. Microarray fold changes fluctuate between 2 and 4 for the vast majority of our differentially expressed genes. However, even small changes in gene expression can have substantial and biologically significant phenotypic effects (CARROLL 2005, WRAY 2007). For example, a 15% reduction in mouse *Gpc3* (a glypican involved in morphogenesis and growth regulation) gene expression level underlies a quantitative trait locus (QTL) responsible for the response to selection on growth, causing a change of body mass of approximately 20% (OLIVER *et al.* 2005). Also, when *PDYN* (a gene that encodes prodynorphin, the precursor of a neuropeptide with critical roles in regulating perception, behavior, and memory) expression is compared between humans and chimpanzees, a slight up-regulation is detected when transcription is induced by intracellular calcium release. The nucleotide changes responsible for this difference are located in the human promoter, which shows signs of positive selection, suggesting that it is possible that minor changes in gene regulation played a significant role in the evolution of human traits (ROCKMAN *et al.* 2005).

Several reasons can explain why the detected expression differences in the *D. melanogaster* experimental system or between *D. buzzatii* chromosomal arrangements are small. (1) Different spatial expression patterns in embryos of *CG13617* and the analyzed differentially expressed genes; (2) Variation of expression levels within chromosomal arrangements; and (3) Subtle differences are more likely to persist in natural populations. Each point is explained in more detail next.


In the first place, according to FlyExpress , a database of embryonic expression patterns in *D. melanogaster*, the studied DNA replication genes for which there are available expression data seem to be expressed in multiple tissues during embryonic development, while *CG13617* exhibits a more restricted spatial expression pattern, at least in *D. buzzatii* (SUPPORTING FIGURE 7 of PUIG *et al.* 2004). We expect that the expression level of the DNA replication genes will only be altered by *CG13617* silencing in those cells where both genes are being co-expressed. Therefore, the cells that normally do not express *CG13617* will remain unaffected by its silencing and will express the DNA replication genes at the usual levels. As a result, changes in gene expression may occur in only a small fraction of the cells expressing a given transcript. Since expression is quantified in whole embryos, the cells expressing DNA replication genes normally are pooled and measured together with those that present an altered expression caused by *CG13617* silencing. It is thus likely that this dilution effect due to the proper expression of the DNA replication genes in most tissues contributes to attenuate the differences found in microarray and real-time experiments for both species. In those particular groups of cells where *CG13617* expression has been reduced 2.5-5 times, as happens in *2j* embryos, the effects on DNA replication genes expression levels could be much higher than the detected fold changes.


Second, a great degree of variation in gene-expression levels has been found within arrangements in *D. buzzatii* measurements. This variation is mainly due to two lines. The combination of one *2st* line with unusually low expression levels in all the analyzed genes, together with a *2j* line that systematically exhibited a higher expression level than the rest of the lines carrying the inversion, resulted in average expression differences that are not statistically significant in many cases. Removing these two outlier lines from the analysis the differences become highly significant for all genes tested. Nonetheless, there is no basis to

exclude these two lines, since they are part of the natural populations of *D. buzzatii* and they represent the variability found within these populations. In addition, these two lines show levels of *CG13617* expression that are similar to those of the rest of lines with the same chromosomal arrangement. To explain the uncommon expression levels detected in these two lines (with respect to the other lines carrying the same arrangement) it is important to take into account that the expression of the DNA replication genes does not depend uniquely on *CG13617* activity, but rather multiple factors control the expression of these genes. Since all the analyzed genes follow the same pattern, with the same two outlier lines in each chromosomal arrangement, the cause of this different expression is likely to be a variant in *trans*, so that a single change is able to affect the expression of several DNA replication genes simultaneously (WITTKOPP *et al.* 2008). Several mechanisms like differences in the affinity, activity or availability of transcription factors (WITTKOPP 2010) involved in the regulation of the expression of the DNA replication genes could account for changes in the expression levels of these genes independently of the levels of *CG13617* protein. In these cases, the functional polymorphisms may lie in the coding or non-coding regions of genes producing direct regulators of these DNA replication genes, or in genes producing indirect regulators that modify the activity of direct regulators. In any case, the existence of such variants could influence greatly the expression level of a transcript in a given line and contribute to alter the mean expression values for each chromosomal arrangement. Possibly, the use of a greater number of samples for each arrangement would increase the statistical power to detect significant differences, since it would reduce the weight of each individual measurement in the mean expression values.

Finally, we need to take into account that inversion $2j$ is a natural inversion that has passed the filter of selection in *D. buzzatii* natural populations. For this reason, we expect to find subtle differences between arrangements rather than dramatic changes, since these are more likely to be deleterious for the individuals carrying them and would have been eliminated by purifying selection. Modifying slightly the expression level of a gene at a certain developmental stage (even if as a consequence, the expression level of another set of genes is readjusted as well) will have less pleiotropic effects (and therefore will be less harmful) than completely inactivating a coding sequence either by disrupting it or by introducing amino acid changes in the encoded protein, which affects all the tissues where the gene is expressed and

might entail negative consequences (CARROLL 2005, WRAY 2007). For example, loss-of-function mutations of gene *Gpc3* cause Simpson-Golabi-Behmel syndrome in humans, a disorder with numerous phenotypic effects like overgrowth, skeletal and renal abnormalities, an increased frequency of embryonic cancers and neonatal mortality. *Gpc3* knock-out mice show similar phenotypes, but when this same gene is simply down-regulated it leads to an increased body size in this species with no pathological side effects (OLIVER *et al.* 2005). Therefore, the same gene that causes a disease when completely inactivated, can also generate favorable and selectable phenotypes when its regulation is modified. In the case of gene *CG13617* and inversion *2j*, it seems that the expression level of some genes involved in DNA replication could be affected as well, and we have to take into consideration that DNA replication is a very important process during development, especially in the embryonic stages when extensive cell proliferation takes place. So, probably DNA replication is not a biological function that can be easily altered without further consequences, and larger fold changes of these key genes could seriously impair the individuals' ability to survive.

The only exception to small fold changes in the differentially-expressed genes list are four genes (*CG8087*, *CG13135*, *CG32198* and *CG14850*) with unknown function that belong to a putative family formed by 31 genes in the *D. melanogaster* genome, according to the PANTHER  database (THOMAS *et al.* 2003). These genes show the largest expression changes in the microarrays with expression levels decreasing 4-14 times in the samples where *CG13617* is silenced (the exact fold changes depend on the gene and the analysis tool, see TABLE 6). Even though the different members of this protein family are quite different at nucleotide and amino acid level, some of these genes are grouped in tandem in *Drosophila* genomes and their copy numbers in a given genomic location vary when several *Drosophila* species are compared. Unfortunately, this repetitive and copy-number-variable nature, together with the large phylogenetic distances between the available sequenced *Drosophila* subgenus genomes, prevented us from identifying reliably in the *D. buzzatii* genome the coding region of *CG8087* (among these four genes, the one with the best supported expression change). As a consequence, gene-expression levels could not be compared between *2st* and *2j* lines. However, two of these genes, *CG8087* and *CG14850*, have been found to be up-regulated in microarray experiments analyzing *trithorax* (*trx*) *D. melanogaster* mutants and together with two other genes of the same family are considered to form a cluster of co-

expressed genes (BLANCO *et al.* 2008). The fact that *CG8087* and *CG14850* are part of the same family and that they are co-expressed in response to another different situation (*trx* deficiency), apart from *CG13617* silencing, suggests that these genes are often expressed under the same circumstances (possibly by sharing regulatory elements) and that they are indeed functionally related. While these four genes do not have any gene ontology terms assigned and their sequences do not show homology to any known proteins in similarity searches, PANTHER  database describes this protein family (PTHR23246) as structural proteins involved in cell structure maintenance. This agrees with the fact that most of the co-expressed gene clusters found by BLANCO *et al.* (2008) to change their expression levels in *trx* mutants consist of genes that encode structural proteins involved in cuticle formation. Therefore, it does not seem likely that these genes are also involved in the DNA replication process, but probably they are part of another biological pathway in which *CG13617* protein might play some kind of regulatory role (see below).

4.2.1.2 Specificity of *CG13617* silencing effect

Since all the detected differentially-expressed genes are down-regulated, it could be argued that this is due to an unspecific action of the injected dsRNA in *D. melanogaster*. Upon injection, the long ~600-bp dsRNA molecule is processed into multiple 21-nt fragments (siRNAs) able to target the complementary mRNA to be degraded (see BOX 3 for more details). Each of these fragments could also silence genes other than *CG13617* if their sequences are similar enough to bind to other transcripts. Besides, several other silencing mechanisms (such as translational repression mediated by miRNAs) do not require a perfect complementarity between the transcript and the small RNAs (CARTHEW and SONTHEIMER 2009). As a result, we could observe these unspecific targets as genes with decreased expression in the subsequent microarray experiments, but they would not correspond to genes being regulated by *CG13617*. In addition, it is also possible that there is some kind of general response caused by the introduction of exogenous nucleic acids in the cells.

Several reasons suggest that this situation is not likely to be happening in our experiments. First, we selected the 587-bp dsRNA because a BLASTN similarity search of this

sequence against the *D. melanogaster* genome yields a single highly significant hit corresponding to gene *CG13617* (100% identity and coverage, E-value = 0.0). Only seven short stretches (19-36 bp long) of the dsRNA sequence also present secondary hits with other parts of the *D. melanogaster* genome (identities 84-100%, E-values = 0.13-5.6). However, four of these fragments show similarity to intergenic regions that are not transcribed, and therefore not susceptible to be silenced through RNAi mechanisms. The remaining three match coding regions, but none of the corresponding genes is included in the differentially expressed list. Second, the Affymetrix oligonucleotide microarrays probes that interrogate *CG13617* expression cover the same region of the gene that the injected dsRNA. In fact, thirteen of the fourteen 25-nt probes specific for *CG13617* overlap the dsRNA sequence. The reason for this coincidence is that both dsRNA and oligonucleotide probes were inadvertently designed in the same part of the gene with the same purpose: maximum specificity. This region spans the fourth and fifth exons and does not contain the more conserved initial fraction of the protein or any of the putative protein motifs identified in *CG13617* (which might cause cross-hybridization with other mRNA sequences) and probably corresponds to the most distinctive sequences able to identify *CG13617* transcripts unequivocally. Third, as we have seen, most of the affected genes are functionally related, and it is unlikely that several genes silenced by mistake based on the accidental homology of some siRNAs turn out to participate in the same biological process when they are unrelated from the point of view of sequence. However, our observations still could be explained if a positive regulator shared by a group of genes involved in a certain process was unspecifically affected by *CG13617* dsRNA. Finally, the fact that the expression levels of the differentially expressed genes detected in *D. melanogaster* are also reduced in *D. buzzatii* 2j lines points clearly to *CG13617* silencing having a role in causing this effect. The possibility that siRNAs generated in *D. buzzatii* as a consequence of the processing of the dsRNA formed by *CG13617* mRNA and the antisense transcript silence unspecifically and by chance the same genes that in *D. melanogaster* seems highly unlikely.

In summary, the reduction of the expression level of several genes involved in DNA replication and cell cycle regulation seems to be indeed a consequence of *CG13617* silencing for two main reasons: (1) The expression changes in these genes were first detected in *D. melanogaster* between samples only differentiated by the presence or absence of *CG13617* expression, but expression changes in the same direction were also found for those same

genes in *D. buzzatii* 2j embryos, where *CG13617* expression is naturally decreased by an antisense RNA. *CG13617* silencing is the only factor shared between both systems (*D. melanogaster* RNAi samples and *D. buzzatii* 2st/2j chromosomal arrangements). (2) In *D. buzzatii*, the expression changes in DNA replication genes occur only in embryos, when the differences in *CG13617* expression level between 2st and 2j lines are larger. In first instar larvae, when the *CG13617* expression difference between arrangements is small and no antisense RNA is produced, no effects were detected on the expression of any of the other genes.

4.2.2 Gene *CG13617* structure and function

CG13617 was predicted computationally as a potential ORF with unknown function during the initial annotation of the *D. melanogaster* genome (ADAMS *et al.* 2000). In this work we have shown that this gene produces a 2.3-kb mRNA that is expressed throughout the whole *D. buzzatii* life cycle and encodes a 734-aa protein, thus confirming that it is indeed a functional gene. *CG13617* has been sequenced in two *D. buzzatii* lines, st-1 and j-1, as well as in *D. martensis*, another *repleta* group species. Moreover, the sequencing of 12 *Drosophila* genomes (CLARK *et al.* 2007) gave us the opportunity to identify genes orthologous to *CG13617* in all these species. *CG13617* is located in *D. buzzatii* chromosome 2 (which corresponds to *D. melanogaster* chromosome arm 3R) in a syntenic block between genes *Pp1 α -96A* and *nAcR β -96A*, forming a conserved group of genes that are found in the same order and orientation in the 12 sequenced genomes as well as in *D. buzzatii*. Comparative DNA and protein sequence analysis of *CG13617* allowed us to determine the following: (1) *CG13617* shows changes in its exonic structure in the analyzed *Drosophila* species; (2) All *Drosophila* proteins share several putative functional domains, which include a zinc finger, three coiled coil regions, two PEST sequences, and putative nuclear localization and export signals, which give hints of their possible function as regulatory proteins; (3) Proteins with similar structure and sequence to *CG13617* seem to be present in multiple animal species ranging from insects to humans, and show some homology to the zebrafish Iguana and human DZIP1 proteins.

4.2.2.1 Changes in gene structure within the *Drosophila* genus

Although *CG13617* coding and protein sequence is conserved (TABLE 11), gene structure varies in some of these species. *CG13617* coding region contains 5 exons and 4 introns in all the *Sophophora* subgenus species except for *D. pseudoobscura* and *D. persimilis* where the third intron has been lost (FIGURE 27). Conversely, all the *Drosophila* subgenus species have 4 exons separated by 3 introns (FIGURE 27). This structure results from the elimination in the *Drosophila* subgenus of *D. melanogaster* intron 2, fusing together what in this species are exons 2 and 3, or, alternatively, the gain of this same intron in the *Sophophora* branch, which would break an originally large exon 2 in two different parts. Therefore, the evolution of *CG13617* gene structure in *Drosophila* requires at least one event of intron loss (intron 3) in the ancestor of *D. pseudoobscura* and *D. persimilis*, and either another intron loss or an intron gain (intron 2) in the *Drosophila* or *Sophophora* subgenus ancestors, respectively. For intron 2, we can not distinguish between these two scenarios because the closest outgroup species have a quite different gene structure that complicates comparison. The *Anopheles gambiae* homologous gene (TABLE 13) has a single intron corresponding to the first *Drosophila* intron (although probably this protein is incomplete at the C-terminal end, see below), and the available coding sequence from another mosquito species, *Culex quinquefasciatus*, is composed of only three exons: two short first and third exons and a large second exon that comprises the most part of the protein, and where the introns lost or gained in *Drosophila* would be located. In fact, variation in intron number does not seem to be an infrequent event in *Drosophila*. In a study where 28933 *D. melanogaster* introns were mapped in the remaining 11 *Drosophila* sequenced genomes, a total of 1944 introns (6.7%) were missing from one or more species (COULOMBE-HUNTINGTON and MAJEWSKI 2007). A total of 1754 intron loss events and 213 gain events could be inferred from these data by maximizing parsimony over the phylogenetic tree. An additional 220 differences were found between the two oldest branches leading to *Sophophora* and *Drosophila* subgenera, where losses and gains can not be distinguished due to the lack of an outgroup. In this analysis, 82% of the missing introns were exactly cut out (or inserted). In the *CG13617* case, intron 2 was precisely inserted or deleted without modifying any exonic nucleotides, whereas the poor conservation of the last part of the protein together with the larger phylogenetic distance with respect to the other available *Drosophila* subgenus sequences,

do not allow to determine if intron 3 was cleanly eliminated in the *D. pseudoobscura* and *D. persimilis* ancestor or not.


Several mechanisms have been proposed for intron loss, such as the genomic deletion of an individual intron (for example through recombination between short direct repeats at the intron ends) or the recombination of an intronless cDNA with the genomic version of the gene, resulting in the deletion of one or more introns (a mechanism that requires the reverse transcription of a processed mRNA into cDNA) (RODRÍGUEZ-TRELLES *et al.* 2006). Also, different mechanisms could explain intron gain, like a TE insertion inside an exon coupled with the use of splice sites contained within the inserted sequence, or the duplication of a pre-existing intron via reverse-splicing into a new location in the same (or another) mRNA followed by recombination of a reversely-transcribed copy with the corresponding genomic DNA (reverse-splicing can represent a rate-limiting step in this case) (RODRÍGUEZ-TRELLES *et al.* 2006). Even improper recombination events between homologous copies of genes and the creation of new splice sites within exons through single nucleotide mutations could be involved in the generation of new introns (COULOMBE-HUNTINGTON and MAJEWSKI 2007). It is important to note that intron gain/loss rates fluctuate across lineages and evolutionary time. Also, some introns seem to have a biological function, like the binding of specific transcription factors, and this could explain their observed higher level of conservation and the fact that they are less likely to be lost (COULOMBE-HUNTINGTON and MAJEWSKI 2007).

4.2.2.2 Protein sequence analysis

The presence of certain motifs in a protein sequence can provide clues about its molecular function inside the cell, as well as about the biological processes it might be involved in. In particular, several known motifs have been identified in the protein encoded by the *D. buzzatii* CG13617 gene. First, the existence of a well-characterized C2H2 zinc finger suggests that the CG13617 protein may bind to DNA to perform its function. However, a single zinc finger domain, like in this case, is in itself considered not sufficient for high-affinity binding to a specific DNA target sequence (KLUG and SCHWABE 1995). Zinc fingers were first identified as a DNA-binding motif, but it is now known that they can also bind to RNA

(HALL 2005) or protein targets (BRAYER and SEGAL 2008). Therefore, CG13617 zinc finger could act binding to other proteins, although cases of proteins with a single zinc finger capable to bind DNA with high affinity have also been reported, like *Drosophila* transcription factor GAGA (PEDONE *et al.* 1996) or *Arabidopsis* SUPERMAN protein (DATHAN *et al.* 2002) and this possibility should not be ruled out without further studies. Coiled coil domains are oligomerization motifs (BURKHARD *et al.* 2001), and their presence suggests that CG13617 interacts with other proteins. The detection of PEST sequences may be indicating that CG13617 has a rapid turnover (a short intracellular half-life), since these signals target the protein for degradation (ROGERS *et al.* 1986). Finally, the identification of nuclear localization and export signals (NLS and NES) points to CG13617 being a protein capable of entering and exiting the nucleus (GAMA-CARVALHO and CARMO-FONSECA 2001). In addition, these functional motifs are conserved in all the analyzed *Drosophila* species with minor variations (see Results for more details), which reinforces their computational prediction. However, despite the conservation of these functional features, the overall sequence conservation of the CG13617 protein among the different *Drosophila* species is not very high (the overall amino acid identity between *D. buzzatii* and *D. melanogaster* is 59.7%, see TABLE 11 for more details). As can be seen in the conservation graph below the multiple alignment of the 14 *Drosophila* CG13617 protein sequences in FIGURE 30 and in FIGURE 31, the first third of the protein is the part that presents a higher level of conservation. The rest of the protein is quite divergent, to the point of being difficult to align among distant species (in the second half of the alignment less than 20% of all positions are identical in the 14 sequences and 27% include gaps in at least one species). The better conserved fragment of the protein includes the two first coiled coil regions separated by the C2H2 zinc finger and two groups of three amino acid residues that are conserved not only within the *Drosophila* genus but in all the proteins similar to CG13617 obtained in the similarity searches in multiple species (see below), which indicates that these features must be essential for the proper functioning of this protein.

The collection of protein motifs identified in the same protein could be consistent with CG13617 carrying out transcription factor functions (which is one of the reasons why we looked for expression changes associated to its silencing). However, transcription factors usually have multiple zinc fingers and, although short coiled-coil domains are also frequently found as homo- and heterodimerization motifs in this kind of proteins (BORNBERG-BAUER *et*

al. 1998), the most common form are leucine zippers, a class of coiled coils that includes leucines every seven residues in a repeated pattern, and CG13617 coiled coils do not fit into this model. In fact, FlyTF , a *Drosophila* transcription factor database, includes CG13617 in a list of 754 putative site-specific transcription factors, but it is not found in a reduced list of 294 well-supported trusted candidates. The reason for this is probably related to the fact that the functional features explained above do not seem to correspond exactly (in structure or number) with the domains we would expect to find in transcription factors, so other types of functions can not be completely ruled out for CG13617 and should be considered.

4.2.2.3 CG13617 homologous proteins in other species

Luckily, even though CG13617 has not been studied in any of the other *Drosophila* species, the search of similar proteins in other organisms provided some useful insights about possible functions for this gene. An extensive search for *CG13617* orthologs in species other than *Drosophila* in the GenBank database revealed the presence of homologous genes in other insects (three mosquito species, a beetle and the honey bee), an echinoderm (sea urchin), two cnidarians (a sea anemone and a hydrozoan), a tunicate, a cephalochordate (lancelet), and several vertebrate species, including two amphibians (frogs), a fish (zebrafish), a bird (chicken), and nine mammal species ranging from marsupials to rodents and primates (TABLE 13 and FIGURE 37). So, animals belonging to very different animal classes seem to possess genes with significant similarity to *D. buzzatii* CG13617.

The region of homology of most of these sequences is restricted to the initial portion of the protein that comprises from the initial amino acids to the zinc finger or the second coiled coil, depending on the species considered. This agrees with the observation that in *Drosophila* species, the level of conservation is variable along the protein and the most conserved part is the first third of the sequence, where the detected functional domains are located (FIGURE 30). In fact, all the proteins retrieved in the searches include a C2H2-type zinc finger (TABLE 13), except for the *Apis mellifera* sequence, which presents a much shorter protein. However, the *Apis* protein is probably improperly annotated due to the presence of multiple gaps in the genomic sequence that could result in internal parts of the gene being

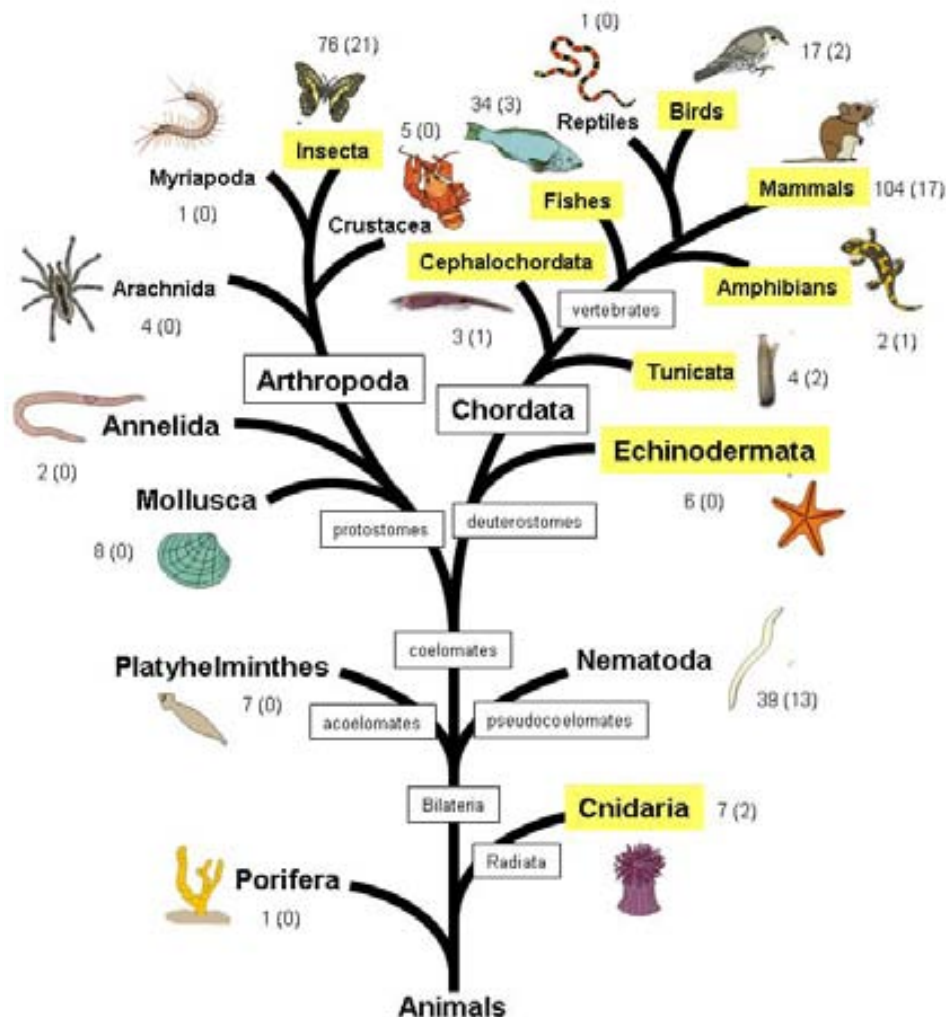


FIGURE 37 | Animal groups with proteins homologous to CG13617. In this tree representing all animals, those classification groups where at least one species seems to possess a protein showing similarity to CG13617 are highlighted in yellow. These results are based on the BLAST searches performed using *D. buzzatii* CG13617 protein sequence as a query. See TABLE 13 for more details. For each group it is indicated the number of genomes already sequenced or currently being sequenced as well as the number of complete available genomes (in parentheses), according to the GOLD genomes online database [↑](#).

missing. The *A. gambiae* protein sequence, which shows the highest similarity and represents the only case where the alignment with *D. buzzatii* CG13617 extends through the whole 607 aa of the mosquito protein, is also reported to be incomplete on the C-terminal end in the GenBank reference sequence. Moreover, the VectorBase [↑](#) database information about this

gene in the mosquitoes *A. gambiae* and *C. quinquefasciatus* states that both proteins possess 4 and 5 coiled coil regions, respectively, which is coincident with the results obtained in the *Drosophila* protein sequence analysis. Unfortunately, the function of many of these orthologous proteins remains unknown. Most of these genes are predicted annotations made in the respective sequenced genomes, although they seem to be indeed related among them since many have been tentatively named DAZ interacting protein 1-like (*DZIP1L*) because of their similarity to human *DZIP1* gene in several species.

4.2.2.4 Is CG13617 a component of the Hedgehog signaling pathway?

The only two proteins retrieved from sequence databases showing similarity with CG13617 that have been analyzed in some detail are those encoded by the genes *iguana* (*igu*) in zebrafish and *DZIP1* in humans. Although amino acid identity is low among them because *Drosophila*, zebrafish and humans are distantly related species, Iguana and DZIP1 share several protein motifs with CG13617 that suggest a common cellular function (FIGURE 38). At least Iguana (and possibly also DZIP1) contains a single zinc finger, coiled-coil domains, PEST sequences and nuclear localization and export signals, the same motifs detected in CG13617. This coincidence of sequence similarity and structural motifs points to *CG13617* being the most likely *Drosophila* ortholog of *igu* and *DZIP1*, as has been suggested by WOLFF *et al.* (2004).

Functional studies indicate that Iguana/DZIP1 is a component of the Hedgehog (Hh) signaling pathway (SEKIMIZU *et al.* 2004, WOLFF *et al.* 2004). This pathway controls several important developmental processes in animals (see INGHAM AND MCMAHON (2001) for a review). Reduced Hh signaling can cause serious developmental defects, and inappropriate activation contributes to certain forms of cancer in humans. Much of the current knowledge about the Hh signaling pathway is based upon genetic analysis in *Drosophila*. The Hh protein that acts as signal is released by the Dispatched (Disp) protein from secreting cells and binds to the cell surface receptor Patched (Ptc) in responding cells. This interaction activates the transmembrane protein Smoothed (Smo) that signals to the transcription factor Cubitus interruptus (Ci), which results in its release from the protein complex it was part of together with Fused kinase (Fu) and Costal2 (Cos2). Active Ci is transported into the nucleus where it

```

Dbuz CG13617 -----MGYKNN--FPQVMREAG-----
Drer Iguana -----MPFYDNVYYPSPDPG-----THSSAGIPSLSSPQSQPSSGSQSRPAPSTMSGLPTS
Hsap DZIP1 MQAEAADWFSMPFKHVYYPPLASGPEGPDVAVAAAAAGASMACAP-----PSAASGPL--
      * : . : : *

```

```

Dbuz CG13617 -----FKLRQYRDGPLDWRQMGSYETERILREQNLEVVDEALQHLSEAPLGTMLETHI---LDGG
Drer Iguana SGASTSIPPPFKFRSRREN-VWRINAVDVRVACEMDFQALQEHINAVTFCVVEGERCHRCQSPVDPA
Hsap DZIP1 -----PPFQFRPRLS-VWRRLSAIDVVKVAGAVDVLTLQENIMNITFCKLEDEKCPHCQSGVDPV
      * : * : . : * * * : : : : : . : : * : : : :

```

```

Dbuz CG13617 IAKYFIMSQYAIQYLLCCRTYLDESVDLREAEHISQOEIAKLRKSLSESNNVQLHKKITQIETI---
Drer Iguana LIKLFRLAQLTVEYLLHSQDCLSLSQAAEERLLAEAREREQICVQLQKKTQDAKALKEELKQRKIIAS
Hsap DZIP1 LLKLRLAQFTIEYLLHSQEFLLTSQLHTLEERLRLSHCDGQSKLLTKQAGEIKTLKKECKRRKMMIST
      : * : : * : : * * * . . * : : * : : : * : : : * : : : :

```

```

Dbuz CG13617 -----REVVFPCHLCTKNFISNEALNVHIGRKH-----RMGTPSSIVAGGATVRDK---ENDMO
Drer Iguana QQAMFSAGISANYHKCQHCEKAFMNASFLQSHMQRHPSEFDMKLMTDNQKKIQTVKLQDEINKLQEQLT
Hsap DZIP1 QQLMIEA--KANYQCCHFCDKAFMNAFLQSHIQRHTEE---NSHFEYQKNAQIEKLRSEIVLVKLEELQ
      . * : * * * : . * : * : * * . . . : : : : :

```

```

Dbuz CG13617 LINT-----IKMELEIKQLKERLNAEARNIKERSGSRSHLQLQDQTTSTMRDVG
Drer Iguana LVTSQMETQKKDYTAKQEKEL---IQRQEEFKRQLEIWKKEEKMRMNSKIDEVKQACQRDMDSINQRNRN
Hsap DZIP1 LTRSELEAAHSAVRFKSEYEMQKTKEDFLKLFDRWKEEKEKLVDEMEKVKEMFMKEFKELTSKNSA
      * : : : : : : : * : . . . : : : :

```

```

Dbuz CG13617 IQSNLADYKEKD--D-VSSEATESEATERKE-----QLHGLAERLNSFEA-WQAQLKQSNEDF----
Drer Iguana LETELKLLQKKNIQESMOSVQTQPNASTSNE---HWQEVVKLQKQLHKQEVKWTGKMQKMKEDHDREKS
Hsap DZIP1 LEYQLSEIQKSNMQI-KSNIGTLKDAHEFKEDRSPYPQDFHNVMLLDSQESKWTARVQAIHQEHKKEKG
      : : * . : : : . . * : * : * : . : : : * * * * : : : : :

```

```

Dbuz CG13617 -----IRGINQKLED---LITIALEQTKRNASKGAAPTAEERVATPC
Drer Iguana LLQEELCKVSAVSEGMEES---RRQVQELSHRLQEQQIITSQNKQMKQISSKPTITVQREGVSTPS
Hsap DZIP1 RLLSHIEKLRSTMIDDLNASNVFYKRIEELGQRLQEQNELIITQRQQIKDFTCNPLN-SISEPKVNAPA
      : : : : * : : : : : : : : : : * : *

```

```

Dbuz CG13617 LEDLERILTEKVAEIGKVSANKLEEVVHQLETTYKEKLEALERDLRRLSLRQTPAAEVQQAASTSKIIPK
Drer Iguana PETKAKV---VVEQSNVSHKLDPIV-ELSEEDKSSSISSEPTENRSWQK---EVQELLNKGLRR
Hsap DZIP1 LHTLETKSSLPVMEHQAFSSHILEPIE-ELSEEEKRENEQKLNKMMHLRK-----ALKSNSLTK
      . . * : * : : : * . . : : . . * . . . .

```

```

Dbuz CG13617 ALPKKEEN--MDRIKRL-----VETEFLKAKRDDDTY--SIEPPPPPEEHVVTH
Drer Iguana DMRLLAQHNLDDRLQSLGKIG-VSGLSKNLYKSSMTQIISDRRKKLEEDPVYRRALKEISHKLEQRVKER
Hsap DZIP1 GLRTMVEQNLMEKLETLGINADIRGISDQLHRVLKSVESERHKQEREIPNFHQIREFLEHQVSCKEEK
      : : : : * : : : * : : : : : : : :

```

```

Dbuz CG13617 ---VQEQPSGNSGSHPTYTKP--APAPAPARDEVKPKATTSE---PE-----QSEATDI---
Drer Iguana ---NTEQPVKSKLHEQVVQSRPRSSFPSTVTRVMGSPASKQRTQPQV-----PRSRTNVPHKT
Hsap DZIP1 ALLSSDQCSVSMQDMLTSTGEVPMKIQLPSKNRQLIRQKAVSTDRTSVPKIKKNVMDPPFRKSSITIT--
      : * . . . * * : : * . : * : :

```

```

Dbuz CG13617 -----SESLSGEE---SISDEGSEVLTSEPERQVFMSPK---IKSSLKTKLPPKPLTRKDKARKLIN
Drer Iguana STPLQHRRTPPFFSDEDSSEEEEEEEEEESDEESPMQKKTVLVNSSTAKAQNTAKTQSTAQSVRSVAV
Hsap DZIP1 -----TPPFSSEE---EQEDDDLIRAYASPGPLVPPFQ---NKGSF-----
      : : * : * . : : : : : : : : :

```

```

Dbuz CG13617 QKLSPHGFNMKSKTISNTTAKRVSael-----AQQRARLK-LDYPNFYTRNRIRKRVFKLSAKMP
Drer Iguana ALTSAEPTNVTTLSDSWDTDGSEMEEINLSQLHKHTDQNGNLKNVTHSNVKALGKSLEKQLAAR-GPKKP
Hsap DZIP1 -----GKN-TVKSDADGTEGSEIED-----TDDSPKPAQVA---VKTPTKEVKEKMFPHRKNVNP
      * . : : : * : : : : : : : : : : * . : : *

```

```

Dbuz CG13617 ---ERAQILLKNTPLQPMVVKSRNLSLTTDDDELNEGSDVTSASQGVVEDEEIEKASTSSKRQQK
Drer Iguana AGGVNT---FLEKPTDVRNTRQNAKELKYSDDDDDDDDD--WDISSLEDV---PAVAKPTQCPVPVRK
Hsap DZIP1 VGGTNPVEMFIKKE-----ELQELKCADVEDED-----WDISSLEEBEISLGKKSQKQKPPPAKN
      : : : : : * : : : * * : * : : * : : :

```

```

Dbuz CG13617 NFKAQLEQLLAKPAAHVSKPKLVQIQAKPMLPRKRVMFNTEGSRKNNEDSAE-
Drer Iguana SLDKSQDTSVWGSSTGKGHKPLTDAGTASTLKSSLVTVS-----DWDSDSEI
Hsap DZIP1 --EPHFAHVLNAGAFNPKGPKGEGLEQENESSTLKSSLVTVT-----DWDSDSDV
      . : . * : . * . * . . . : : * : :

```

FIGURE 38 | Alignment of *D. buzzatii* CG13617 with human DZIP1 and zebrafish Iguana proteins. Asterisks below the alignment indicate identical amino acids and dots conservative changes. Residues enclosed in a green box correspond to the C2H2 zinc finger. The two cysteine residues and the two histidines are clearly conserved in the three proteins (red asterisks). The orange boxes denote two blocks of three amino acids that present a striking level of conservation since they are found intact in all the sequences retrieved by the BLAST similarity searches (TABLE 13). Yellow boxes correspond to coiled coil regions. Purple boxes indicate the NLS and blue ones the NES. PEST sequences are enclosed in pink boxes. All protein motifs details were obtained from SEKIMIZU *et al.* (2004) and WOLFF *et al.* (2004) for Iguana, and MOORE *et al.* (2004) for DZIP1. Detailed information about some of the protein motifs in the human sequence could not be found in the literature, although preliminary analyses suggest that possibly it also contains coiled coils and nuclear signals. For each of these species, the protein isoform that aligned better with *D. buzzatii* CG13617 protein sequence has been represented here. A black arrow indicates the position of an alternative exon in human DZIP1 protein. The grey-shaded amino acids are an alternatively spliced intron in zebrafish Iguana protein sequence. GenBank accession numbers of the aligned sequences are: NP_055749.1 (isoform 1) for human DZIP1, Q7T019.2 (long isoform) for zebrafish Iguana, and AAT52021 for *D. buzzatii* st-1 line CG13617 protein sequence. Dbuz, *Drosophila buzzatii*; Hsap, *Homo sapiens*; Drer, *Danio rerio*.

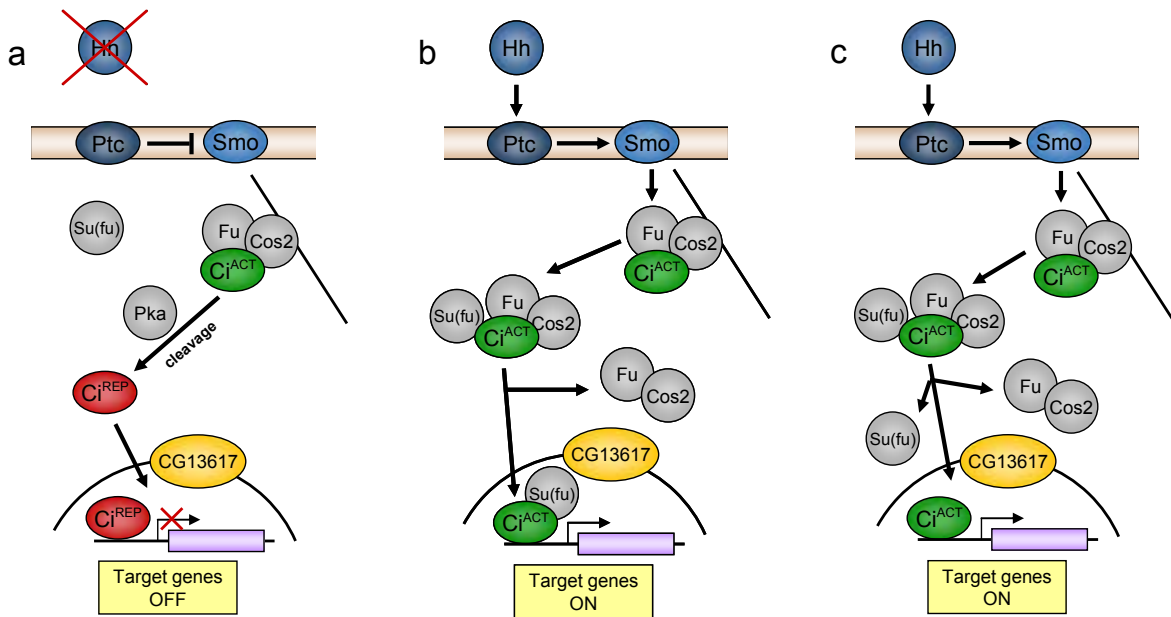


FIGURE 39 | Hedgehog signaling pathway in *Drosophila*. Blue and grey circles represent the different components of the Hh signaling pathway. See main text for the complete protein names. Ci transcription factor is depicted in green (active form) and red (repressor form). (a) When there is no Hh signal, the protein complex containing Ci remains anchored to the microtubules (straight black line) and Ci is cleaved to release the repressor form that accumulates into the nucleus and represses target genes. (b) When there are intermediate levels of Hh, the protein complex dissociates from microtubules and Ci binds Su(fu), an interaction that restricts its nuclear import and its activator activity. (c) High Hh levels promote dissociation of Su(fu) stimulating Ci transport to the nucleus where it activates transcription of target genes. CG13617, in yellow, has been added by the authors of this work in the step of the pathway where it might perform its function: the regulation of the nuclear access of both Ci isoforms. Based on Figure 8 of INGHAM and MCMAHON (2001).

can activate Hh target genes (FIGURE 39c). In the absence of Hh ligands, cAMP-dependent protein kinase 1 (Pka) phosphorylates Ci to promote its proteolytic cleavage, generating a repressor isoform able to repress expression of Hh target genes when transported into the nucleus (FIGURE 39a). The nuclear import of both the activator and repressor forms of Ci is regulated by Suppressor of Fused [Su(Fu)] (see FIGURE 39 for more details). The balance between these activator and repressor forms of Ci within the nucleus determines the specific target genes that the cell expresses in response to a particular level of Hh signaling.

Mutations in *igu* lead to a reduced expression of Hh target genes in the ventral neural tube of the zebrafish, but at the same time, cause an expanded expression pattern for Hh target genes in somites. This is thought to be due to different threshold responses to Hh signals. The ability to fully activate Hh target genes seems to be impaired in all tissues in *igu* mutants but the remaining low levels of Hh signaling appear to be sufficient to activate target genes in some tissues although not in the neural tube, where higher levels of Hh signaling are required. The two available studies on zebrafish Iguana (SEKIMIZU *et al.* 2004, WOLFF *et al.* 2004) propose that Iguana/DZIP1 plays a role in regulating the nuclear levels of Gli (the ortholog of *Drosophila* Ci in zebrafish) and therefore the ratios between the active and repressor forms. Iguana/DZIP1 has domains like the zinc finger and the coiled coil regions capable of mediating protein-protein interactions, and seems to be able to shuttle between the cytoplasm and the nucleus in a manner correlated with Hh pathway activity (WOLFF *et al.* 2004). So, Iguana/DZIP1 may act directly in nuclear import or indirectly by sequestering Gli factors in the cytoplasm, affecting both positive and negative regulation of Hh signaling (VOKES and MCMAHON 2004). Su(fu) is another protein involved in regulation of the nuclear import of Ci/Gli and mutants show phenotypes similar to *igu* deficient individuals. The hypersensitivity of *igu* mutants to Su(fu) alterations suggests a dual role or a certain redundancy for Iguana/DZIP1 and Su(fu) in mediating nuclear transport of the two forms of Ci/Gli (VOKES and MCMAHON 2004).

CG13617 could be performing a similar function to vertebrate Iguana/DZIP1 in *Drosophila*. Although in zebrafish Iguana seems to be required to achieve the complete activation of Hh target genes, we do not know if the same happens in *Drosophila*. It is noteworthy that according to the microarray results, even though most of the well-known

targets of Hh signaling pathway (*ptc*, *en*, *vg*, *dpp*, *ato*) are expressed in the studied *D. melanogaster* samples, they do not alter significantly their expression level when *CG13617* expression is silenced (results not shown). The only differentially expressed gene that is known to respond to Hh signaling at least in some tissues is *CycE* (DUMAN-SCHEEL *et al.* 2002), which is down-regulated both in the *D. melanogaster* RNAi experiments and in *D. buzzatii* 2j individuals. Cyclin E is an important regulator of S phase of the cell cycle (when DNA replication takes place) during *Drosophila* development (KNOBLICH *et al.* 1994) (FIGURE 40a). DUMAN-SCHEEL *et al.* (2002) have demonstrated that Ci is capable of inducing *CycE* expression in *Drosophila* eye and wing, through direct binding to three Ci-specific binding sites in the *CycE* promoter. Thus, Hh signaling is able to promote the S phase of the cell cycle by inducing *CycE* expression. If the lack of *CG13617* impairs Hh signaling pathway and causes the down-regulation of *CycE*, this could affect cell cycle progression by hindering the G1/S transition. The observed decrease in expression of the rest of DNA replication genes, which act during S phase, could be an

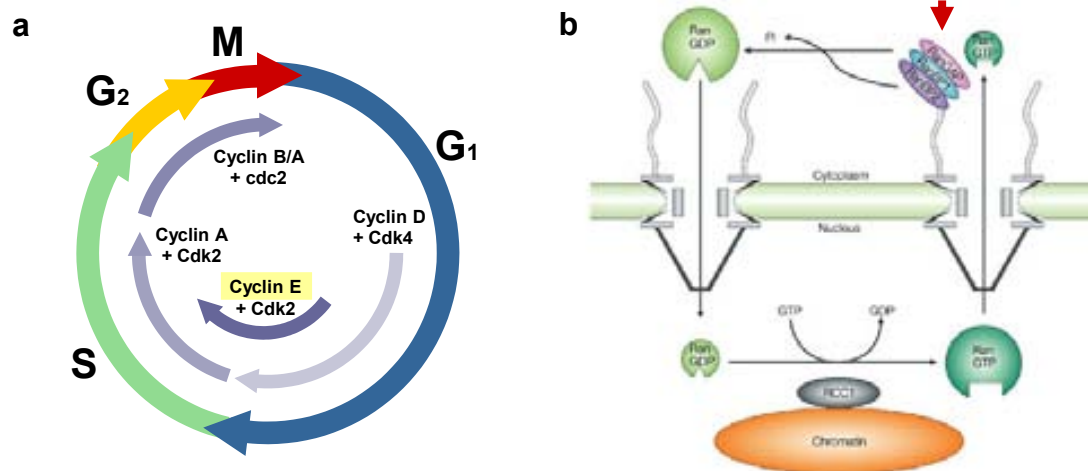


FIGURE 40 | Cyclin E and RanGap cellular functions. (a) Phases of the cell cycle and associated cyclins and cyclin-dependent kinases (Cdk). Cyclin E (highlighted in yellow) binds cyclin-dependent kinase Cdk2 and initiates transition from the growth phase G₁ to the S phase, when DNA replication takes place. (b) Ran protein is involved in the transport of proteins across the nuclear envelope by binding to importins and changing their ability to bind or release cargo molecules. Ran exists in the cell in two nucleotide-bound forms: GDP-bound and GTP-bound. The cytoplasmic localization of RanGap (marked by a red arrowhead) induces GTP hydrolysis in this cellular compartment, while chromatin-bound RCC1 in humans (Bj1 in *Drosophila*) catalyzes the exchange of GTP for GDP on Ran in the nucleus. These processes create a concentration gradient that is essential for some cellular processes like the assembly of the spindle and nuclear envelope during mitosis or to establish the directionality of nuclear transport. Both *RanGap* and *Bj1* transcripts are down-regulated to some extent in samples where *CG13617* gene is silenced (although *Bj1* presents a lower fold change and is not included in the list of 41 differentially expressed genes). Panel b of the figure extracted from FAHRENKROG and AEBI (2003).

indirect consequence of Cyclin E not functioning properly. In this regard, it is remarkable the absence of the nematode *Caenorhabditis elegans* from the list of species that possess genes with significant similarity to CG13617 protein (TABLE 13). As can be seen in FIGURE 37, there is a relatively high number of available nematode genomes (with 13 completed genomes) compared to other groups of species, but no sequence with similarity to CG13617 protein was obtained for this phylum in the similarity searches. This could be related to the fact that the *C. elegans* genome does not have any *hh* ortholog and therefore no Hh signaling (INGHAM and MCMMAHON 2001).

The unaffected expression of many Hh signaling target genes suggests that this pathway is not significantly altered in individuals not expressing *CG13617*. Therefore, it is possible that, if *CG13617* acts during Hh signaling, another protein (for example Su(fu), which is also involved in nuclear trafficking) is able to perform its function in its absence. This redundant activity would agree with the fact that the lack of *CG13617* in embryos does not seem to cause any important defects. Alternatively, *CG13617* may not be part of the Hh signaling route. In this case, *CG13617* could be regulating the nuclear import of other transcription factors different than Ci, possibly transcription factors regulating more directly the expression of genes involved in DNA replication. One example of such a transcription factor would be E2f. This transcription factor plays a very important role in cell division control and activates the transcription of many genes involved in DNA replication (TRIMARCHI and LEES 2002). In a microarray study of E2f-depleted cells, 12 genes of our differentially expressed list (*mus209*, *CycE*, *Mcm2*, *Mcm5*, *Mcm7*, *RnrL*, *RnrS*, *RanGap*, *Ts*, *dnk*, *CG7670* and *ichn*) were found to be down-regulated (DIMOVA *et al.* 2003). An additional 3 genes (*Hel25A*, *RfC40* and *His2Av*) that show fold changes in our experiments between 1.5 and 1.8 (below the cutoff of the differentially expressed genes list) are also regulated by E2f in the same manner. Most of these genes are included in the DNA replication or cell cycle functional categories in the gene ontology classification. The fact that all these genes present a reduced expression level in absence of both *E2f* and *CG13617* expression, makes E2f a good candidate for a transcription factor whose activity could be regulated by *CG13617*, for example, by controlling its access to the nucleus. E2f is not differentially expressed in our microarrays. However, if the molecular function of *CG13617* was related to nuclear transport, we would not expect to find the proteins directly regulated by *CG13617* as expression changes

in the microarray experiments, since mainly alterations in the transcriptional regulation of a gene can be detected as different amounts of transcript. Other regulatory mechanisms acting at post-transcriptional or protein levels that do not translate into changes in the quantity of a given mRNA will go unnoticed.

Nuclear transport is an essential process for the cell since mRNAs have to be exported to the cytoplasm to be translated and nuclear proteins need to be imported into the nucleus in order to perform their functions. It also represents a critical step in the regulation of gene expression, because the nuclear envelope can be used as a barrier to control access of transcriptional regulators to their target genes. A simple way to regulate the transcription of a gene in response to a signaling pathway is to retain transcription factors in the cytoplasm until a specific signal triggers their access to the nucleus, where they can activate the corresponding target genes. In particular, control over nuclear import of transcription factors is a well-established strategy to regulate gene expression downstream of developmental signaling pathways (SISSON *et al.* 2006, DONG *et al.* 2007), which suggests that the nuclear transport machinery can also play important roles during development (MASON and GOLDFARB 2009). For example, proper function of *Bj1* and *RanGap* are essential for *Drosophila* development and several loss-of-function mutations of genes that are part of the Ran pathway result in developmental defects or are lethal (MASON and GOLDFARB 2009).

Transport of either proteins or RNAs across the nuclear envelope can be bi-directional, and an increasing number of proteins have been identified that shuttle continuously back and forth between the nucleus and the cytoplasm. These shuttling proteins are essential factors in conveying information on nuclear and cytoplasmic activities within the cell and they play a critical role in the regulation of cell cycle progression and control of cellular proliferation (GAMA-CARVALHO and CARMO-FONSECA 2001). Nucleocytoplasmic shuttling proteins have typically both NLS and NES, and include transport receptors and adaptors, transcription factors, cell cycle regulators and numerous RNA binding proteins (GAMA-CARVALHO and CARMO-FONSECA 2001). Therefore, the presence of both classes of nuclear signals in CG13617 suggests that this protein could be able to shuttle between the nucleus and cytoplasm, and the zinc finger indicates that it could bind RNA or other proteins. In addition, the overall sequence similarity to Iguana/DZIP1, a protein whose function seems

to be regulating the access of a certain transcription factor to the nucleus, contributes even further to support a molecular function of *CG13617* related to nuclear transport, where it could operate as a regulator, controlling the access of transcription factors to their target genes. Finally, it is suggestive that other proteins involved in nuclear transport, such as RanGap (RanGTPase-activating protein), an important member of the Ran pathway (FIGURE 40b), are down-regulated as well in the absence of *CG13617* expression. Besides RanGap, two other genes performing nuclear transport functions are included in the extended list of differentially expressed genes with lower fold changes. They are *Bjl*, which encodes the other key enzyme (together with RanGap) involved in the maintenance of the Ran gradient (FIGURE 40b), and *Hel25A*, which produces a RNA helicase involved in the export of mRNAs from the nucleus. The reduced expression levels of these genes indicates that *CG13617* could be directly or indirectly regulating the transcription of other genes essential for nuclear transport, a cellular process in which *CG13617* itself might also play a role.

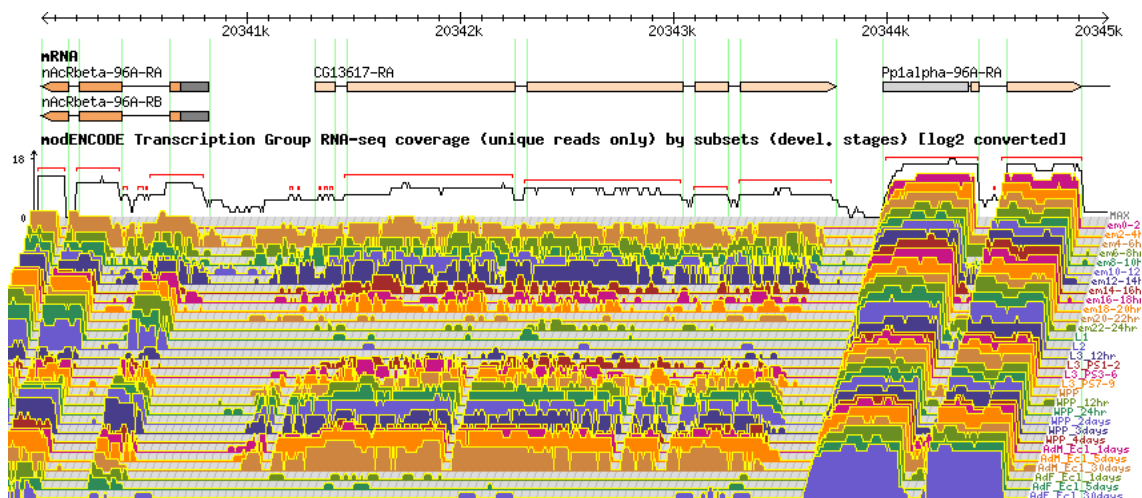



FIGURE 41 | *D. melanogaster* RNA-Seq developmental profile. Preview of data from the Developmental Stage Timecourse Transcriptional Profiling with RNA-Seq (modENCODE Project) (CELNIKER *et al.* 2009) performed on polyA⁺-RNA from 30 developmental stages spanning the life cycle of *D. melanogaster*, from 0-2 h embryos through 30-day adults. On top of the image the coding regions of *CG13617* and part of the flanking genes are represented as light orange rectangles and UTRs as grey-shaded boxes. Below, the transcriptional profile of each developmental stage is shown in a different line and color. The ages of the individuals are shown at the right of each line written in the corresponding color. As can be seen in the graph, *CG13617* is expressed in embryos (except for the earlier stages) and its expression decreases in larvae only to increase again in the pupal and adult stages. It is notable that *CG13617* expression is not detected in adult females (3 last lines) even though it is clearly expressed in males. Image obtained from FlyBase [↑](#).

As mentioned above, the other characterized protein showing similarity to CG13617 is human DZIP1 (DAZ-interacting protein 1) (MOORE *et al.* 2004). The *DZIP1* gene encodes three different protein isoforms that contain a C2H2 zinc finger domain (FIGURE 38). It is expressed predominantly in human embryonic stem cells and fetal and adult germ cells, although expression was also detected in fetal brain (MOORE *et al.* 2004). Dzip1 may associate with DAZ (Deleted in Azoospermia), a factor required early in germ cell development to maintain initial germ cell populations, both in embryonic stem cells and germ cells. Consistent with this, according to FlyAtlas  (CHINTAPALLI *et al.* 2007), a database about gene expression in *D. melanogaster* adults, CG13617 is mainly expressed in testis at this developmental stage, even though its gene-expression pattern in *D. buzzatii* embryos appears to be more general and does not indicate an expression restricted to germ cells or its precursors (SUPPORTING FIGURE 7 of PUIG *et al.* 2004). Similarly, recent RNA-Seq data from the Developmental Stage Timecourse Transcriptional Profiling (modENCODE project) (CELNIKER *et al.* 2009) reveal that CG13617 does not seem to be expressed in adult females, even though expression is detected in males, as well as in embryos and pupae (the level of expression is lower in the larval stages, also consistent with our *D. buzzatii* data) (FIGURE 41). In relation to this, it is noteworthy that the male germ line is one of the few proliferating adult tissues where a great number of cell divisions (and therefore DNA replication) take place. So, this expression pattern in adults could be reinforcing CG13617 role as a protein involved in the DNA replication process.

4.2.3 CG13617 and inversion 2j evolutionary history

Inversion 2j is widespread in *Drosophila* natural populations (HASSON *et al.* 1995) and can be considered evolutionary successful. In this work we have described a position effect associated with the breakpoints of this inversion. The expression change of gene CG13617 provides a molecular mechanism able to link this chromosomal rearrangement with the phenotypical differences in adult body size and developmental time observed between 2st and 2j arrangements (BETRÁN *et al.* 1998) and, consequently, with the adaptive value of inversion 2j.

The reduction in *CG13617* expression level is caused by the presence of an antisense transcript that has been associated to *2j* chromosomes. Two main events have contributed to the generation of an antisense RNA overlapping gene *CG13617*, and therefore, to the detected expression change: the *GalileoK* insertion and inversion *2j* itself. If the insertion of this particular TE copy happened first, then the inversion of a portion of chromosome 2 by ectopic recombination in the adjacent *GalileoG* fragment situated *GalileoK* in the current position and orientation, which allowed the antisense transcription of *CG13617*. Alternatively, *GalileoK* could have inserted in an already inverted chromosome. In this case, the antisense transcription, with the consequent silencing of *CG13617*, might have resulted in a favorable mutation that swept all other *2j* chromosomes in the population, causing also an increase in inversion frequency. This last possibility would explain that the different *2j* alleles at the breakpoints are younger than the inversion itself (CÁCERES *et al.* 2001). Consistent with this, LAAYOUNI *et al.* (2003) state that their nucleotide diversity analysis is compatible with a scenario where historical frequencies of inversion *2j* have remained quite low during most of the time, and that its rise in frequency, likely due to selection, has only occurred recently. Thus, the selected favorable change could be *CG13617* silencing in embryos caused by the insertion of *GalileoK* inside the initial *GalileoG* copy that originated inversion *2j*. However, for such an escenario to be possible, the reduction of *CG13617* expression level in *2j* embryos should cause a phenotypical change in traits affecting fitness.

Developmental time and adult body size are two traits of adaptive importance. Faster development can increase fitness in many insects like *Drosophila* either by an increase in larval survival in wild conditions or a demographic advantage for early reproduction (LEWONTIN 1965). A positive correlation has been reported in *D. buzzatii* between body size and longevity, mating success and fecundity, three major fitness components (SANTOS *et al.* 1992). In *Drosophila*, large flies have also been shown to present lower metabolic rates, higher desiccation tolerance and greater dispersal ability (SANTOS *et al.* 1992). Based on their effects on fitness, developmental time and body size are connected by a trade-off because, everything else being equal, it takes a longer time to grow to a larger size (ROFF 2000). Since a larger body size requires a longer development, a certain individual will experience either the advantages of having a short development or the ones derived from a large body size. In *D. buzzatii*, the phenotypic effects of inversion *2j* are related to these two traits. Inversion *2j* carriers have a

larger adult body size (RUIZ *et al.* 1991) but also a longer development when compared to *2st* individuals (BETRÁN *et al.* 1998). This means that the carriers of the *2st* arrangement are smaller as adults but they benefit from a shorter development. Therefore, the two chromosomal arrangements are thought to be maintained as a balanced polymorphism in *D. buzzatii* natural populations because of a trade-off between developmental time and body size.

We have shown in this work that as a consequence of *CG13617* natural (in *D. buzzatii* *2j* embryos) or experimental (in *D. melanogaster*) silencing, the expression levels of several genes involved in DNA replication and regulation of cell cycle are decreased as well. This indicates that *CG13617* protein participates directly or indirectly in the regulation of these processes (additional pieces of evidence linking gene *CG13617* to these activities come from its putative promoter sequences, which might contain a TFBS for Rfx, a transcription factor also known to bind the promoter of other genes that take part in DNA replication, and maybe a DRE element). The importance of some of the affected genes in both biological processes suggests that it is possible that their lower expression leads to a reduced DNA replication rate. Studies in yeast, plants, mammalian cells or *Drosophila* (WEIGMANN *et al.* 1997, NEUFELD *et al.* 1998) have shown that cells forced to divide more slowly become much larger than controls because cell growth continues for a longer period of time (SU and O'FARRELL 1998). For example, *D. melanogaster giant (gt)* mutants, which exhibit an extended period of growth during the third larval instar that allows them to reach twice the size of wild type larvae and generate giant adults, have a reduced rate of DNA synthesis measured as the incorporation of DNA precursors (NARACHI and BOYD 1985). In addition, *gt* flies also show many single-strand and double-strand breaks, further indicating an altered DNA metabolism (NARACHI and BOYD 1985). Larvae carrying another mutation with similar phenotypical effects, *l(2)gl*, also exhibit an slowed DNA synthesis (but less so than in *gt*). In both cases, these observations correlate with a slowed growth and development. Besides, in *gt* larvae, the increased body size appears to be caused by an increase in cell size and not in cell number (SIMPSON and MORATA 1980). Therefore, the consequences of slowing down DNA replication and cell division could be: (1) larger cells, which could result in larger animals, and (2) longer developmental times, since it would take longer to complete the cell divisions required to generate a fully developed individual. Interestingly, a larger adult body size and longer developmental times are precisely the two phenotypical characteristics that distinguish inversion *2j* carriers from *2st* individuals.

Nonetheless, this hypothesis has some drawbacks. In the first place, *CG13617* expression change in *2j* lines takes place in embryos, a developmental stage where cell proliferation occurs without growth (the ~50000-cell larvae that hatches from the egg has roughly the same size that the initial single-cell embryo) (O'FARRELL 2004). So, most of the embryonic cells decrease in size with each division, rather than doubling their mass, although it is also true that some cell types of the embryo clearly grow, like neuroblasts, which enlarge considerably in the early embryo (EDGAR and NIJHOUT 2004). And in second place, adult body size in *Drosophila* is largely determined by the size of the larva at the time of pupation. Thus, if body size is effectively regulated by the mechanism that controls the timing of the onset of metamorphosis in third instar larvae, how could an expression change that takes place in embryos affect this process? In larvae, if development is delayed, pupation takes place later, giving rise to larger individuals. For example, *Drosophila giant (gt)* mutants have reduced hormonal signals for pupation, slow-growing imaginal discs (groups of cells from which adult structures like wings or legs will develop) and delayed metamorphosis (SCHWARTZ *et al.* 1984). Therefore, these larvae grow to a giant size before undergoing metamorphosis because it takes longer for the discs to achieve their final size. Also, in studies on imaginal disc regeneration in *Drosophila* larvae, it has been shown that regenerating discs are able to delay pupation until regeneration is complete (SIMPSON *et al.* 1980). So, while imaginal discs are growing, larvae can not initiate metamorphosis but they keep growing nonetheless (EDGAR and NIJHOUT 2004). If for some reason imaginal discs growth takes longer to complete, the resulting animal is larger. Since no significant expression differences have been detected so far between *2st* and *2j* lines in larvae, the detected *CG13617* expression change is probably not affecting imaginal disc development. However, similar mechanisms could be operating during embryonic development, so that that slowing down the growth of a certain tissue or group of cells could cause a delay in the development of the whole animal. As a result of a longer developmental time, a total larger size could be reached later on. In the case of inversion *2j*, gene *CG13617* reduced expression in *2j* embryos could be the primary cause of a delay in development starting at the embryonic stage. Also, we need to take into account that the earlier a change occurs in development the more unpredictable and major the consequences can be, since they can be affecting a larger number of cells and in the critical moment of formation of the organism.

So, *CG13617* silencing seems to affect the expression levels of several genes involved in DNA replication and in cell cycle, and this change could be the basis for an increase of both developmental time and adult body size in inversion *2j* carriers. It is not the first time that genes involved in these processes have been associated to variation in size. In humans, large-scale genome-wide association studies have recently identified 44 *loci* known to influence normal variation in height (WEEDON and FRAYLING 2008). Despite that the causal variants at each of these *loci* have not yet been determined, in many cases there is a strong case for a particular gene being responsible for this effect. An analysis of the candidate genes has given some insights into gene groups and molecular processes that are important in normal human growth. As can be seen in FIGURE 42, there is an over-representation of genes involved in cell cycle, nucleic acid metabolism and developmental processes among the height-associated genes (WEEDON and FRAYLING 2008). Within the category of developmental genes, the

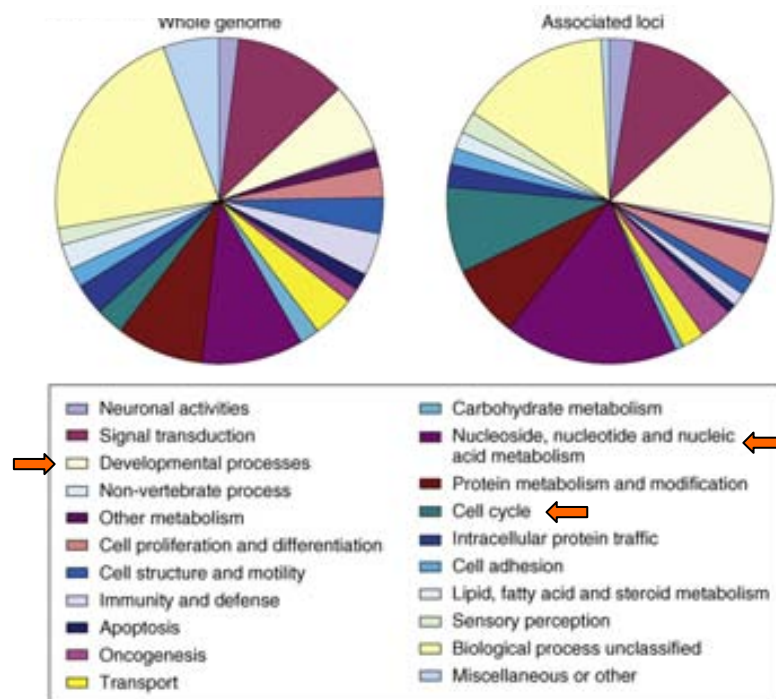


FIGURE 42 | Biological processes in which human genes associated to variation in height are involved compared with all genes in the genome. All genes included between flanking recombination hotspots in 44 *loci* known to influence normal variation in human height are considered height-associated genes. Functional categories are based on PANTHER classification 🏠. There is a clear over-representation of genes in three categories: cell cycle, nucleic acid metabolism and developmental processes. Figure extracted from WEEDON and FRAYLING (2008).

authors highlight that there are several components of the Hedgehog signaling pathway, which is also crucial for vertebrate patterning and development. Remarkably, the over-represented functional categories are the same we found in the list of differentially expressed genes due to *CG13617* silencing (FIGURE 22 and TABLE 7) and the analysis of *CG13617* molecular function by searching possible homologous proteins in other species indicates that it is an ortholog of a member of the Hh signaling pathway. Therefore, it is indeed possible that a modification of the expression level of genes involved in processes like nucleic acid metabolism or cell cycle effectively contributes to a change in size, since in humans, genes belonging to these functional categories have an effect on height too. Unfortunately, the underlying mechanism by which certain alleles of these genes (some of which have been proven to be differentially expressed in humans or mice) are able to affect the size of an organism remains unknown.

In the case of *D. buzzatii*, the phenotypical differences between *2st* and *2j* carriers are small, although it does not mean that they can not be biologically relevant. For example, the thorax length measurements (as an indication of adult body size) show an increment of 1.5-2% in *2j* individuals with respect to *2st* adults, but this difference is consistent and statistically highly-significant (RUIZ *et al.* 1991, NORRY *et al.* 1995, BETRÁN *et al.* 1998, FERNÁNDEZ IRIARTE and HASSON 2000). As previously said, we need to take into consideration that inversion *2j* is a natural inversion that arose in a natural population of *D. buzzatii* and that has passed the filter of natural selection. For this reason, we do not expect the inversion to cause any major phenotypic changes. Any mutation altering fitness-related traits, such as size or developmental time in a dramatic way, is likely to be eliminated. The changes (for example loss-of-function mutations) that result in noticeable phenotypical effects often turn out to be viable in laboratory conditions but they would probably cause a significant reduction of fitness in nature and they would be eliminated by natural selection (FISHER 1930). On the contrary, subtle changes have increased chances of remaining in the population and, if at a given moment they turn out to be useful, they can be positively selected.

The alteration experienced by gene *CG13617* in *2j* chromosomes has a regulatory nature, which means that the silencing is not permanent and can be restricted to embryos by expressing the antisense RNA only in this developmental stage, while its expression can be restored to normal levels in larvae, pupae and adults. The fact that this change affects gene

regulation and not coding sequences reduces greatly its possible pleiotropic effects and facilitates that the positive effects it may have on fitness can be selected, causing an increase in its frequency (CARROLL 2005, WRAY 2007). Besides, the dominant effect of the antisense RNA acting in *trans* to silence both copies of gene *CG13617* in heterokaryotypes (SUPPORTING TABLE 6 of PUIG *et al.* 2004) could facilitate the detection of this expression change by natural selection if it was accompanied by the corresponding phenotypical effect in individuals carrying only one copy of the inverted chromosome. This would provide an advantage for the initial increase of inversion *2j* frequency when compared to recessive mutations, since homozygous individuals carrying a new mutation on which natural selection can act will only appear in the population if a mutation has previously achieved a certain frequency through stochastic processes. Even though the evolutionary potential of gene regulation has long been recognized (KING and WILSON 1975), only recently regulatory changes in expression patterns or levels that contribute to the evolution of specific traits have been identified (AVEROF and PATEL 1997, STERN 1998, SHAPIRO *et al.* 2004, GOMPEL *et al.* 2005, CLARK *et al.* 2006). The reduction in the expression level of gene *CG13617* in *2j* chromosomes caused by the TE-induced antisense RNA could be influencing significant traits such as body size and developmental time and therefore it might represent another example of a regulatory change able to affect the fitness of carrier individuals.

The evidence of position effects in natural inversions is scarce, being the hypotheses based on the maintenance of linkage disequilibrium between *loci* located within the inverted segment traditionally favored as the mechanism by which inversions are able to affect phenotype and be selected. However, we provide an example that shows that a position effect on a single gene caused by a natural inversion can have functional consequences on other genes and trigger a more global response able to affect traits important for fitness. Yet, it is possible that different mechanisms are responsible for the increase in the inversion frequency in different moments of its evolution. For instance, the position effect could have provided an initial advantage but the genetic isolation between rearrangements within the inverted sequences could have facilitated the establishment of coadapted gene complexes that complemented the effect of *CG13617* silencing. Alternatively, initial effects caused by locally adapted alleles found within the inverted segment could have been accentuated by a position effect appearing later on because of the insertion of the *GalileoK* copy that originates antisense

transcription. Therefore, the existence of position effects should also be considered (together with the rest of mechanisms, as well as their combinations) when searching for the genes responsible for the adaptive value of other natural inversions.

In this work we have demonstrated the existence of such a position effect caused by inversion *2j* breakpoint on the adjacent gene *CG13617*. This gene shows a 5-fold reduction of its expression level in embryos carrying the inversion when compared to *2st* ones, a difference caused by the presence of an antisense RNA transcribed from the TEs inserted at the breakpoint junction. In turn, *CG13617* silencing seems to induce the down-regulation of several genes involved in DNA replication and regulation of cell cycle. This generates slightly different expression profiles in embryos of the *2st* and *2j* lines that could be the genetic basis of the phenotypic differences in size and developmental time that distinguish both arrangements, and could represent at least part of the molecular mechanism responsible for the adaptive value of this inversion. Future studies doing deeper comparisons of the expression levels in both rearrangements, a more extensive functional analysis of the differentially expressed genes, and more formal phenotypic characterizations of transiently *CG13617*-depleted individuals would provide a better understanding of these questions.

CONCLUSIONS

Conclusions

The following conclusions can be drawn from this work:

1. *CG13617* is a functional gene expressed through the whole *D. buzzatii* life cycle that is transcribed into a 2.3-kb mRNA and encodes a 734-amino acid protein in this species. This gene is conserved across the genus *Drosophila* although there is variation in gene structure and the overall level of sequence conservation is not very high.
2. *CG13617* shows a 5-fold reduction of its expression level in embryos carrying inversion *2j* with respect to embryos with the *2st* arrangement, but none of the other developmental stages (larvae, pupae and adults) showed significant differences between lines with inverted and non-inverted chromosomes.
3. *CG13617* silencing in *2j* embryos is caused by an antisense RNA that overlaps the whole *CG13617* coding region and presents an expression level that is 5 times higher in *2j* embryos when compared to *2st* individuals. The antisense RNA is transcribed exclusively in the embryonic stage from a *GalileoK* copy inserted at the proximal breakpoint junction that has been found in all analyzed *2j* chromosomes but not in non-inverted ones.
4. *CG13617* silencing causes the down-regulation of several genes involved in DNA replication and regulation of cell cycle both in *CG13617*-depleted *D. melanogaster* and in *D. buzzatii 2j* embryos. However, no differences were detected between arrangements in the expression levels of any of these genes in *D. buzzatii* first instar larvae, where the difference in *CG13617* expression level between *2st* and *2j* lines is lower than in embryos and the antisense RNA is no longer transcribed.

5. *D. buzzatii* CG13617 protein contains a C2H2 zinc finger, three coiled coil regions, two PEST sequences, and putative nuclear localization and export signals. These features are mostly conserved in the CG13617 protein of other *Drosophila* species.
6. CG13617 protein presents sequence similarity and shares the presence of several functional motifs with human DZIP1 and zebrafish Iguana (a component of the Hedgehog signaling pathway) proteins, which suggests that CG13617 might be involved in the nuclear transport of transcription factors.
7. The reduction in *CG13617* expression level in embryos is shared by all *D. buzzatii* *2j* lines and might be at the basis of the phenotypic differences in size and developmental time that distinguish both arrangements. By slowing down DNA replication and cell cycle and causing a delay in embryonic development, *CG13617* silencing might increase both developmental time and the final adult body size, both characteristic of individuals carrying inversion *2j*.

APPENDIX

I

Molecular Characterization of Two Natural Hotspots in the *Drosophila buzzatii* Genome Induced by Transposon Insertions

Mario Cáceres,¹ Marta Puig, and Alfredo Ruiz

Departament de Genètica i de Microbiologia, Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona), Spain

Transposable elements (TEs) have been implicated in the generation of genetic rearrangements, but their potential to mediate changes in the organization and architecture of host genomes could be even greater than previously thought. Here, we describe the naturally occurring structural and nucleotide variation around two TE insertions in the genome of *Drosophila buzzatii*. The studied regions correspond to the breakpoints of a widespread chromosomal inversion generated by ectopic recombination between oppositely oriented copies of a TE named *Galileo*. A detailed molecular analysis by Southern hybridization, PCR amplification, and DNA sequencing of 7.1 kb surrounding the inversion breakpoints in 39 *D. buzzatii* lines revealed an unprecedented degree of restructuring, consisting of 22 insertions of ten previously undescribed TEs, 13 deletions, 1 duplication, and 1 small inversion. All of these alterations occurred exclusively in inverted chromosomes and appear to have accumulated after the insertion of the *Galileo* elements, within or close to them. The nucleotide variation at the studied regions is six times lower in inverted than in noninverted chromosomes, suggesting that most of the observed changes originated in only 84,000 years. *Galileo* elements thus seemed to promote the transformation of these, otherwise normal, chromosomal regions in genetically unstable hotspots and highly efficient traps for transposon insertions. The particular features of two new *Galileo* copies found indicate that this TE belongs to the *Foldback* family. Together, our results strengthen the importance of TEs, and especially DNA transposons, as inducers of genome plasticity in evolution.

[The sequence data described in this paper have been submitted to the GenBank data library under accession nos. AF368842–AF368859 and AF368861–AF368900. In addition, sequences submitted under accession nos. AFI62796–AFI62799 were used as a basis for this study.]

Transposable elements (TEs) are intrinsic components of the genomes of all living organisms, from the simplest prokaryotes to the most complex eukaryotes (Berg and Howe 1989; Capy et al. 1998). They make up a substantial fraction of most studied genomes, although TE content varies widely in different species and tends to be positively correlated with total genome size (Hartl 2000). Current sequencing projects are revealing the precise organization of genomes and how repetitive sequences are distributed and arranged within them. In the euchromatin, TEs are usually found scattered as individual repeats interspersed with single-copy sequences. The chromosomal arms of *Drosophila melanogaster*, for example, contain sporadic TE insertions separated by long stretches of unique DNA (Ashburner et al. 1999; Adams et al. 2000; Benos et al. 2000). In the human genome around 35%–45% of the euchromatic portion is taken up by TEs, mainly SINEs and LINEs, more or less randomly distributed in a short period interspersion pattern (Lander et al. 2001; Venter et al. 2001). Heterochromatic regions located around centromeres and telomeres of eukaryote chromosomes, however, show a very different organization. These regions consist almost exclusively of repeated sequences and harbor a great accumulation of TE sequences. A well-known case is the pericentromeric

heterochromatin of *D. melanogaster*, where, besides simple sequence repeats, there are many different families of mostly rearranged TEs interspersed with very little unique DNA (Gatti and Pimpinelli 1992; Pimpinelli et al. 1995; Adams et al. 2000).

Traditionally, TEs have been considered as junk DNA or mere genomic parasites, exploiting cells for their own propagation (Doolittle and Sapienza 1980; Orgel and Crick 1980). However, though probably as indirect consequences of their existence (Charlesworth et al. 1994), TEs exert a great variety of effects on the genome of their hosts and could have played a very important role in the shaping of the genetic material during evolution (Finnegan 1989; McDonald 1995; Kidwell and Lisch 1997). TEs are a major source of mutation and genetic variation by getting inserted into coding sequences or regulatory regions of genes. These insertions are generally deleterious for the organism, as happens in many *Drosophila* phenotypic mutants (Lindsley and Zimm 1992) and several human genetic diseases (Wallace et al. 1991; Holmes et al. 1994), but some have been involved in new gene expression patterns and even new genes with apparently beneficial effects (Britten 1996, 1997; Lander et al. 2001). Moreover, TEs possess the ability to promote genetic recombination between homologous sequences and can produce large-scale chromosomal rearrangements (Lim and Simmons 1994; Gray 2000). Specifically, TEs have been implicated in the origin of some natural chromosomal inversions in different organisms, such as bacteria (Daveran-Mingot et al. 1998), yeast (Kim et al. 1998),

¹Corresponding author.

E-MAIL caceres@salk.edu; FAX (858) 558-7454.

Article published on-line before print: *Genome Res.*, 10.1101/gr.174001.
Article and publication are at <http://www.genome.org/cgi/doi/10.1101/gr.174001>.

flies (Cáceres et al. 1999), and hominids (Schwartz et al. 1998).

One of the most outstanding examples of natural variation in chromosome structure is the extraordinarily rich inversion polymorphism in the species of the *Drosophila* genus. Hundreds of polymorphic inversions have been described in *Drosophila*, and these inversions do not distribute at random among species or among chromosomal elements within species (Krimbas and Powell 1992). Furthermore, the breakpoints of inversions are not randomly distributed along chromosomes either (Krimbas and Powell 1992; Cáceres et al. 1997). Despite the fact that not all naturally occurring inversions have TEs at their breakpoints (Wesley and Eanes 1994; Cirera et al. 1995), inversion breakpoints have been found to be associated with TE insertion sites in *D. melanogaster* (Lyttle and Haymer 1992; Andolfatto et al. 1999), *D. willistoni* (Regner et al. 1996), and the *D. virilis* group (Evgen'ev et al. 2000), and direct evidence for the implication of TEs in the origin of chromosomal inversions has been obtained both in the laboratory (Lim and Simmons 1994) and in nature (Cáceres et al. 1999). Therefore, it has been suggested that TEs could be responsible for the hotspots where repeated breaks have been observed (Krimbas and Powell 1992; Evgen'ev et al. 2000). However, the molecular confirmation of the existence of the hotspots and the elucidation of their anatomy have remained elusive.

Recently, we cloned and sequenced the breakpoints of a highly successful chromosomal inversion of *D. buzzatii*, inversion *2j*, that was originated by ectopic recombination between oppositely oriented copies of a TE (Cáceres et al. 1999). This inversion inverted a central segment of the 2 standard (*2st*) chromosomal arrangement, the ancestral arrangement of chromosome 2 for all of the *D. buzzatii* cluster species (Ruiz and Wasserman 1993), comprising around one-fourth of its euchromatic fraction. In all *2j* chromosomes both inversion breakpoints were found to contain large insertions that were absent from the noninverted *2st* chromosomes. Because these insertions fulfilled all characteristic features of TEs (Capy et al. 1998), they were considered copies of a new transposon that was named *Galileo*. However, the insertion at the proximal

breakpoint exhibited a very complex structure, with copies of several different internal repeats in an apparently chaotic arrangement. In addition, a preliminary study revealed that some variation in the structure of both breakpoint insertions existed among inverted chromosomes. Thus, the further characterization of the *2j* breakpoints offered the opportunity to get a deeper insight into the molecular nature of inversion breakpoints and to investigate the long-term effects that TE insertions raised up to a high frequency might have on the organization of the genome.

Here, an exhaustive molecular analysis of the *2j* breakpoint regions in 9 lines with *2st* chromosomes and 30 lines with the *2j* inversion has uncovered an amazing degree of naturally occurring structural variation among *2j* chromosomes, caused by the insertion of multiple TEs inside each other, deletions, and other small DNA rearrangements. The observed structural diversity contrasts with the low level of nucleotide variation, suggesting that the structural changes have accumulated in a short period of time. Therefore, the breakpoints of inversion *2j* appear to be highly variable hotspots.

RESULTS

Structural Variation at Inversion *2j* Breakpoint Regions

Figure 1 shows the breakpoint regions of inversion *2j* in the two *D. buzzatii* lines that were previously characterized, *st-1* and *j-1* (Cáceres et al. 1999). In *2st* chromosomes the breakpoint regions have been designated as *AB* (distal breakpoint) and *CD* (proximal breakpoint). Inversion *2j* took place between *A* and *B* sequences and between *C* and *D* sequences, and the breakpoint regions in *2j* chromosomes consist of *AC* (distal breakpoint) and *BD* (proximal breakpoint). Large insertions not present in *2st* chromosomes are found in the chromosomes with the inversion between *A* and *C* sequences and between *B* and *D* sequences. In this study, several molecular techniques with increasing resolution power and accuracy were sequentially used to examine the structure of the *2j* breakpoints in other *2st* and *2j* lines: Southern blot hybrid-

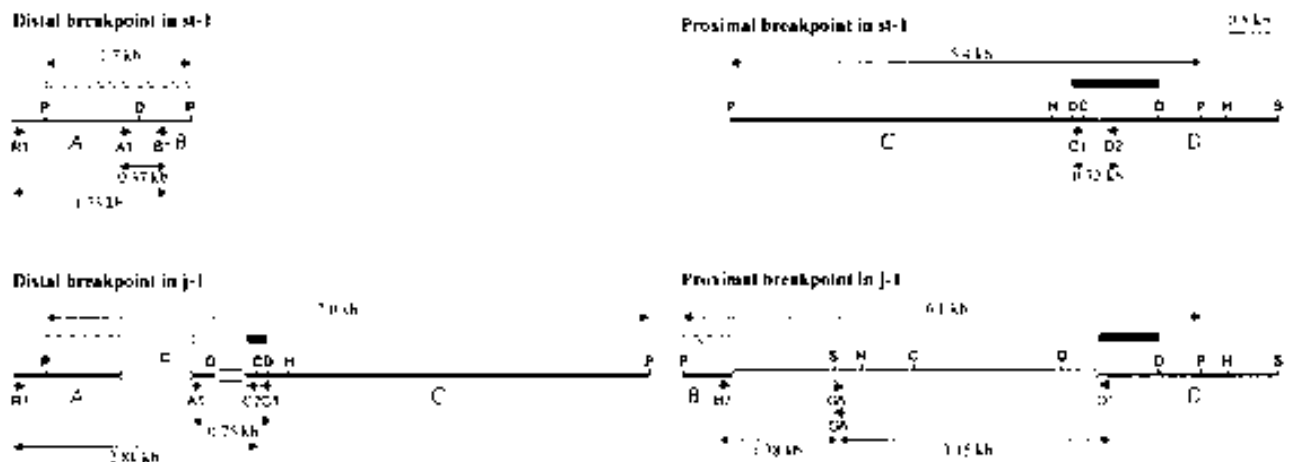


Figure 1 Physical map of the distal and proximal *2j* breakpoint regions in the *st-1* and *j-1* lines. Thick lines represent the single-copy *A*, *B*, *C*, and *D* sequences. TE insertions are represented as empty boxes. Hatched and black rectangles correspond, respectively, to the *AB* and *CD* probes used for the Southern hybridization analysis. Small arrows represent primers used in the PCR amplification. Some of the restriction sites found in this region are shown: *C*, *Clal*; *D*, *Dral*; *H*, *HindIII*; *P*, *PstI*; *S*, *Sall*.

ization, PCR amplification of different segments, restriction mapping of the PCR products, and DNA sequencing.

No structural variation in the *AB* or *CD* regions was found between nine *2st* lines of diverse geographic origins. Southern blot hybridization of *Pst*I-digested genomic DNA with *AB* and *CD* probes revealed in all *2st* lines the same bands of 1.7 kb and 5.4 kb, respectively, corresponding to the distal and proximal *2j* breakpoint regions (Fig. 1). PCR amplification of the 1.73-kb R1–B1 and 0.37-kb A1–B1 segments (distal breakpoint) or the 0.32-kb C1–D2 segment (proximal breakpoint) did not show any size variation between the *2st* lines either. Restriction mapping of the PCR products corroborated the absence of differences within each segment.

Clearly contrasting results were found in *2j* chromosomes. First, variation in the restriction map of the breakpoint regions in 30 *2j* lines was analyzed by Southern blot hybridization. Genomic DNA of all *2j* lines was digested with *Pst*I and hybridized with a *CD* probe. Two hybridization bands were observed in each of the *2j* lines, corresponding to the proximal and distal breakpoints with their respective insertions, and remarkable variation was detected among them: There were 11 bands of different sizes for the proximal breakpoint, whereas there were 6 different bands for the distal breakpoint (Table 1). For those lines whose *Pst*I hybridization

pattern did not coincide with that of *j*-1 (Fig. 1), a more detailed restriction map of the breakpoint region was elaborated by repeated Southern hybridization using additional restriction enzymes (*Cl*aI, *Dra*I, *Eco*RI, *Eco*RV, *Hin*dIII, *Sal*I, and *Xba*I) and *AB* and *CD* probes. This resulted in the identification of nine main structural types in the proximal breakpoint and six in the distal breakpoint (Table 1).

In the PCR analysis of the *2j* lines, smaller regions, containing just the breakpoint insertions and the adjacent single-copy DNA, were studied. Primer pairs B2–G6 and G5–D1 (proximal breakpoint) and R1–C2 and A1–C1 (distal breakpoint) were used with genomic DNA of all *2j* lines (Fig. 1). The PCR products of each line were compared by gel electrophoresis and were digested with restriction enzymes to detect and map any variation existing between them (Table 1). The PCR results revealed a small difference between two lines (*j*-16 and *jz*³-4) belonging to one of the previous nine structural types defined in the proximal breakpoint and between several lines previously ascribed to the same structural type of the distal breakpoint, but otherwise confirmed the restriction maps obtained from the Southern hybridizations. However, two problems arose in the PCR amplifications. First, *Taq* DNA polymerase sometimes jumped between distant parts of certain DNA templates, causing an excision of the intervening segment. By

Table 1. Molecular Analysis by Southern Blot Hybridization and PCR Amplification of the *2j* Breakpoint Regions of the 30 *2j* Lines Used in This Study

Name	Geographic origin	Hybridization bands (kb)		PCR products (kb)			
		Proximal	Distal	B2-G6	G5-D1	R1-C2	A1-C1
<i>j</i> -1	Carboneras (Spain)	6.1	7.0	1.38	1.92	2.81	0.75
<i>j</i> -2	Carboneras (Spain)	6.1	7.0	1.38	1.92	2.81	0.75
<i>j</i> -3	Carboneras (Spain)	6.1	7.0	1.38	1.92	2.81	0.75
<i>j</i> -4	Carboneras (Spain)	6.1	7.0	1.38	1.92	2.81	0.75
<i>j</i> -5	Carboneras (Spain)	6.1	7.0	1.38	1.92	2.81	0.75
<i>j</i> -6	Carboneras (Spain)	6.1	7.0	1.38	1.92	2.81	0.75
<i>j</i> -7	Caldetas (Spain)	6.1	7.0	1.38	1.92	2.81	0.75
<i>j</i> -8	San Luis (Argentina)	8.5	7.0	4.15	2.13	2.83	0.77
<i>j</i> -9	Quilmes (Argentina)	5.0	8.5	1.32	2.07	4.34	2.28
<i>j</i> -10	Palo Labrado (Argentina)	5.1	9.0	1.38	2.13	—	—
<i>j</i> -11	Los Negros (Bolivia)	8.8	7.0	1.32	2.07	2.83	0.77
<i>j</i> -12	Guaritas (Brazil)	8.8	7.0	1.32	2.07	2.83	0.77
<i>j</i> -13	Guaritas (Brazil)	8.8	7.0	1.32	2.07	2.81	0.75
<i>j</i> -14	Laboratory (Australia)	6.1	7.0	1.38	1.92	2.83	0.77
<i>j</i> -15	Catamarca (Argentina)	6.1	7.0	1.38	1.92	2.81	0.75
<i>j</i> -16	Salta (Argentina)	12.1	7.0	1.38	—	2.83	0.77
<i>j</i> -17	Tilcara (Argentina)	6.0	7.0	1.38	—	2.83	0.77
<i>j</i> -18	Termas Rio Hondo (Argentina)	5.0	7.0	1.32	2.07	2.83	0.77
<i>j</i> -19	Ticucho (Argentina)	10.3	8.9	1.32	2.25	—	—
<i>j</i> -20	Hemmant Australia)	6.1	7.0	1.38	1.92	2.81	0.75
<i>j</i> -21	Hemmant (Australia)	6.1	7.0	1.38	1.92	2.81	0.75
<i>j</i> -22	Trinkey (Australia)	8.8	7.0	1.32	2.07	2.83	0.77
<i>jz</i> ³ -1	Carboneras (Spain)	9.9	7.0	1.32	3.11	2.83	0.77
<i>jz</i> ³ -2	Carboneras (Spain)	9.9	7.0	1.32	3.11	2.81	0.75
<i>jz</i> ³ -3	Kariouan (Tunisia)	9.9	7.0	1.32	3.11	2.83	0.77
<i>jz</i> ³ -4	Tilcara (Argentina)	8.3	9.2	1.34	—	—	—
<i>jq</i> ⁷ -1	Carboneras (Spain)	7.5	7.0	1.36	—	2.81	0.75
<i>jq</i> ⁷ -2	Mogan, Canary Islands (Spain)	7.5	11.0	1.36	—	3.62	1.56
<i>jq</i> ⁷ -3	Caldetas (Spain)	7.5	7.0	1.36	—	2.81	0.75
<i>jq</i> ⁷ -4	Otamendi (Argentina)	6.1	7.0	1.38	1.92	2.83	0.77

Hybridization bands are those obtained by Southern hybridization of *Pst*I-digested genomic DNA of each line with the *CD* probe. Proximal and distal refer to the proximal and distal breakpoint, respectively. Proximal breakpoint bands indicated in boldface include a 3.8-kb extra segment due to a polymorphism in a *Pst*I site. Products of each PCR were digested with different restriction enzymes: B2-G6, *Bam*HI–*Eco*RI; G5-D1, R1-C2, and A1-C1, *Dra*I.

sequencing the G5-D1 PCR products of lines j-1 and j-19 we showed that two different ~1-kb deletions have occurred during the amplification. In both cases the deletions were found to take place between short homologous sequences repeated in direct orientation that were contained within long inverted repeats. Thus, the PCR excision mechanism resembles that of spontaneous deletion by slippage during DNA replication (Farabaugh et al. 1978; Albertini et al. 1982), which is stimulated by the formation of stem-loop secondary structures (Egner and Berg 1981). On the other hand, no amplification occurred in some of the 2*j* lines (Table 1) and other combinations of primers different of the previous ones were assayed. Nevertheless, a few breakpoint segments could not be amplified either with the new combinations of primers or with PCR conditions specially designed for the amplification of difficult templates (see Methods).

As a final step, we sequenced the regions that were found to differ between 2*j* lines (Fig. 2). Fragments showing varying restriction patterns were cloned and sequenced completely from the corresponding PCR products. However, when two or more 2*j* lines did not show any variation in the restriction map of a particular region, only the DNA of one of them was sequenced as representative. A thorough effort was made to isolate and characterize all segments in which differences have been detected. Therefore, for those segments that were not PCR-amplified or that suffered deletions during PCR, we turned to traditional cloning. Two λ genomic libraries of the j-19 and jz³⁻⁴ lines were constructed and in both lines the two breakpoints of inversion 2*j* were isolated. Those segments differing with regard to the other 2*j* lines in each breakpoint were cloned and sequenced.

Altogether, the Southern blot hybridization and PCR data allowed us to infer the structures present at the breakpoints of the 30 2*j* lines studied, and DNA sequencing let us fully identify the changes that differentiate them (Fig. 2). Ten different structural types were found in the proximal breakpoint and seven in the distal breakpoint, and most of them were related by relatively simple changes, such as insertions or deletions of DNA segments. Thus, with this information we were able to postulate a plausible evolutionary sequence of changes between the breakpoint structures. To better illustrate the changes, five hypothetical variants (Hyp) have been represented as intermediaries between the observed ones. Also, for the sake of simplicity, we have considered that all insertions occurred independently, although a few of them could have originated in a single event. In the proximal breakpoint, the simplest structure is that of Hyp-P1, which contains a *Galileo* insertion between B and D sequences with three other TEs inserted inside (Fig. 2A). All of the TEs inside *Galileo* are flanked by direct repeats, presumably generated by the duplication of the target site during the insertion event, with the only exception of *BuTI*. In the latter case, the absence of

the outermost nucleotide of the right inverted terminal repeat (ITR), suggests that a deletion after the *BuTI* insertion removed its last base pair, the right target site duplication, and part of the left long ITR of *Galileo* (see below). From Hyp-P1, eight large insertions of seven different TEs, eight deletions, and the inversion of an internal segment are required to generate the structural diversity actually seen in the proximal breakpoint (see Fig. 2A for details). In the distal breakpoint, the simplest structure is that of j-12, formed by a 392-bp *Galileo* insertion between A and C sequences and an *ISBu1* insertion in A (Fig. 2B). From here, eight insertions of seven different TEs, five deletions and a small duplication should have occurred to explain the other six structural variants observed (see Fig. 2B for details).

The most important features of the 22 large insertions (named from i1 to i22) found at the breakpoints of inversion 2*j* are summarized in Table 2. The target site duplications flanking most insertions, the presence of multiple copies, and the variation found among lines identify the inserted DNA sequences as TEs (Capy et al. 1998). According to sequence similarities between the inserted sequences, we have recognized ten different previously undescribed TEs (that will be described in detail elsewhere). Apart from the original *Galileo*-1 and *Galileo*-2 insertions that were implicated in the generation of inversion 2*j* (Cáceres et al. 1999), there are two more *Galileo* copies inserted at the 2*j* breakpoints, *Galileo*-3 and *Galileo*-4. These new *Galileo* copies are basically composed of very long ITRs, with a relatively small and heterogeneous central region that does not seem to encode any protein involved in their transposition. Like the first two copies, they do not show homology to any known sequence in the available databases, but they display significant structural similarity to the *Foldback* elements described in many organisms (Bingham and Zachar 1989; Hoffman-Liebermann et al. 1989; Hankeln and Schmidt 1990; Yuan et al. 1991; Rebatchouk and Narita 1997), including the ability to form stable secondary structures when denatured (as indicated by the difficulties encountered in the PCR amplification of the segments containing these elements). Five other insertions corresponding to two closely related TEs (average sequence identity 84%) also show similarities to *Foldback* elements. These new elements have been named *Kepler* and *Newton* and share many of their characteristics with *Galileo* (average sequence identity 73%), suggesting that they belong to the same family: (1) The terminal 40 bp of their ITRs are identical (except for one single nucleotide difference); (2) all of them tend to duplicate 7 bp of the target site upon insertion (Table 2); and (3) *Newton* elements exhibit very long ITRs resembling those of *Galileo* elements. Moreover, insertions i10 to i17 correspond to four different TEs that can be ascribed to Class II (Finnegan 1989; Capy et al. 1998) and have been designated as *D. buzzatii* transposons or BuTs. Based on sequence ho-

Figure 2 Schematic representation of the structures found at the proximal (A) and distal (B) breakpoints of inversion 2*j* in the 30 2*j* lines studied. All different structures are shown, except for that of j-16 in the proximal breakpoint, which differs from jz³⁻⁴ by the absence of d6 deletion. Thick lines represent the single-copy A, B, C, and D sequences. TEs are represented as colored boxes and sharp ends correspond to the ITRs. Insertions and deletions are delimited by green and red lines, respectively, and are named with an i or a d followed by a number. Target site duplications flanking the insertions are shown above them. Blue lines indicate the inversion of an internal segment. Arrows below the diagrams inform on the orientation of some homologous segments. Segments sequenced in each structure are enclosed within clear rectangles. Only the *D. buzzatii* lines representative of each structural variant are shown. Lines sharing the same structure in the proximal breakpoint are jq⁷⁻¹, jq⁷⁻², and jq⁷⁻³; j-1, j-2, j-3, j-4, j-5, j-6, j-7, j-14, j-15, j-20, j-21, and jq⁷⁻⁴; j-9, j-11, j-12, j-13, j-18, and j-22 (deletion d2 was detected during j-12 sequencing and we do not know whether it is present in other lines or not); jz³⁻¹, jz³⁻², and jz³⁻³. Lines sharing the same structure in the distal breakpoint are j-1, j-2, j-3, j-4, j-5, j-6, j-7, j-13, j-15, j-20, j-21, jz³⁻², jq⁷⁻¹, and jq⁷⁻³; j-8, j-11, j-12, j-14, j-16, j-17, j-18, j-22, jz³⁻¹, jz³⁻³, and jq⁷⁻⁴. Hyp are hypothetical structures not found in our sample of 2*j* lines. Small black arrows are PCR primers used in the study.

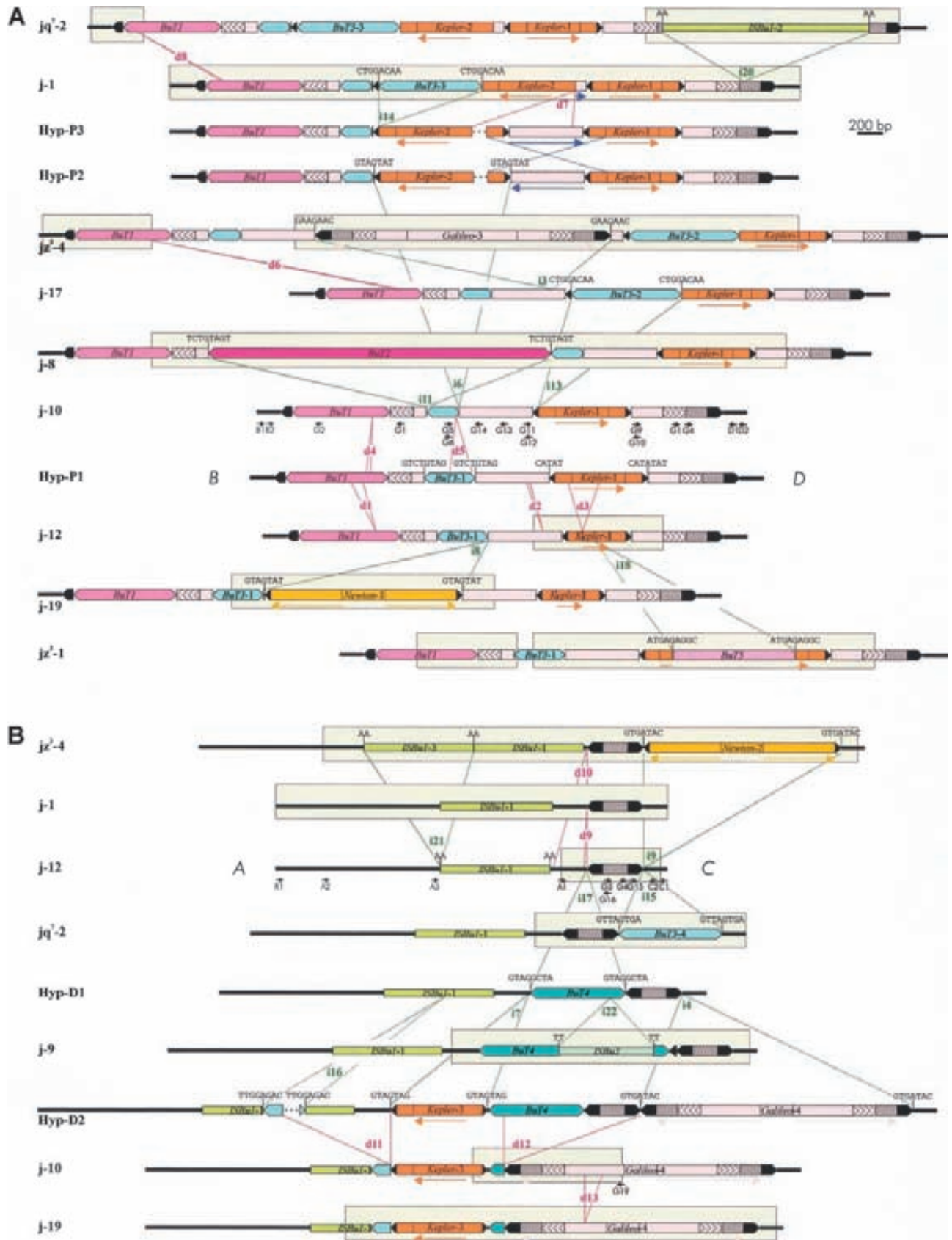


Figure 2 (See facing page for legend.)

Table 2. TE Insertions at the Breakpoint Regions of Inversion 2j of *Drosophila buzzatii*

Insertion	TE	Size (bp)	ITRs (bp)	Target site (bp)	BP
<i>Foldback-like elements</i>					
i1	<i>Galileo-1</i>	1589	228/443	7	P
i2	<i>Galileo-2</i>	392	106	7	D
i3	<i>Galileo-3</i>	2204	683/684	7	P
i4	<i>Galileo-4</i>	2083	918/916	ND	D
i5	<i>Kepler-1</i>	722	150	5	P
i6	<i>Kepler-2</i>	735	ND	7	P
i7	<i>Kepler-3</i>	692	20	ND	D
i8	<i>Newton-1</i>	1510	572/575	7	P
i9	<i>Newton-2</i>	1512	575/574	7	D
<i>hobo, Activator, Tam3 (hAT) elements</i>					
i10	<i>BuT1</i>	801	15/14	ND	P
i11	<i>But2</i>	2775	12	8	P
i12	<i>BuT3-1</i>	413	23	8	P
i13	<i>BuT3-2</i>	844	23	8	P
i14	<i>BuT3-3</i>	798	23	8	P
i15	<i>BuT3-4</i>	795	23	8	D
i16	<i>BuT3-5</i>	147	ND	ND	D
i17	<i>BuT4</i>	721	24/23	8	D
Unclassified elements					
i18	<i>BuT5</i>	1039	3	9	P
i19	<i>ISBu1</i>	841	—	2	D
i20	<i>ISBu2</i>	1467	—	2	P
i21	<i>ISBu3</i>	853	—	2	D
i22	<i>ISBu2</i>	726	—	2	D

Elements have been classified by structural and sequence similarities with described TEs according to Capy et al. (1998). When different, the size of the left and right inverted terminal repeats (ITRs) are indicated. BP refers to the location of the element in the proximal (P) or distal (D) breakpoint. ND, data that could not be determined due to deletions.

mologies they have been included in the *hAT* superfamily (Calvi et al. 1991). *BuT1* and *BuT2* show similarity to the element *Gandalf* of *D. koepferae* (Marín and Fontdevila 1995), whereas *BuT3* and *BuT4* are related to the element *Hopper* of *Bactrocera dorsalis* (Handler and Gomez 1997). Finally, five insertions could not be neatly classified into any of the previously known TE families. *BuT5* ends in ITRs of just three base pairs (followed by subterminal imperfect inverted repeats of 17 bp), generates 9-bp duplications during insertion, shows a moderately repetitive pattern by in situ hybridization to *D. buzzatii* polytene chromosomes (J.M. Ranz, pers. comm.), and has been tentatively considered a Class II TE. The other four insertions belong to a new class of highly repetitive mobile elements, whose members do not possess ITRs and seem to duplicate two base pairs upon insertion. We have called them *ISBu* elements because of their structural and sequence similarity to the IS elements of the species of the *obscura* group of *Drosophila* (Hagemann et al. 1998).

Several other types of genetic rearrangements besides the multiple TE insertions have been found at the 2j breakpoints. We have detected 13 deletions of more than 17 bp (Fig. 2): d1, 93 bp; d2, 24 bp; d3, 238 bp; d4, 32 bp; d5, 179 bp; d6, 41 bp; d7, >536 bp; d8, 20 bp; d9, 17 bp; d10, 248 bp; d11, >649 bp;

d12, 1023 bp; and d13, 136 bp (the lengths of d7 and d11 are minimum estimates, as the real size of the deleted fragments is not known). Five of these deletions seem to have originated by the well-established mechanism of slipped-strand mispairing (Farabaugh et al. 1978; Albertini et al. 1982): d2, d3, and d6 took place between two repeated sequences of 3–4 bp, eliminating one of them and the intervening DNA; d8 and d13 removed one copy of a sequence of 20 bp and 136 bp, respectively, duplicated in tandem. A similar mechanism could also have generated the tandem duplication of the terminal 41 bp of *Galileo-2* in j-9 (Fig. 2B). Finally, in some of the 2j lines we have found a change of orientation of a 55-bp *Galileo-1* internal fragment, which suggests that an inversion has occurred inside the proximal breakpoint insertion (Fig. 2A). This inversion spanned ~600 bp and was probably generated by recombination between the oppositely oriented ITRs of *Kepler-1* and *Kepler-2* in Hyp-P2.

Nucleotide Variation at Inversion 2j Breakpoint Regions

In addition to the structural variation study, we sequenced 596 bp corresponding to the A, B, C, and D single-copy sequences in the nine 2st lines and 12 2j lines representing the diversity of structural types found. For comparison, we obtained the nucleotide sequence of the same regions in *D. martensis*, another species of the *D. buzzatii* complex (Ruiz and Wasserman 1993). These are seemingly noncoding intergenic regions, located 0.5–3.7 kb apart from the *rox8* (A), *Pp1 α -96A* (C), and *nAcR β -96A* (D) coding sequences (Cáceres et al. 1999). However, the last 112 bp of D show homology to a putative *D. melanogaster* ORF recently discovered (Adams et al. 2000) that would require further investigation. In the 12 2j lines we sequenced also 839 bp of the distal breakpoint insertion and the ends of the proximal breakpoint insertion. Figure 3 summarizes the 81 polymorphic sites found and Table 3 shows the estimates of the nucleotide diversity, π (Nei 1987), calculated ignoring sites with alignment gaps or missing data only in pairwise comparisons.

Considering the four single-copy regions together, nucleotide diversity is six times lower in 2j chromosomes than in 2st chromosomes (Table 3). We carried out computer simulations of the coalescent process using the *DnaSP* program (Rozas and Rozas 1999) to assess whether the nucleotide variation in each chromosomal arrangement was significantly different. Ten thousand trees were generated assuming the average number of nucleotide differences of 2st chromosomes, constant population size and no recombination, and a statistically significant probability of 0.01 of obtaining nucleotide diversity values as the one observed in 2j chromosomes or lower was found. In addition, 2st and 2j chromosomes exhibit a great number of fixed differences, including 17 nucleotide substitutions and six indels of 1–4 bp (TE insertions and target site duplications excluded). Using *D. martensis* as outgroup, a neighbor-joining tree (Saitou and Nei 1987) was built with the single-copy sequences of 2st and 2j lines (Fig. 4). All 2j sequences formed a monophyletic cluster of high bootstrap value, clearly separated from that of 2st sequences, confirming the proposed unique origin of the inversion (Cáceres et al. 1999).

No significant departures from the neutral model were found with the Tajima (1989) and Fu and Li (1993) tests, and nucleotide variation was used to date the origin of the inversion and of the sampled 2st and 2j alleles. The age of the

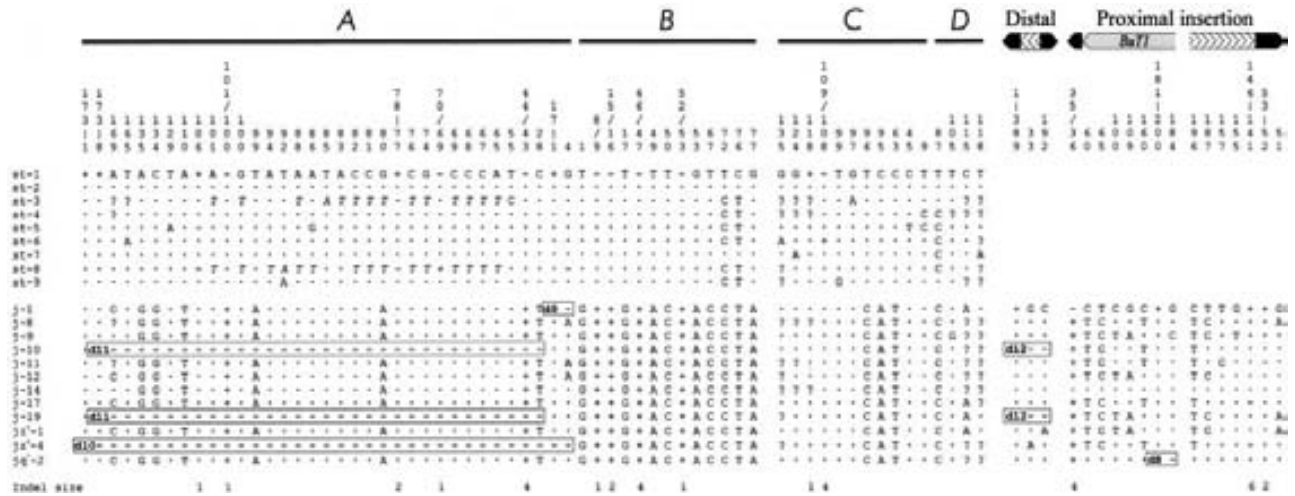


Figure 3 Nucleotide polymorphism at the breakpoint regions of inversion 2j. Nucleotide position is represented above the sequences. The breakpoints are taken as start point of A, B, C, D, distal breakpoint insertion, and proximal breakpoint insertion sequences. Nucleotides identical to the first sequence are indicated by a dot and missing data by a question mark. Deletions and insertions are indicated by minus and plus signs, respectively, and their size in base pairs is shown below. Gross deletions affecting the sequenced regions are named as in Fig. 2 and are included in rectangles. TE insertions and target site duplications are not shown. In 2st lines there is a 18-bp stretch between A and B sequences resembling *Galileo* footprints (Cáceres et al. 1999) that is not represented here either. Positions A65 to A101 in st-3 and st-8 accumulate multiple nucleotide changes with regard to the other lines and are shown in italics.

inversion was estimated from the fixed differences between 2st and 2j chromosomes. The average number of nucleotide differences, d_{xy} (Nei 1987), between 2st and 2j chromosomes is 0.0353 and between *D. buzzatii* and *D. martensis* is 0.1094. Subtracting from both figures the intraspecific polymorphism (0.0197), the net average number of nucleotide substitutions is obtained (Nei 1987). Combining the available information (Russo et al. 1995; Rodríguez-Trelles et al. 2000), we have estimated the divergence time between *D. buzzatii* and *D. martensis* as 5.8 million years (Myr) and this results in a rate of 7.7×10^{-9} nucleotide substitutions per site and per year for the breakpoint regions. Therefore, the 2j inversion should be ~1 Myr old, which is consistent with its widespread distribution through most *D. buzzatii* populations. The coalescence

Table 3. Nucleotide Variation in the Breakpoint Regions of Inversion 2j of *Drosophila buzzatii*

Region	Total (N = 21)			2st (N = 9)		2j (N = 12)	
	m	S	π	S	π	S	π
ABCD	596	35	0.0197	15	0.0075	3	0.0013
A	179	13	0.0251	5	0.0063	2	0.0189
B	143	9	0.0320	2	0.0076	0	0
C	155	9	0.0167	6	0.0104	0	0
D	119	4	0.0045	2	0.0055	1	0.0015
Insertions	839	—	—	—	—	13	0.0066
proximal	447	—	—	—	—	11	0.0096
distal	392	—	—	—	—	2	0.0007

Positions A65 to A101 of st-3 and st-8 lines, probably originated by some sort of genetic exchange, have been excluded from the estimation of the nucleotide diversity.
 N, number of sequences considered; m, maximum number of nucleotides sequenced in each region; S, number of segregating sites; π , average number of pairwise differences between sequences per nucleotide.

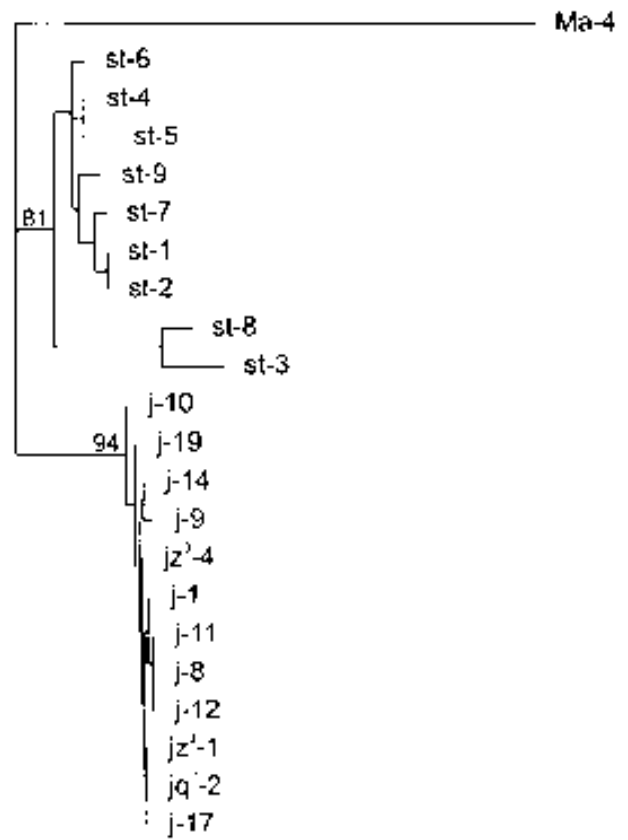


Figure 4 Neighbor-joining phylogenetic tree of the breakpoint sequences of inversion 2j based on the A, B, C, and D sequence data for the nine 2st and 12 2j *Drosophila buzzatii* lines. The Ma-4 *Drosophila martensis* line was used as outgroup. Bootstrap values in percentage out of 500 replicates are indicated for the main nodes.

time of *2st* and *2j* alleles was estimated from the average number of pairwise differences between the sequences of each chromosomal arrangement (Rozas et al. 1999). Accordingly, the sampled *2st* alleles are estimated to be 485,000 years old and the sampled *2j* alleles 84,000 years old.

Finally, we have used the Kreitman and Hudson's homogeneity test to detect differences in polymorphism levels between the studied regions (Kreitman and Hudson 1991). In the pooled set of 21 *2st* and *2j* sequences no significant differences in polymorphism across *A*, *B*, *C*, and *D* regions were found ($X^2_L = 2.86$, $df = 3$, $P = 0.41$). However, the TE sequences inserted at the proximal breakpoint accumulate strikingly higher nucleotide variation between *2j* chromosomes than the single-copy regions and the distal breakpoint insertion ($X^2_L = 8.61$, $df = 2$, $P = 0.01$). The difference between the polymorphism levels between *2j* chromosomes at the TE insertions of each breakpoint ($X^2_L = 4.00$, $df = 1$, $P = 0.04$), which are expected to be equally selectively constrained, suggests that there could be an intrinsic increased rate of nucleotide change at the proximal breakpoint insertion.

DISCUSSION

Our detailed analysis of the breakpoints of inversion *2j* has allowed us to characterize and reconstruct the evolutionary sequence of changes that has occurred in these regions. This study has revealed a great extent of genetic rearrangement at the breakpoints, consisting of 22 insertions of 10 different TEs, 13 deletions, a duplication, and an internal inversion. The low level of nucleotide variation at the single-copy sequences among *2j* chromosomes suggests that the different structures in each breakpoint were generated gradually from a common ancestor in a short period of time. According to the coalescence time of the sampled *2j* alleles, the changes that differentiate them, that is, 16 of the TE insertions, the 13 deletions, the duplication, and the internal inversion, are estimated to have occurred <84,000 years ago. Together with the inversion *2j* itself, this represents a rapid degree of genome restructuring never found before in nature and qualifies the *2j* breakpoints as genetically unstable hotspots.

Typically, the density of TE insertions in *D. melanogaster* euchromatin is low. The 2.9-Mb sequence from the *Adh* region (Ashburner et al. 1999) and the 2.6-Mb sequence from the tip of the X chromosome (Benos et al. 2000) display just

one insertion every 171 kb and 155 kb on average, respectively. These values coincide with the previous observed frequencies of polymorphic insertions in particular gene regions of *D. melanogaster* and other *Drosophila* species (Table 4). The frequency of insertions found at the *2j* breakpoints in *D. buzzatii* *2j* chromosomes is, however, ~100 times higher than the *D. melanogaster* average and ~40 times bigger than the highest frequency of insertions ever found in the genus *Drosophila*, that of the *vermillion* locus of *D. ananassae* (Table 4). This complex array of broken and rearranged TEs accumulated in the *2j* breakpoints in *2j* chromosomes clearly differs from the expected organization of ordinary euchromatin and resembles more closely some *D. melanogaster* heterochromatic regions (Miklos et al. 1988; Vaury et al. 1989; Devlin et al. 1990; Locke et al. 1999).

What is the cause of these hotspots? The structural diversity in *2j* chromosomes contrasts sharply with the lack of TE insertions and structural variation in the homologous regions of *2st* chromosomes and points to an effect of the inversion or of the initial *Galileo* insertions as most likely explanations for the hotspots. It has been argued that TEs should accumulate around inversion breakpoints because the reduction of recombination protects them from being eliminated by deleterious ectopic exchanges (Montgomery et al. 1987; Eanes et al. 1992; Sniegowski and Charlesworth 1994), and this could in part account for the insertions at the *2j* breakpoints. However, we think that the former explanation does not agree completely with our observations. First, TE insertions accumulate exclusively in very small regions around the *2j* inversion breakpoints. Of the 12.3 kb corresponding to the studied region in the *2j* ancestral chromosome, all TE insertions have accumulated just in the 5.1 kb comprised by the *Galileo-1*, *Galileo-2*, and *ISBu1-1* elements and none in the surrounding single-copy DNA. In the two other polymorphic inversions in which variation around the breakpoints was analyzed, *In(3L)P* and *In(2L)t* of *D. melanogaster*, only two TE insertions were found in 2.5 kb and 5 kb studied, respectively (Hasson and Eanes 1996; Andolfatto et al. 1999). Second, although differences in mobility levels may be involved, the complete absence among the TEs inserted in the *2j* breakpoints of retrotransposons, which seem to constitute the majority of TEs in *Drosophila* (Arkhipova et al. 1995), is noteworthy. Third, given the actual intermediate frequency

of inversion *2j*, the reduction in recombination is expected to affect *2st* and *2j* chromosomes in a similar way. Finally, the recombination reduction hypothesis does not account for deletions and other chromosomal rearrangements.

Accordingly, we favor the idea that the *Galileo* insertions were probably the main inducers of the generation of the hotspots. It is particularly remarkable that *Galileo* elements seem to belong to the *Fold-back* family. These elements have a distinctive internally repeated structure and the *FB* elements of *D. melanogaster* are characterized by the production of extremely unstable mutations and chromosomal rearrangements at unusually high frequencies in laboratory popula-

Table 4. Frequency of Naturally Occurring Insertions in Different *Drosophila* Species

Species	DNA analyzed	Frequency of insertions (insertions/kb/chromosome)	Reference
<i>D. buzzatii</i>			
<i>2st</i> chromosomes	7.1 kb ^a	0	This study
<i>2j</i> chromosomes	7.1 kb ^a	0.601 ^b	This study
<i>D. melanogaster</i>	578 kb	0.005	Charlesworth and Langley 1991
<i>D. melanogaster</i>	229 kb	0.004	Aquadro 1993
<i>D. simulans</i>	165 kb	0.0005	Aquadro 1993
<i>D. pseudoobscura</i>	32 kb	0	Aquadro 1993
<i>D. ananassae</i>			
<i>forked</i> locus	18 kb	0.004	Stephan and Langley 1989
<i>vermillion</i> locus	18 kb	0.017	Stephan and Langley 1989

^aFor *2st* and *2j* chromosomes, the length of the single-copy region analyzed by Southern hybridization of *Pst*I-digested DNA in *2st* chromosomes was considered.

^bOnly those insertions known to have occurred independently were computed.

tions (Bingham and Zachar 1989; Lovering et al. 1991). TE insertions, deletions, and the other DNA rearrangements are not distributed uniformly along the studied regions in *2j* chromosomes. Instead, they appear to have occurred after *Galileo-1* and *Galileo-2* insertions, within or very close to them (Fig. 2). Fourteen TEs out of 20 are inserted within *Galileo-1* or *Galileo-2* elements and all of the observed deletions occurred inside or at the ends of pre-existing *Galileo* or *Galileo*-like elements. The fact that all *2j* chromosomes share three TE insertions and one hypothetical deletion inside the *Galileo-1* element and an *ISBu1* insertion at the distal breakpoint is suggestive of the hotspots predating the origin of the *2j* inversion, but a population bottleneck affecting *2j* chromosomes could also be invoked.

There are several cases of nested insertion of TEs inside *Foldback* elements (Bingham and Zachar 1989; Hoffman-Liebermann et al. 1989). This sometimes has been interpreted as a mechanism to direct TE insertion outside of gene coding regions to reduce the damage inflicted to the host by their mobilization (Kidwell and Lisch 1997). Among Class II TEs, insertion site preference has been examined only for *D. melanogaster* *P* elements, which show some tendency to insert into accessible chromatin regions in the 5' end of genes and into pre-existing *P* copies (Engels 1996; Liao et al. 2000). Nevertheless, many more examples are known among retrotransposons. In *Saccharomyces cerevisiae*, *Ty1*, *Ty2*, *Ty3*, and *Ty4* elements are mostly located in regions upstream of tRNA genes and other genes transcribed by RNA polymerase III, whereas *Ty5* prefers to integrate near silent chromatin at the telomeres (Ji et al. 1993; Zou and Voytas 1997; Boeke and Devine 1998; Kim et al. 1998). In addition, blocks of nested retrotransposons are formed in the intergenic regions of the maize genome by repeated insertion of them inside each other. In particular, 14 of the 23 retrotransposons found in the *adh1-F* region were inserted within other retrotransposons (SanMiguel et al. 1996, 1998). Finally, there are also retrotransposons that seem to preferentially target heterochromatic regions, such as the *KERV-1* element of kangaroos (Vaugh O'Neill et al. 1998) or the *I* element of *D. melanogaster* (Dimitri et al. 1997).

On the other hand, TEs, and especially DNA transposons, are largely known to mediate the production of various types of genetic rearrangements, including deletions, duplications, and inversions, with high efficiency. In laboratory studies, *P* elements have been found to promote deletions and duplications of the flanking genomic sequences (Preston et al. 1996) and internal deletions of *P* DNA (Staveley et al. 1995), whereas deletions recovered from *mariner* elements usually affect the ITR of the element and the DNA where is inserted (Lohe et al. 2000). In both cases, extra DNA appears sometimes between the deletion endpoints, as happens in our d4 and d5 deletions, which were accompanied by the introduction of a new nucleotide. In addition, TEs are involved in promoting genetic recombination between homologous sequences (Sved et al. 1990; McCarron et al. 1994; Lohe et al. 2000). We have already shown that recombination between *Galileo* copies was implicated in the generation of inversion *2j* (Cáceres et al. 1999), and several other naturally occurring inversions in Diptera could have originated by a similar mechanism as well (Lyttle and Haymer 1992; Mathiopoulos et al. 1998; Andolfatto et al. 1999). At the molecular level, genetic instability might result from the presence of inverted repeats or the mechanism of transposition of the TEs inserted at the *2j* breakpoints. Excluding *ISBu1* and *ISBu2*, all of the other ele-

ments are thought to transpose by a conservative cut-and-paste mechanism (Finnegan 1989; Capy et al. 1998), in which DNA breaks induced by the transposase at the transposon ends could be aberrantly repaired by host repair functions, producing many different types of DNA alterations (Lohe et al. 2000). Either an increased mutation rate attributable to repeated repair events or an increased frequency of genetic exchange with other copies of the element could account for the higher nucleotide variation observed at the TE insertion of the proximal breakpoint.

Several lessons can be drawn from this work. We have been able to follow the effects of particular TE insertions on the genome through evolutionary time and to see how these TEs seem to have altered the dynamics of ordinary euchromatic regions, transforming them into highly unstable heterochromatin-like structures. Previously, insertion and expansion of *P* transposon transgenes in the *D. melanogaster* genome was found to induce local formation of heterochromatin and this was proposed to be caused by the pairing of adjacent repeats (Dorer and Henikoff 1994). Also, the TE clustering at the *2j* breakpoints is consistent with the retrotransposon associations found in *D. virilis* chromosomes by *in situ* hybridization (Evgen'ev et al. 2000) but challenges the prototypical picture of the *Drosophila* genome provided by *D. melanogaster* (Ashburner et al. 1999; Adams et al. 2000; Benos et al. 2000). An analogous disparity in TE distribution is found between two plant species with very different genome sizes, *Arabidopsis thaliana* and *Zea mays*. Similar to *D. melanogaster*, *A. thaliana* has a relatively small genome and is atypical in that most TEs are located in the pericentromeric region (Lin et al. 1999; Mayer et al. 1999). Our results are reminiscent of the explosive accumulation of 23 retrotransposons in the originally 80-kb *adh-1* region of maize over the last 6 Myr that resulted in the triplication of its size (SanMiguel et al. 1996, 1998). However, the TE insertion rate observed in the 7.1-kb *2j* breakpoint regions of *D. buzzatii* is even faster. The important effects that these blocks of TEs could have on genome evolution and the possibility that *Galileo* or other *Foldback* elements could be involved in analogous hotspots at other locations of the *D. buzzatii* genome are very interesting questions for further investigation.

METHODS

Drosophila Stocks

Thirty-nine lines of *D. buzzatii* and one of *D. martensis* were used in the study. The *D. buzzatii* lines (except *jq⁷⁻³* and *jq⁷⁻⁴*) are isogenic for chromosome 2 and bear one of four different 2 chromosome arrangements: *2st*, *2j*, *2jz³*, or *2jq⁷* (*2jz³* and *2jq⁷* derive from the *2j* arrangement and carry inversions *2z³* and *2q⁷*, respectively). These lines were isolated from different natural populations covering the whole range of the species distribution. The geographic origins of the *2st* lines are: st-1 and st-2, Carboneras (Spain); st-3, Vipos (Argentina); st-4, Guaritas (Brazil); st-5, Catamarca (Argentina); st-6, Salta (Argentina); st-7, Termas de Rio Hondo (Argentina); st-8, Ticucho (Argentina); and st-9, Trinkey (Australia). The geographic origin of the *2j* lines is given in Table 1. The *D. martensis* line (Ma-4) is from Guaca (Venezuela).

Southern Hybridization and Construction of Genomic Libraries

Southern hybridization was carried out by standard methods as described previously (Ranz et al. 1999). Two probes were used for the analysis of the *2j* breakpoint regions (Fig. 1). The

AB probe consists of a 1.7-kb *Pst*I fragment containing 1178 bp of *A* and 510 bp of *B* sequences, whereas the *CD* probe consists of a 0.9-kb *Dra*I fragment containing 242 bp of *C* and 715 bp of *D* sequences (Cáceres et al. 1999). Two genomic libraries of the *j*-19 and *j*³⁻⁴ *D. buzzatii* lines were constructed in the λGEM-11 vector (Promega) as described in Cáceres et al. (1999). To isolate the clones containing the *2j* breakpoints, these libraries were screened by plaque hybridization with the *AB* and *CD* probes.

PCR Amplification

For the PCR amplification, different pairs of oligonucleotide primers covering the entire regions of study were designed (see Table 5, available as an on-line supplement at <http://www.genome.org>, for sequence of primers). To specifically amplify the breakpoint insertions, primers that anneal to inserted repetitive sequences were always used in combination with primers located on the flanking nonrepetitive DNA. PCRs were carried out in a volume of 50 μl, including 100–200 ng of genomic DNA of each line, 20 pmoles of the different primers, 200 μM dNTPs, 1.5 mM MgCl₂, and 1–1.5 units of *Taq* DNA polymerase. Typical temperature cycling conditions were 30 rounds of 30 sec at 94°C, 30 sec at 50–70°C (depending on the primer pair used), and 60–180 sec at 72°C. Difficult templates that were not amplified with the normal PCR conditions were assayed with the GC-Rich PCR System (Roche), using 0.5–2 M GC-Rich resolution solution and an elongation temperature of 68°C.

DNA Sequencing and Sequence Analysis

DNA fragments of interest coming from restriction enzyme digestion or PCR amplification were cloned into Bluescript II SK (Stratagene) or pGEM-T (Promega) vectors, respectively. These fragments were sequenced on an ALFexpress (Amersham Pharmacia Biotech) or an ABI 373 A (Perkin-Elmer) automated DNA sequencer, using M13 universal and reverse primers. Nucleotide sequences were analyzed with the Wisconsin Package (Genetics Computer Group). *Bestfit* was used to align pairs of homologous sequences in different lines to detect inserted or deleted segments. Similarity searches through the GenBank/EMBL databases using *FASTA*, *BLASTX*, and *TBLASTX* were carried out to identify the inserted sequences. To analyze the nucleotide variation at the *2j* breakpoints, we sequenced the same regions as in Cáceres et al. (1999) in six additional *2st* lines and seven additional *2j* lines. Both strands of PCR-generated templates were sequenced completely with different pairs of primers (Table 5, available as an on-line supplement at <http://www.genome.org>). Sequences were multiply aligned with *Clustal W* (Thompson et al. 1994). Polymorphism analysis was performed using the *DnaSP* program (Rozas and Rozas 1999). Phylogenetic analysis was performed using the *PHYLP* software package (J. Felsenstein).

ACKNOWLEDGMENTS

We are deeply indebted to J.M. Ranz for the data on the repetitive nature of *BuT5* and general advice at all stages of this work. J.S.F. Barker kindly provided us with 15 of the *D. buzzatii* stocks used. J. Rozas greatly contributed to improve the nucleotide variation analysis. We also thank A. Barbadilla for helpful discussion of results, and M. Ashburner, A. Berry, P. Capy, F. Casares, A. Navarro, and D. Petrov for valuable comments and suggestions. Work was supported by grant PB98-0900-C02-01 from the Dirección General de Investigación Científica y Técnica (Ministerio de Educación y Cultura, Spain) awarded to A.R. and a doctoral FI fellowship from the Comissionat per a Universitats i Recerca (Generalitat de Catalunya, Spain) awarded to M.C.

The publication costs of this article were defrayed in part

by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

- Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., et al. 2000. The genome sequence of *Drosophila melanogaster*. *Science* **287**: 2185–2195.
- Albertini, A.M., Hofer, M., Calos, M.P., and Miller, J.H. 1982. On the formation of spontaneous deletions: The importance of short sequence homologies in the generation of large deletions. *Cell* **29**: 319–328.
- Andolfatto, P., Wall, J.D., and Kreitman, M. 1999. Unusual haplotype structure at the proximal breakpoint of *In(2L)t* in a natural population of *Drosophila melanogaster*. *Genetics* **153**: 1297–1311.
- Aquadro, C.F. 1993. Molecular population genetics of *Drosophila*. In *Molecular approaches to fundamental and applied entomology* (eds. J. Oakeshott and M.J. Whitten), pp. 222–266. Springer-Verlag, New York.
- Arkhipova, I.R., Lyubomirskaya, N.V., and Ilyin, Y.V. 1995. *Drosophila retrotransposons*. Springer-Verlag, Heidelberg.
- Ashburner, M., Misra, S., Roote, J., Lewis, S., Blazej, R., Davis, T., Doyle, C., Galle, R., George, R., Harris, N., et al. 1999. An exploration of the sequence of a 2.9-Mb region of the genome of *Drosophila melanogaster*: The *Adh* region. *Genetics* **153**: 179–219.
- Benos, P.V., Gatt, M.K., Ashburner, M., Murphy, L., Harris, D., Barrell, B., Ferraz, C., Vidal, S., Brun, C., Demailles, J., et al. 2000. From sequence to chromosome: The tip of the X chromosome of *Drosophila melanogaster*. *Science* **287**: 2220–2222.
- Berg, D.E. and Howe, M.M. 1989. *Mobile DNA*. American Society for Microbiology, Washington, D.C.
- Bingham, P.M. and Zachar, Z. 1989. Retrotransposons and the *FB* transposon from *Drosophila melanogaster*. In *Mobile DNA* (eds. D.E. Berg and M.M. Howe), pp. 485–502. American Society for Microbiology, Washington, D.C.
- Boeke, J.D. and Devine, S.E. 1998. Yeast retrotransposons: Finding a nice quiet neighborhood. *Cell* **93**: 1087–1089.
- Britten, R.J. 1996. DNA sequence insertion and evolutionary variation in gene regulation. *Proc. Natl. Acad. Sci.* **93**: 9374–9377.
- . 1997. Mobile elements inserted in the distant past have taken on important functions. *Gene* **205**: 177–182.
- Cáceres, M., Barbadilla, A., and Ruiz, A. 1997. Inversion length and breakpoint distribution in the *Drosophila buzzatii* species complex: Is inversion length a selected trait? *Evolution* **51**: 1149–1155.
- Cáceres, M., Ranz, J.M., Barbadilla, A., Long, M., and Ruiz, A. 1999. Generation of a widespread *Drosophila* inversion by a transposable element. *Science* **285**: 415–418.
- Calvi, B.R., Hong, T.J., Findley, S.D., and Gelbart, W.M. 1991. Evidence for a common evolutionary origin of inverted repeat transposons in *Drosophila* and plants: *hobo*, *Activator*, and *Tam3*. *Cell* **66**: 465–471.
- Capy, P., Bazin, C., Higuier, D., and Langin, T. 1998. *Dynamics and evolution of transposable elements*. Springer-Verlag, Heidelberg.
- Charlesworth, B. and Langley, C.H. 1991. Population genetics of transposable elements in *Drosophila*. In *Evolution at the molecular level* (eds. R.K. Selander, A.G. Clark, and T.S. Whittam), pp. 150–176. Sinauer, Sunderland, MA.
- Charlesworth, B., Sniegowski, P., and Stephan, W. 1994. The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature* **371**: 215–220.
- Cirera, S., Martín-Campos, J.M., Segarra, C., and Aguadé, M. 1995. Molecular characterization of the breakpoints of an inversion fixed between *Drosophila melanogaster* and *D. subobscura*. *Genetics* **139**: 321–326.
- Daveran-Mingot, M.-L., Campo, N., Ritzenthaler, P., and le Bourgeois, P. 1998. A natural large chromosomal inversion in *Lactococcus lactis* is mediated by homologous recombination between two insertion sequences. *J. Bacteriol.* **180**: 4834–4842.
- Devlin, R.H., Bingham, B., and Wakimoto, B.T. 1990. The organization and expression of the *light* gene, a heterochromatic gene of *Drosophila melanogaster*. *Genetics* **125**: 129–140.
- Dimitri, P., Arcà, B., Berghella, L., and Mei, E. 1997. High genetic instability of heterochromatin after transposition of the LINE-like *I* factor in *Drosophila melanogaster*. *Proc. Natl. Acad. Sci.* **94**: 8052–8057.

- Doolittle, W.F. and Sapienza, C. 1980. Selfish genes, the phenotype paradigm, and genome evolution. *Nature* **284**: 601–603.
- Dorer, D.R. and Henikoff, S. 1994. Expansions of transgene repeats cause heterochromatin formation and gene silencing in *Drosophila*. *Cell* **77**: 993–1002.
- Eanes, W.F., Wesley, C., and Charlesworth, B. 1992. Accumulation of *P* elements in minority inversions in natural populations of *Drosophila melanogaster*. *Genet. Res.* **59**: 1–9.
- Egner, C. and Berg, D.E. 1981. Excision of transposon Tn5 is dependent on the inverted repeats but not on the transposase function of Tn5. *Proc. Natl. Acad. Sci.* **78**: 459–463.
- Engels, W.R. 1996. *P* elements in *Drosophila*. In *Transposable elements* (eds. H. Saedler and A. Gierl), pp. 103–123. Springer-Verlag, Berlin.
- Evgen'ev, M.B., Zelentsova, H., Poluectova, H., Lyozin, G.T., Veleikodvorskaja, V., Pyatkov, K.I., Zhivotovsky, L.A., and Kidwell, M.G. 2000. Mobile elements and chromosomal evolution in the *virilis* group of *Drosophila*. *Proc. Natl. Acad. Sci.* **94**: 7704–7711.
- Farabaugh, P.J., Schmeissner, U., Hofer, M., and Miller, J.H. 1978. Genetic studies of the *lac* repressor. VII. On the molecular nature of spontaneous hotspots in the *lacI* gene of *Escherichia coli*. *J. Mol. Biol.* **126**: 847–863.
- Finnegan, D.J. 1989. Eukaryotic transposable elements and genome evolution. *Trends Genet.* **5**: 103–107.
- Fu, Y.-X. and Li, W.-H. 1993. Statistical tests of neutrality of mutations. *Genetics* **133**: 693–709.
- Gatti, M. and Pimpinelli, S. 1992. Functional elements in *Drosophila melanogaster* heterochromatin. *Amu. Rev. Genet.* **26**: 239–275.
- Gray, Y.H.M. 2000. It takes two transposons to tango: Transposable-element-mediated chromosomal rearrangements. *Trends Genet.* **16**: 461–468.
- Hagemann, S., Miller, W.J., Haring, E., and Pinsker, W. 1998. Nested insertions of short mobile sequences in *Drosophila P* elements. *Chromosoma* **107**: 6–16.
- Handler, A.M. and Gomez, S.P. 1997. A new *hobo*, *Ac*, *Tam3* transposable element, *hopper*, from *Bactrocera dorsalis* is distantly related to *hobo* and *Ac*. *Gene* **185**: 133–135.
- Hankeln, T. and Schmidt, E.R. 1990. New *Foldback* transposable element *TFB1* found in histone genes of the midge *Chironomus thummi*. *J. Mol. Biol.* **215**: 477–482.
- Hartl, D.L. 2000. Molecular melodies in high and low C. *Nat. Rev. Genet.* **1**: 145–149.
- Hasson, E. and Eanes, W.F. 1996. Contrasting histories of three gene regions associated with *In(3L)Payne* of *Drosophila melanogaster*. *Genetics* **144**: 1565–1575.
- Hoffman-Liebermann, B., Liebermann, D., and Cohen, S.N. 1989. *TU* elements and *Puppy* sequences. In *Mobile DNA* (eds. D.E. Berg and M.M. Howe), pp. 575–592. American Society for Microbiology, Washington, D.C.
- Holmes, S.E., Dombroski, B.A., Krebs, C.M., Boehm, C.D., and Kazazian, H.H. 1994. A new retrotransposable human *L1* element from the *LRE2* locus on chromosome 1q produces a chimaeric insertion. *Nature Genet.* **7**: 143–148.
- Ji, H., Moore, D.P., Blomberg, M.A., Braitterman, L.T., Voytas, D.F., Natsoulis, G., and Boeke, J.D. 1993. Hotspots for unselected *Ty1* transposition events on yeast chromosome III are near tRNA genes and LTR sequences. *Cell* **73**: 1007–1018.
- Kidwell, M.G. and Lisch, D. 1997. Transposable elements as sources of variation in animals and plants. *Proc. Natl. Acad. Sci.* **94**: 7704–7711.
- Kim, J.M., Vanguri, S., Boeke, J.D., Gabriel, A., and Voytas, D.F. 1998. Transposable elements and genome organization: A comprehensive survey of retrotransposons revealed by the complete *Saccharomyces cerevisiae* genome sequence. *Genome Res.* **8**: 464–478.
- Kreitman, M. and Hudson, R.R. 1991. Inferring the evolutionary histories of the *Adh* and *Adh-dup* loci in *Drosophila melanogaster* from patterns of polymorphism and divergence. *Genetics* **127**: 565–582.
- Krimbas, C.B. and Powell, J.R. 1992. Introduction. In *Drosophila inversion polymorphism* (eds. C.B. Krimbas and J.R. Powell), pp. 1–52. CRC Press, Boca Raton, FL.
- Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921.
- Liao, G.-C., Rehm, E.J., and Rubin, G.M. 2000. Insertion site preferences of the *P* transposable element in *Drosophila melanogaster*. *Proc. Natl. Acad. Sci.* **97**: 3347–3351.
- Lim, J.K. and Simmons, M.J. 1994. Gross chromosome rearrangements mediated by transposable elements in *Drosophila melanogaster*. *BioEssays* **16**: 269–275.
- Lin, X., Kaul, S., Rounsley, S., Shea, T.P., Benito, M.-I., Town, C.D., Fujii, C.Y., Mason, T., Bowman, C.L., Barnstead, M., et al. 1999. Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. *Nature* **402**: 761–777.
- Lindsley, D.L. and Zimm, G.G. 1992. *The genome of Drosophila melanogaster*. Academic Press, San Diego.
- Locke, J., Podemski, L., Roy, K., Pilgrim, D., and Hodgetts, R. 1999. Analysis of two cosmid clones from chromosome 4 of *Drosophila melanogaster* reveals two new genes amid an unusual arrangement of repeated sequences. *Genome Res.* **9**: 137–149.
- Lohe, A.R., Timmons, C., Beerman, I., Lozovskaya, E.R., and Hartl, D.L. 2000. Self-inflicted wounds, template-directed gap repair and a recombination hotspot: Effects of the *mariner* transposase. *Genetics* **154**: 647–656.
- Lovering, R., Harden, N., and Ashburner, M. 1991. The molecular structure of *TE146* and its derivatives in *Drosophila melanogaster*. *Genetics* **128**: 357–372.
- Lyttle, T.W. and Haymer, D.S. 1992. The role of the transposable element *hobo* in the origin of endemic inversions in wild populations of *Drosophila melanogaster*. *Genetica* **86**: 113–126.
- Marin, I. and Fontdevila, A. 1995. Characterization of *Gandalf*, a new inverted-repeat transposable element of *Drosophila koepferae*. *Mol. Gen. Genet.* **248**: 423–433.
- Mathiopoulou, K.D., della Torre, A., Predazzi, V., Petrarca, V., and Coluzzi, M. 1998. Cloning of inversion breakpoints in the *Anopheles gambiae* complex traces a transposable element at the inversion junction. *Proc. Natl. Acad. Sci.* **95**: 12444–12449.
- Mayer, K., Schüller, C., Wambutt, R., Murphy, G., Volckaert, G., Pohl, T., Düsterhöft, A., Stiekema, W., Entian, K.-D., Terryn, N., et al. 1999. Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*. *Nature* **402**: 769–777.
- McCaron, M., Duttaroy, A., Dougherty, G., and Chovnick, A. 1994. *Drosophila P* element transposase induces male recombination additively and without a requirement for *P* element excision or insertion. *Genetics* **136**: 1013–1023.
- McDonald, J.F. 1995. Transposable elements: Possible catalysts of organismic evolution. *Trends Ecol. Evol.* **10**: 123–126.
- Miklos, G.L.G., Yamamoto, M.-T., Davies, J., and Pirrotta, V. 1988. Microcloning reveals a high frequency of repetitive sequences characteristic of chromosome 4 and the β -heterochromatin of *Drosophila melanogaster*. *Proc. Natl. Acad. Sci.* **85**: 2051–2055.
- Montgomery, E.A., Charlesworth, B., and Langley, C.H. 1987. A test for the role of natural selection in the stabilization of transposable element copy number in a population of *Drosophila melanogaster*. *Genet. Res.* **49**: 31–41.
- Nei, M. 1987. *Molecular evolutionary genetics*. Columbia University Press, New York.
- Orgel, L.E. and Crick, F.H.C. 1980. The ultimate parasite. *Nature* **284**: 604–607.
- Pimpinelli, S., Berloco, M., Fanti, L., Dimitri, P., Bonaccorsi, S., Marchetti, E., Caizzi, R., Caggese, C., and Gatti, M. 1995. Transposable elements are stable structural components of *Drosophila melanogaster* heterochromatin. *Proc. Natl. Acad. Sci.* **92**: 3804–3808.
- Preston, C.R., Sved, J.A., and Engels, W.R. 1996. Flanking duplications and deletions associated with *P*-induced male recombination in *Drosophila*. *Genetics* **144**: 1623–1638.
- Ranz, J.M., Cáceres, M., and Ruiz, A. 1999. Comparative mapping of cosmids and gene clones from a 1.6 Mb chromosomal region of *Drosophila melanogaster* in three species of the distantly related subgenus *Drosophila*. *Chromosoma* **108**: 32–45.
- Rebatchouk, D. and Narita, J.O. 1997. *Foldback* transposable elements in plants. *Plant. Mol. Biol.* **34**: 831–835.
- Regner, L.P., Pereira, M.S.O., Alonso, C.E.V., Abdelhay, E., and Valente, V.L.S. 1996. Genomic distribution of *P* elements in *Drosophila willistoni* and a search for their relationship with chromosomal inversions. *J. Hered.* **87**: 191–198.
- Rodríguez-Trelles, F., Alarcón, L., and Fontdevila, A. 2000. Molecular evolution and phylogeny of the *buzzatii* complex (*Drosophila repleta* group): A maximum-likelihood approach. *Mol. Biol. Evol.* **17**: 1112–1122.
- Roza, J. and Roza, R. 1999. DnaSP version 3: An integrated program for molecular population genetics and molecular evolution analysis. *Bioinformatics* **15**: 174–175.
- Roza, J., Segarra, C., Ribó, G., and Aguadé, M. 1999. Molecular population genetics of the *rp49* gene region in different

- chromosomal inversions of *Drosophila subobscura*. *Genetics* **151**: 189–202.
- Ruiz, A. and Wasserman, M. 1993. Evolutionary cytogenetics of the *Drosophila buzzatii* species complex. *Heredity* **70**: 582–596.
- Russo, C.A.M., Takezaki, N., and Nei, M. 1995. Molecular phylogeny and divergence times of drosophilid species. *Mol. Biol. Evol.* **12**: 391–404.
- Saitou, N. and Nei, M. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**: 406–425.
- SanMiguel, P., Tikhonov, A., Jin, Y.-K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P.S., Edwards, K.J., Lee, M., Avramova, Z., and Bennetzen, J.L. 1996. Nested retrotransposons in the intergenic regions of the maize genome. *Science* **274**: 765–768.
- SanMiguel, P., Gaut, B.S., Tikhonov, A., Nakajima, Y., and Bennetzen, J.L. 1998. The paleontology of intergene retrotransposons of maize. *Nat. Genet.* **20**: 43–45.
- Schwartz, A., Chan, D.C., Brown, L.G., Alagappan, R., Pettay, D., Distech, C., McGillivray, B., de la Chapelle, A., and Page, D.C. 1998. Reconstructing hominid Y evolution: X-homologous block, created by X-Y transposition, was disrupted by Yp inversion through LINE-LINE recombination. *Hum. Mol. Genet.* **7**: 1–11.
- Sniegowski, P.D. and Charlesworth, B. 1994. Transposable element numbers in cosmopolitan inversions from a natural population of *Drosophila melanogaster*. *Genetics* **137**: 815–827.
- Staveley, B.E., Heslip, T.R., Hodgetts, R.B., and Bell, J.B. 1995. Protected P-element termini suggests a role for inverted-repeat-binding protein in transposase-induced gap repair in *Drosophila melanogaster*. *Genetics* **139**: 1321–1329.
- Stephan, W. and Langley, C.H. 1989. Molecular genetic variation in the centromeric region of the X chromosome in three *Drosophila ananassae* populations. I. Contrasts between the *vermillion* and *forked* loci. *Genetics* **121**: 89–99.
- Sved, J.A., Eggleston, W.B., and Engels, W.R. 1990. Germ-line and somatic recombination induced by *in vitro* modified P elements in *Drosophila melanogaster*. *Genetics* **124**: 331–337.
- Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**: 585–595.
- Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**: 4673–4680.
- Vaury, C., Bucheton, A., and Pelisson, A. 1989. The β -heterochromatic sequences flanking the *I* elements are themselves defective transposable elements. *Chromosoma* **98**: 215–224.
- Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. 2001. The sequence of the human genome. *Science* **291**: 1304–1351.
- Wallace, M.R., Andersen, L.B., Saulino, A.M., Gregory, P.E., Glover, T.W., and Collins, F.S. 1991. A de novo *Alu* insertion results in neurofibromatosis type 1. *Nature* **353**: 864–866.
- Waugh O'Neill, R.J., O'Neill, M.J., and Marshall Graves, J.A. 1998. Undermethylation associated with retroelement activation and chromosome remodelling in an interspecific mammalian hybrid. *Nature* **393**: 68–72.
- Wesley, C.S. and Eanes, W.F. 1994. Isolation and analysis of the breakpoint sequences of chromosome inversion *In(3L)Payne* in *Drosophila melanogaster*. *Proc. Natl. Acad. Sci.* **91**: 3132–3136.
- Yuan, J., Finney, M., Tsung, N., and Horvitz, H.R. 1991. *Tc4*, a *Caenorhabditis elegans* transposable element with an unusual fold-back structure. *Proc. Natl. Acad. Sci.* **88**: 3334–3338.
- Zou, S. and Voytas, D.F. 1997. Silent chromatin determines target preference of the *Saccharomyces* retrotransposon *Ty5*. *Proc. Natl. Acad. Sci.* **94**: 7412–7416.

Received December 11, 2000; accepted in revised form May 14, 2001.

APPENDIX

II

Evolution of genes and genomes on the *Drosophila* phylogeny

Drosophila 12 Genomes Consortium*

Comparative analysis of multiple genomes in a phylogenetic framework dramatically improves the precision and sensitivity of evolutionary inference, producing more robust results than single-genome analyses can provide. The genomes of 12 *Drosophila* species, ten of which are presented here for the first time (*sechellia*, *simulans*, *yakuba*, *erecta*, *ananassae*, *persimilis*, *willistoni*, *mojavensis*, *virilis* and *grimshawi*), illustrate how rates and patterns of sequence divergence across taxa can illuminate evolutionary processes on a genomic scale. These genome sequences augment the formidable genetic tools that have made *Drosophila melanogaster* a pre-eminent model for animal genetics, and will further catalyse fundamental research on mechanisms of development, cell biology, genetics, disease, neurobiology, behaviour, physiology and evolution. Despite remarkable similarities among these *Drosophila* species, we identified many putatively non-neutral changes in protein-coding genes, non-coding RNA genes, and cis-regulatory regions. These may prove to underlie differences in the ecology and behaviour of these diverse species.

As one might expect from a genus with species living in deserts, in the tropics, on chains of volcanic islands and, often, commensally with humans, *Drosophila* species vary considerably in their morphology, ecology and behaviour¹. Species in this genus span a wide range of global distributions: the 12 sequenced species originate from Africa, Asia, the Americas and the Pacific Islands, and also include cosmopolitan species that have colonized the planet (*D. melanogaster* and *D. simulans*) as well as closely related species that live on single islands (*D. sechellia*)². A variety of behavioural strategies is also encompassed by the sequenced species, ranging in feeding habit from generalist, such as *D. ananassae*, to specialist, such as *D. sechellia*, which feeds on the fruit of a single plant species.

Despite this wealth of phenotypic diversity, *Drosophila* species share a distinctive body plan and life cycle. Although only *D. melanogaster* has been extensively characterized, it seems that the most important aspects of the cellular, molecular and developmental biology of these species are well conserved. Thus, in addition to providing an extensive resource for the study of the relationship between sequence and phenotypic diversity, the genomes of these species provide an excellent model for studying how conserved functions are maintained in the face of sequence divergence. These genome sequences provide an unprecedented dataset to contrast genome structure, genome content, and evolutionary dynamics across the well-defined phylogeny of the sequenced species (Fig. 1).

Genome assembly, annotation and alignment

Genome sequencing and assembly. We used the previously published sequence and updated assemblies for two *Drosophila* species, *D. melanogaster*^{3,4} (release 4) and *D. pseudoobscura*⁵ (release 2), and generated DNA sequence data for 10 additional *Drosophila* genomes by whole-genome shotgun sequencing^{6,7}. These species were chosen to span a wide variety of evolutionary distances, from closely related pairs such as *D. sechellia*/*D. simulans* and *D. persimilis*/*D. pseudoobscura* to the distantly related species of the *Drosophila* and *Sophophora* subgenera. Whereas the time to the most recent common ancestor of the sequenced species may seem small on an evolutionary timescale, the evolutionary divergence spanned by the genus *Drosophila* exceeds

that of the entire mammalian radiation when generation time is taken into account, as discussed further in ref. 8. We sequenced seven of the new species (*D. yakuba*, *D. erecta*, *D. ananassae*, *D. willistoni*, *D. virilis*, *D. mojavensis* and *D. grimshawi*) to deep coverage (8.4× to 11.0×) to produce high quality draft sequences. We sequenced two species, *D. sechellia* and *D. persimilis*, to intermediate coverage (4.9× and 4.1×, respectively) under the assumption that the availability of a sister species sequenced to high coverage would obviate the need for deep sequencing without sacrificing draft genome quality. Finally, seven inbred strains of *D. simulans* were sequenced to low coverage (2.9× coverage from *w*⁵⁰¹ and ~1× coverage of six other strains) to provide population variation data⁹. Further details of the sequencing strategy can be found in Table 1, Supplementary Table 1 and section 1 in Supplementary Information.

We generated an initial draft assembly for each species using one of three different whole-genome shotgun assembly programs (Table 1). For *D. ananassae*, *D. erecta*, *D. grimshawi*, *D. mojavensis*, *D. virilis* and *D. willistoni*, we also generated secondary assemblies; reconciliation of these with the primary assemblies resulted in a 7–30% decrease in the estimated number of misassembled regions and a 12–23% increase in the N50 contig size¹⁰ (Supplementary Table 2). For *D. yakuba*, we generated 52,000 targeted reads across low-quality regions and gaps to improve the assembly. This doubled the mean contig and scaffold sizes and increased the total fraction of high quality bases (quality score (Q) > 40) from 96.5% to 98.5%. We improved the initial 2.9× *D. simulans* *w*⁵⁰¹ whole-genome shotgun assembly by filling assembly gaps with contigs and unplaced reads from the ~1× assemblies of the six other *D. simulans* strains, generating a 'mosaic' assembly (Supplementary Table 3). This integration markedly improved the *D. simulans* assembly: the N50 contig size of the mosaic assembly, for instance, is more than twice that of the initial *w*⁵⁰¹ assembly (17 kb versus 7 kb).

Finally, one advantage of sequencing genomes of multiple closely related species is that these evolutionary relationships can be exploited to dramatically improve assemblies. *D. yakuba* and *D. simulans* contigs and scaffolds were ordered and oriented using pairwise alignment to the well-validated *D. melanogaster* genome

*A list of participants and affiliations appears at the end of the paper.

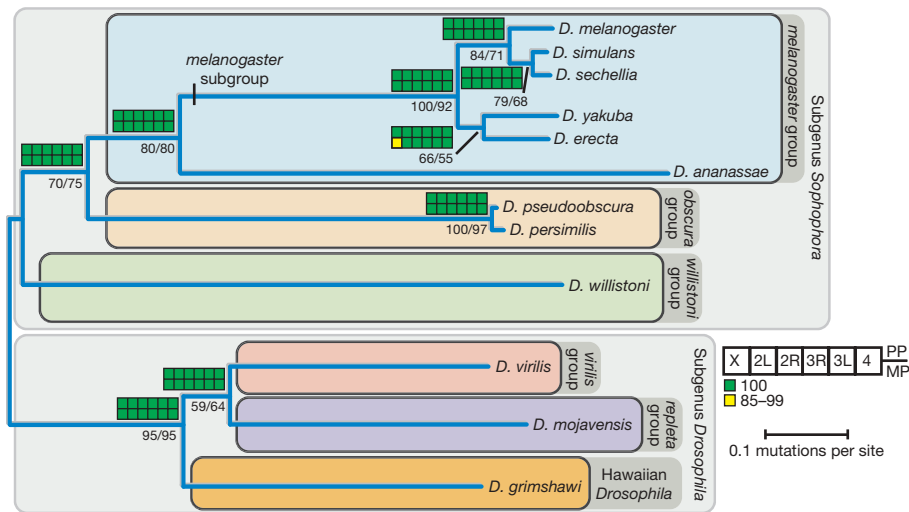


Figure 1 | Phylogram of the 12 sequenced species of *Drosophila*. Phylogram derived using pairwise genomic mutation distances and the neighbour-joining method^{152,153}. Numbers below nodes indicate the per cent of genes supporting a given relationship, based on evolutionary distances estimated from fourfold-degenerate sites (left of solidus) and second codon positions (right of solidus). Coloured blocks indicate support from bayesian

(posterior probability (PP), upper blocks) and maximum parsimony (MP; bootstrap values, lower blocks) analyses of data partitioned by chromosome arm. Branch lengths indicate the number of mutations per site (at fourfold-degenerate sites) using the ordinary least squares method. See ref. 154 for a discussion of the uncertainties in the *D. yakuba/D. erecta* clade.

sequence (Supplementary Information section 2). Likewise, the 4–5× *D. persimilis* and *D. sechellia* assemblies were improved by assisted assembly using the sister species (*D. pseudoobscura* and *D. simulans*, respectively) to validate both alignments between reads and linkage information. For the remaining species, comparative syntenic information, and in some cases linkage information, were also used to pinpoint locations of probable genome mis-assembly, to assign assembly scaffolds to chromosome arms and to infer their order and orientation along euchromatic chromosome arms, supplementing experimental analysis based on known markers (A. Bhutkar, S. Russo, S. Schaeffer, T. F. Smith and W. M. Gelbart, personal communication) (Supplementary Information section 2).

The mitochondrial (mt)DNA of *D. melanogaster*, *D. sechellia*, *D. simulans* (siII), *D. mauritiana* (maII) and *D. yakuba* have been previously sequenced^{11,12}. For the remaining species (except *D. pseudoobscura*, the DNA from which was prepared from embryonic nuclei), we were able to assemble full mitochondrial genomes, excluding the A+T-rich control region (Supplementary Information section 2)¹³. In addition, the genome sequences of three *Wolbachia* endosymbionts (*Wolbachia wSim*, *Wolbachia wAna* and *Wolbachia wWil*) were assembled from trace archives, in *D. simulans*, *D. ananassae* and *D. willistoni*, respectively¹⁴. All of the genome sequences described here are available in FlyBase (www.flybase.org) and GenBank (www.ncbi.nlm.nih.gov) (Supplementary Tables 4 and 5).

Repeat and transposable element annotation. Repetitive DNA sequences such as transposable elements pose challenges for

whole-genome shotgun assembly and annotation. Because the best approach to transposable element discovery and identification is still an active and unresolved research question, we used several repeat libraries and computational strategies to estimate the transposable element/repeat content of the 12 *Drosophila* genome assemblies (Supplementary Information section 3). Previously curated transposable element libraries in *D. melanogaster* provided the starting point for our analysis; to limit the effects of ascertainment bias, we also developed *de novo* repeat libraries using PILER-DF^{15,16} and ReAS¹⁷. We used four transposable element/repeat detection methods (RepeatMasker, BLASTER-TX, RepeatRunner and CompTE) in conjunction with these transposable element libraries to identify repetitive elements in non-*melanogaster* species. We assessed the accuracy of each method by calibration with the estimated 5.5% transposable element content in the *D. melanogaster* genome, which is based on a high-resolution transposable element annotation¹⁸ (Supplementary Fig. 1). On the basis of our results, we suggest a hybrid strategy for new genome sequences, employing translated BLAST with general transposable element libraries and RepeatMasker with species-specific ReAS libraries to estimate the upper and lower bound on transposable element content.

Protein-coding gene annotation. We annotated protein-coding sequences in the 11 non-*melanogaster* genomes, using four different *de novo* gene predictors (GeneID¹⁹, SNAP²⁰, N-SCAN²¹ and CONTRAST²²); three homology-based predictors that transfer annotations from *D. melanogaster* (GeneWise²³, Exonerate²⁴, GeneMapper²⁵); and one predictor that combined *de novo* and homology-based evidence (Gnomon²⁶). These gene prediction sets

Table 1 | A summary of sequencing and assembly properties of each new genome

Final assembly	Genome centre	Q20 coverage (×)	Assembly size (Mb)	No. of contigs ≥2 kb	N50 contig ≥2 kb (kb)	Per cent of base pairs with quality >Q40
<i>D. simulans</i>	WUGSC*	2.9	137.8	10,843	17	90.3
<i>D. sechellia</i>	Broad†	4.9	166.6	9,713	43	90.6
<i>D. yakuba</i>	WUGSC*	9.1	165.7	6,344	125	98.5
<i>D. erecta</i>	Agencourt‡	10.6	152.7	3,283	458	99.2
<i>D. ananassae</i>	Agencourt‡	8.9	231.0	8,155	113	98.5
<i>D. persimilis</i>	Broad†	4.1	188.4	14,547	20	93.3
<i>D. willistoni</i>	JCVI‡	8.4	235.5	6,652	197	97.4
<i>D. virilis</i>	Agencourt‡	8.0	206.0	5,327	136	98.7
<i>D. mojavensis</i>	Agencourt‡	8.2	193.8	5,734	132	98.6
<i>D. grimshawi</i>	Agencourt‡	7.9	200.5	9,632	114	97.1

Contigs, contiguous sequences not interrupted by gaps; N50, the largest length *L* such that 50% of all nucleotides are contained in contigs of size ≥*L*. The Q20 coverage of contigs is based on the number of assembled reads, average Q20 readlength and the assembled size excluding gaps. Assemblers used: *PCAP6, †ARACHNE4.5 and ‡Celera Assembler 7.

Table 2 | A summary of annotated features across all 12 genomes

	Protein-coding gene annotations			Non-coding RNA annotations				Repeat coverage (%) [*]	Genome size (Mb; assembly [†] /flow cytometry [‡])
	Total no. of protein-coding genes (per cent with <i>D. melanogaster</i> homologue)	Coding sequence/intron (Mb)	tRNA (pseudo)	snoRNA	miRNA	rRNA (5.8S + 5S)	snRNA		
<i>D. melanogaster</i>	13,733 (100%)	38.9/21.8	297 (4)	250	78	101	28	5.35	118/200
<i>D. simulans</i>	15,983 (80.0%)	45.8/19.6	268 (2)	246	70	72	32	2.73	111/162
<i>D. sechellia</i>	16,884 (81.2%)	47.9/21.9	312 (13)	242	78	133	30	3.67	115/171
<i>D. yakuba</i>	16,423 (82.5%)	50.8/22.9	380 (52)	255	80	55	37	12.04	127/190
<i>D. erecta</i>	15,324 (86.4%)	49.1/22.0	286 (2)	252	81	101	38	6.97	134/135
<i>D. ananassae</i>	15,276 (83.0%)	57.3/22.3	472 (165)	194	76	134	29	24.93	176/217
<i>D. pseudoobscura</i>	16,363 (78.2%)	49.7/24.0	295 (1)	203	73	55	31	2.76	127/193
<i>D. persimilis</i>	17,325 (72.6%)	54.0/21.9	306 (1)	199	75	80	31	8.47	138/193
<i>D. willistoni</i>	15,816 (78.8%)	65.4/23.5	484 (164)	216	77	76	37	15.57	187/222
<i>D. virilis</i>	14,680 (82.7%)	57.9/21.7	279 (2)	165	74	294	31	13.96	172/364
<i>D. mojavensis</i>	14,849 (80.8%)	57.8/21.9	267 (3)	139	71	74	30	8.92	161/130
<i>D. grimshawi</i>	15,270 (81.3%)	54.9/22.5	261 (1)	154	82	70	32	2.84	138/231

^{*} Repeat coverage calculated as the fraction of scaffolds >200 kb covered by repeats, estimated as the midpoint between BLASTER-tx + PILER and RepeatMasker + ReAS (Supplementary Information section 3). [†]Total genome size estimated as the sum of base pairs in genomic scaffold >200,000 bp. [‡]Genome size estimates based on flow cytometry³⁸.

were combined using GLEAN, a gene model combiner that chooses the most probable combination of start, stop, donor and acceptor sites from the input predictions^{27,28}. All analyses reported here, unless otherwise noted, relied on a reconciled consensus set of predicted gene models—the GLEAN-R set (Table 2, and Supplementary Information section 4.1).

Quality of gene models. As the first step in assessing the quality of the GLEAN-R gene models, we used expression data from microarray experiments on adult flies, with arrays custom-designed for *D. simulans*, *D. yakuba*, *D. ananassae*, *D. pseudoobscura*, *D. virilis* and *D. mojavensis*²⁹ (GEO series GSE6640; Supplementary Information section 4.2). We detected expression significantly above negative controls (false-discovery-rate-corrected Mann–Whitney U (MWU) $P < 0.001$) for 77–93% of assayed GLEAN-R models, representing 50–68% of the total GLEAN-R predictions in each species (Supplementary Table 6). Evolutionarily conserved gene models are much more likely to be expressed than lineage-specific ones (Fig. 2). Although these data cannot confirm the detailed structure of gene models, they do suggest that the majority of GLEAN-R models contain sequence that is part of a poly-adenylated transcript. Approximately 20% of transcription in *D. melanogaster* seems to be unassociated with protein-coding genes³⁰, and our microarray experiments fail to detect conditionally expressed genes. Thus,

transcript abundance cannot conclusively establish the presence or absence of a protein-coding gene. Nonetheless, we believe these expression data increase our confidence in the reliability of the GLEAN-R models, particularly those supported by homology evidence (Fig. 2).

Because the GLEAN-R gene models were built using assemblies that were not repeat masked, it is likely that some proportion of gene models are false positives corresponding to coding sequences of transposable elements. We used RepeatMasker with *de novo* ReAS libraries and PFAM structural annotations of the GLEAN-R gene set to flag potentially transposable element-contaminated gene models (Supplementary Information section 4.2). These procedures suggest that 5.6–32.3% of gene models in non-*melanogaster* species correspond to protein-coding content derived from transposable elements (Supplementary Table 7); these transposable element-contaminated gene models are almost exclusively confined to gene predictions without strong homology support (Fig. 2). Transposable element-contaminated gene models are excluded from the final gene prediction set used for subsequent analysis, unless otherwise noted.

Homology assignment. Two independent approaches were used to assign orthology and paralogy relationships among euchromatic *D. melanogaster* gene models and GLEAN-R predictions. The first approach was a fuzzy reciprocal BLAST (FRB) algorithm, which is an

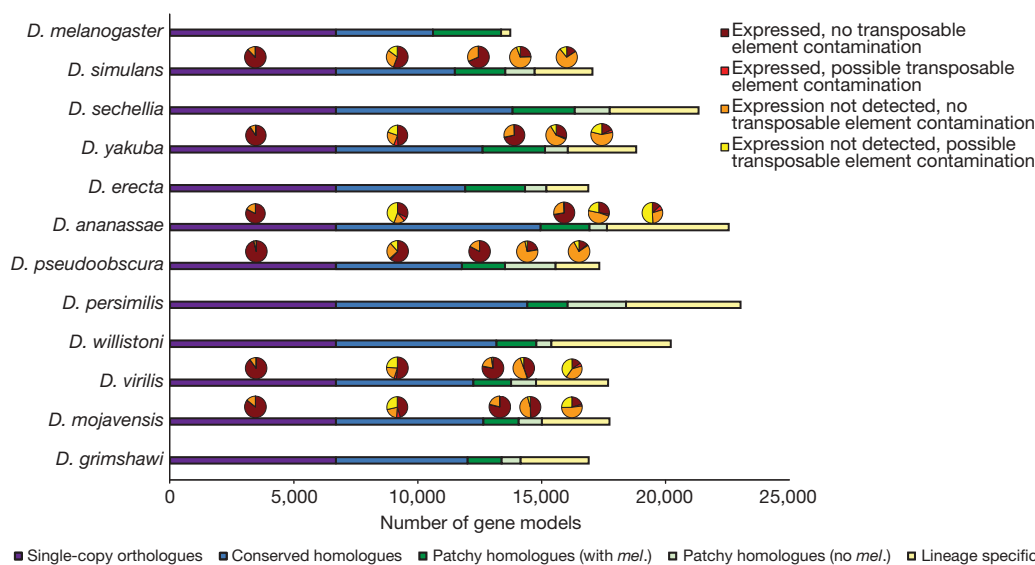


Figure 2 | Gene models in 12 *Drosophila* genomes. Number of gene models that fall into one of five homology classes: single-copy orthologues in all species (single-copy orthologues), conserved in all species as orthologues or paralogues (conserved homologues), a *D. melanogaster* homologue, but not found in all species (patchy homologues with *mel.*), conserved in at least two

species but without a *D. melanogaster* homologue (patchy homologues, no *mel.*), and found only in a single lineage (lineage specific). For those species with expression data²⁹, pie charts indicate the fraction of genes in each homology class that fall into one of four evidence classes (see text for details).

extension of the reciprocal BLAST method³¹ applicable to multiple species simultaneously (Supplementary Information section 5.1). Because the FRB algorithm does not integrate syntenic information, we also used a second approach based on Synpipe (Supplementary Information section 5.2), a tool for synteny-aided orthology assignment³². To generate a reconciled set of homology calls, pairwise Synpipe calls (between each species and *D. melanogaster*) were mapped to GLEAN-R models, filtered to retain only 1:1 relationships, and added to the FRB calls when they did not conflict and were non-redundant. This reconciled FRB + Synpipe set of homology calls forms the basis of our subsequent analyses. There were 8,563 genes with single-copy orthologues in the *melanogaster* group and 6,698 genes with single-copy orthologues in all 12 species; similar numbers of genes were also obtained with an independent approach³³. Most single-copy orthologues are expressed and are free from potential transposable element contamination, suggesting that the reconciled orthologue set contains robust and high-quality gene models (Fig. 2).

Validation of homology calls. Because both the FRB algorithm and Synpipe rely on BLAST-based methods to infer similarities, rapidly evolving genes may be overlooked. Moreover, assembly gaps and poor-quality sequence may lead to erroneous inferences of gene loss. To validate putative gene absences, we used a synteny-based GeneWise pipeline to find potentially missed homologues of *D. melanogaster* proteins (Supplementary Information section 5.4). Of the 21,928 cases in which a *D. melanogaster* gene was absent from another species in the initial homology call set, we identified plausible homologues for 13,265 (60.5%), confirmed 4,546 (20.7%) as genuine absences, and were unable to resolve 4,117 (18.8%). Because this approach is conservative and only confirms strongly supported absences, we are probably underestimating the number of genuine absences.

Coding gene alignment and filtering. Investigating the molecular evolution of orthologous and paralogous genes requires accurate multi-species alignments. Initial amino acid alignments were generated using TCOFFEE³⁴ and converted to nucleotide alignments (Supplementary Table 8). To reduce biases in downstream analyses, a simple computational screen was developed to identify and mask problematic regions of each alignment (Supplementary Information section 6). Overall, 2.8% of bases were masked in the *melanogaster* group alignments, and 3.0% of bases were masked in the full 12 species alignments, representing 8.5% and 13.8% of alignment columns, respectively. The vast majority of masked bases are masked in no more than one species (Supplementary Fig. 3), suggesting that the masking procedure is not simply eliminating rapidly evolving regions of the genome. We find an appreciably higher frequency of masked bases in lower-quality *D. simulans* and *D. sechellia* assemblies, compared to the more divergent (from *D. melanogaster*) but higher-quality *D. erecta* and *D. yakuba* assemblies, suggesting a higher error rate in accurately predicting and aligning gene models in lower-quality assemblies (Supplementary Information section 6 and Supplementary Fig. 3). We used masked versions of the alignments, including only the longest *D. melanogaster* transcripts for all subsequent analysis unless otherwise noted.

Annotation of non-coding (nc)RNA genes. Using *de novo* and homology-based approaches we annotated over 9,000 ncRNA genes from recognized ncRNA classes (Table 2, and Supplementary Information section 7). In contrast to the large number of predictions observed for many ncRNA families in vertebrates (due in part to large numbers of ncRNA pseudogenes^{35,36}), the number of ncRNA genes per family predicted by RFAM and tRNAscan in *Drosophila* is relatively low (Table 2). This suggests that ncRNA pseudogenes are largely absent from *Drosophila* genomes, which is consistent with the low number of protein-coding pseudogenes in *Drosophila*³⁷. The relatively low numbers of some classes of ncRNA genes (for example, small nucleolar (sno)RNAs) in the *Drosophila* subgenus are likely to be an artefact of rapid rates of evolution in these types

of genes and the limitation of the homology-based methods used to annotate distantly related species.

Evolution of genome structure

Coarse-level similarities among Drosophilids. At a coarse level, genome structure is well conserved across the 12 sequenced species. Total genome size estimated by flow cytometry varies less than three-fold across the phylogeny, ranging from 130 Mb (*D. mojavensis*) to 364 Mb (*D. virilis*)³⁸ (Table 2), in contrast to the order of magnitude difference between *Drosophila* and mammals. Total protein-coding sequence ranges from 38.9 Mb in *D. melanogaster* to 65.4 Mb in *D. willistoni*. Intronic DNA content is also largely conserved, ranging from 19.6 Mb in *D. simulans* to 24.0 Mb in *D. pseudoobscura* (Table 2). This contrasts dramatically with transposable element-derived genomic DNA content, which varies considerably across genomes (Table 2) and correlates significantly with euchromatic genome size (estimated as the summed length of contigs > 200 kb) (Kendall's $\tau = 0.70$, $P = 0.0016$).

To investigate overall conservation of genome architecture at an intermediate scale, we analysed synteny relationships across species using Synpipe³² (Supplementary Information section 9.1). Synteny block size and average number of genes per block varies across the phylogeny as expected, with the number of blocks increasing and the average size of blocks decreasing with increasing evolutionary distance from *D. melanogaster* (A. Bhutkar, S. Russo, T. F. Smith and W. M. Gelbart, personal communication) (Supplementary Fig. 4). We inferred 112 syntenic blocks between *D. melanogaster* and *D. sechellia* (with an average of 122 genes per block), compared to 1,406 syntenic blocks between *D. melanogaster* and *D. grimshawi* (with an average of 8 genes per block). On average, 66% of each genome assembly was covered by syntenic blocks, ranging from 68% in *D. sechellia* to 58% in *D. grimshawi*.

Similarity across genomes is largely recapitulated at the level of individual genes, with roughly comparable numbers of predicted protein-coding genes across the 12 species (Table 2). The majority of predicted genes in each species have homologues in *D. melanogaster* (Table 2, Supplementary Table 9). Moreover, most of the 13,733 protein-coding genes in *D. melanogaster* are conserved across the entire phylogeny: 77% have identifiable homologues in all 12 genomes, 62% can be identified as single-copy orthologues in the six genomes of the *melanogaster* group and 49% can be identified as single-copy orthologues in all 12 genomes. The number of functional non-coding RNA genes predicted in each *Drosophila* genome is also largely conserved, ranging from 584 in *D. mojavensis* to 908 in *D. ananassae* (Table 2).

There are several possible explanations for the observed interspecific variation in gene content. First, approximately 700 *D. melanogaster* gene models have been newly annotated since the FlyBase Release 4.3 annotations used in the current study, reducing the discrepancy between *D. melanogaster* and the other sequenced genomes in this study. Second, because low-coverage genomes tend to have more predicted gene models, we suspect that artefactual duplication of genomic segments due to assembly errors inflates the number of predicted genes in some species. Finally, the non-*melanogaster* species have many more predicted lineage-specific genes than *D. melanogaster*, and it is possible that some of these are artefactual. In the absence of experimental evidence, it is difficult to distinguish genuine lineage-specific genes from putative artefacts. Future experimental work will be required to fully disentangle the causes of interspecific variation in gene number.

Abundant genome rearrangements during Drosophila evolution. To study the structural relationships among genomes on a finer scale, we analysed gene-level synteny between species pairs. These synteny maps allowed us to infer the history and locations of fixed genomic rearrangements between species. Although *Drosophila* species vary in their number of chromosomes, there are six fundamental chromosome arms common to all species. For ease of denoting

chromosomal homology, these six arms are referred to as ‘Muller elements’ after Hermann J. Muller, and are denoted A–F. Although most pairs of orthologous genes are found on the same Muller element, there is extensive gene shuffling within Muller elements between even moderately diverged genomes (Fig. 3, and Supplementary Information section 9.1).

Previous analysis has revealed heterogeneity in rearrangement rates among close relatives: careful inspection of 29 inversions that differentiate the chromosomes of *D. melanogaster* and *D. yakuba* revealed that 28 were fixed in the lineage leading to *D. yakuba*, and only one was fixed on the lineage leading to *D. melanogaster*³⁹. Rearrangement rates are also heterogeneous across the genome among the 12 species: simulations reject a random-breakage model, which assumes that all sites are free to break in inversion events, but fail to reject a model of coldspots and hotspots for breakpoints (S. Schaeffer, personal communication). Furthermore, inversions seem to have played important roles in the process of speciation in at least some of these taxa⁴⁰.

One particularly striking example of the dynamic nature of genome micro-structure in *Drosophila* is the homeotic *homeobox* (*Hox*) gene cluster(s)⁴¹. *Hox* genes typically occur in genomic clusters, and this clustering is conserved across many vertebrate and invertebrate taxa, suggesting a functional role for the precise and collinear arrangement of these genes. However, several cluster splits have been previously identified in *Drosophila*^{42,43}, and the 12 *Drosophila* genome sequences provide additional evidence against the functional importance of *Hox* gene clustering in *Drosophila*. There are seven different gene arrangements found across 13 *Drosophila* species (the 12 sequenced genomes and *D. buzzatii*), with no species retaining the inferred ancestral gene order⁴⁴. It thus seems that, in *Drosophila*, *Hox* genes do not require clustering to maintain proper function, and are a powerful illustration of the dynamism of genome structure across the sequenced genomes.

Transposable element evolution. Mobile, repetitive transposable element sequences are a particularly dynamic component of eukaryotic genomes. Transposable element/repeat content (in scaffolds >200 kb) varies by over an order of magnitude across the genus, ranging from ~2.7% in *D. simulans* and *D. grimshawi* to ~25% in *D. ananassae* (Table 2, and Supplementary Fig. 1). These data support the lower euchromatic transposable element content in *D. simulans* relative to *D. melanogaster*⁴⁵, and reveal that euchromatic transposable element/repeat content is generally similar within the *melanogaster* subgroup. Within the *Drosophila* subgenus,

D. grimshawi has the lowest transposable element/repeat content, possibly relating to its ecological status as an island endemic, which may minimize the chance for horizontal transfer of transposable element families. Finally, the highest levels of transposable element/repeat content are found in *D. ananassae* and *D. willistoni*. These species also have the highest numbers of pseudo-transfer (t)RNA genes (Table 2), indicating a potential relationship between pseudo-tRNA genesis and repetitive DNA, as has been established in the mouse genome³⁶.

Different classes of transposable elements can vary in abundance owing to a variety of host factors, motivating an analysis of the intragenomic ecology of transposable elements in the 12 genomes. In *D. melanogaster*, long terminal repeat (LTR) retrotransposons have the highest abundance, followed by LINE (long interspersed nuclear element)-like retrotransposons and terminal inverted repeat (TIR) DNA-based transposons¹⁸. An unbiased, conservative approach (Supplementary Information section 3) for estimating the rank order abundance of major transposable element classes suggests that these abundance trends are conserved across the entire genus (Supplementary Fig. 5). Two exceptions are an increased abundance of TIR elements in *D. erecta* and a decreased abundance of LTR elements in *D. pseudoobscura*; the latter observation may represent an assembly artefact because the sister species *D. persimilis* shows typical LTR abundance. Given that individual instances of transposable element repeats and transposable element families themselves are not conserved across the genus, the stability of abundance trends for different classes of transposable elements is striking and suggests common mechanisms for host–transposable element co-evolution in *Drosophila*.

Although comprehensive analysis of the structural and evolutionary relationships among families of transposable elements in the 12 genomes remains a major challenge for *Drosophila* genomics, some initial insights can be gleaned from analysis of particularly well-characterized transposable element families. Previous analysis has shown variable dynamics for the most abundant transposable element family (*DINE-1*)⁴⁶ in the *D. melanogaster* genome^{18,47}: although inactive in *D. melanogaster*⁴⁸, *DINE-1* has experienced a recent transpositional burst in *D. yakuba*⁴⁹. Our analysis confirms that this element is highly abundant in all of the other sequenced genomes of *Drosophila*, but is not found outside of Diptera^{50,51}. Moreover, the inferred phylogenetic relationship of *DINE-1* paralogues from several *Drosophila* species suggests vertical transmission as the major mechanism for *DINE-1* propagation. Likewise, analysis of the *Galileo*

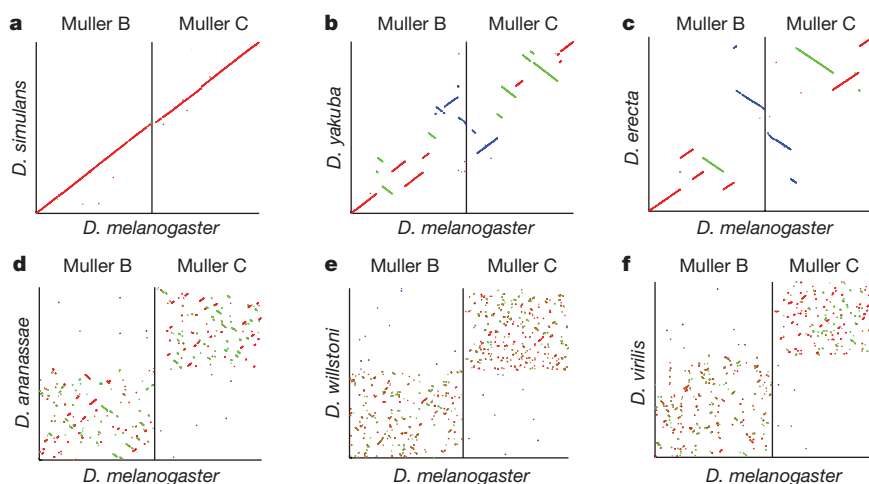


Figure 3 | Synteny plots for Muller elements B and C with respect to *D. melanogaster* gene order. The horizontal axis shows *D. melanogaster* gene order for Muller elements B and C, and the vertical axis maps homologous locations^{32,155} in individual species (a–f in increasing evolutionary distance from *D. melanogaster*). Left to right on the x axis is

from telomere to centromere for Muller element B, followed by Muller element C from centromere to telomere. Red and green lines represent syntenic segments in the same or reverse orientation along the chromosome relative to *D. melanogaster*, respectively. Blue segments show gene transposition of genes from one element to the other.

and 1360 transposons reveals a widespread but discontinuous phylogenetic distribution for both families, notably with both families absent in the geographically isolated Hawaiian species, *D. grimshawi*⁵². These results are consistent with an ancient origin of the *Galileo* and 1360 families in the genus and subsequent horizontal transfer and/or loss in some lineages.

The use of these 12 genomes also facilitated the discovery of transposable element lineages not yet documented in *Drosophila*, specifically the P instability factor (*PIF*) superfamily of DNA transposons. Our analysis indicates that there are four distinct lineages of this transposon in *Drosophila*, and that this element has indeed colonized many of the sequenced genomes⁵³. This superfamily is particularly intriguing given that *PIF*-transposase-like genes have been implicated in the origin of at least seven different genes during the *Drosophila* radiation⁵³, suggesting that not only do transposable elements affect the evolution of genome structure, but that their domestication can play a part in the emergence of novel genes.

D. melanogaster maintains its telomeres by occasional targeted transposition of three telomere-specific non-LTR retrotransposons (*HeT-A*, *TART* and *TAHRE*) to chromosome ends^{54,55} and not by the more common mechanism of telomerase-generated G-rich repeats⁵⁶. Multiple telomeric retrotransposons have originated within the genus, where they now maintain telomeres, and recurrent loss of most of the ORF2 from telomeric retrotransposons (for example, *TAHRE*) has given rise to half-telomeric-retrotransposons (for example, *HeT-A*) during *Drosophila* evolution⁵⁷. The phylogenetic relationship among these telomeric elements is congruent with the species phylogeny, suggesting that they have been vertically transmitted from a common ancestor⁵⁷.

ncRNA gene family evolution. Using ncRNA gene annotations across the 12-species phylogeny, we inferred patterns of gene copy number evolution in several ncRNA families. Transfer RNA genes are the most abundant family of ncRNA genes in all 12 genomes, with 297 tRNAs in *D. melanogaster* and 261–484 tRNA genes in the other species (Table 2). Each genome encodes a single selenocysteine tRNA, with the exception of *D. willistoni*, which seems to lack this gene (R. Guigo, personal communication). Elevated tRNA gene counts in *D. ananassae* and *D. willistoni* are explained almost entirely by pseudo-tRNA gene predictions. We infer from the lack of pseudo-tRNAs in most *Drosophila* species, and from similar numbers of tRNAs obtained from an analysis of the chicken genome ($n = 280$)⁵⁸, that the minimal metazoan tRNA set is encoded by ~300 genes, in contrast to previous estimates of 497 in human and 659 in *Caenorhabditis elegans*^{59,60}. Similar numbers of snoRNAs are predicted in the *D. melanogaster* subgroup ($n = 242$ –255), in which sequence similarity is high enough for annotation by homology, with fewer snoRNAs ($n = 194$ –216) annotated in more distant members of the *Sophophora* subgenus, and even fewer snoRNAs ($n = 139$ –165) predicted in the *Drosophila* subgenus, in which annotation by homology becomes much more difficult.

Of 78 previously reported micro (mi)RNA genes, 71 (91%) are highly conserved across the entire genus, with the remaining seven genes (*mir-2b-1*, -289, -303, -310, -311, -312 and -313) restricted to the subgenus *Sophophora* (Supplementary Information section 7.2). All the species contain similar numbers of spliceosomal snRNA genes (Table 2), including at least one copy each of the four U12-dependent (minor) spliceosomal RNAs, despite evidence for birth and death of these genes and the absence of stable subtypes⁶¹. The unusual, lineage-specific expansion in size of U11 snRNA, previously described in *Drosophila*^{61,62}, is even more extreme in *D. willistoni*. We annotated 99 copies of the 5S ribosomal (r)RNA gene in a cluster in *D. melanogaster*, and between 13 and 73 partial 5S rRNA genes in clusters in the other genomes. Finally, we identified members of several other classes of ncRNA genes, including the RNA components of the RNase P (1 per genome) and the signal recognition particle (SRP) RNA complexes (1–3 per genome), suggesting that these functional RNAs are involved in similar biological processes throughout the

genus. We were only able to locate the *roX* (RNA on X)^{63,64} genes involved in dosage compensation using nucleotide homology in the *melanogaster* subgroup, although analyses incorporating structural information have identified *roX* genes in other members of the genus⁶⁵.

We investigated the evolution of rRNA genes in the 12 sequenced genomes, using trace archives to locate sequence variants within the transcribed portions of these genes. This analysis revealed moderate levels of variation that are not distributed evenly across the rRNA genes, with fewest variants in conserved core coding regions, more variants in coding expansion regions, and higher still variant abundances in non-coding regions. The level and distribution of sequence variation in rRNA genes are suggestive of concerted evolution, in which recombination events uniformly distribute variants throughout the rDNA loci, and selection dictates the frequency to which variants can expand⁶⁶.

Protein-coding gene family evolution. For a general perspective on how the protein-coding composition of these 12 genomes has changed, we examined gene family expansions and contractions in the 11,434 gene families (including those of size one in each species) predicted to be present in the most recent common ancestor of the two subgenera. We applied a maximum likelihood model of gene gain and loss⁶⁷ to estimate rates of gene turnover. This analysis suggests that gene families expand or contract at a rate of 0.0012 gains and losses per gene per million years, or roughly one fixed gene gain/loss across the genome every 60,000 yr⁶⁸. Many gene families (4,692 or 41.0%) changed in size in at least one species, and 342 families showed significantly elevated ($P < 0.0001$) rates of gene gain and loss compared to the genomic average, indicating that non-neutral processes may play a part in gene family evolution. Twenty-two families exhibit rapid copy number evolution along the branch leading to *D. melanogaster* (eighteen contractions and four expansions; Supplementary Table 10). The most common Gene Ontology (GO) terms among families with elevated rates of gain/loss include 'defence response', 'protein binding', 'zinc ion binding', 'proteolysis', and 'trypsin activity'. Interestingly, genes involved in 'defence response' and 'proteolysis' also show high rates of protein evolution (see below). We also found heterogeneity in overall rates of gene gain and loss across lineages, although much of this variation could result from interspecific differences in assembly quality⁶⁸.

Lineage-specific genes. The vast majority of *D. melanogaster* proteins that can be unambiguously assigned a homology pattern (Supplementary Information section 5) are inferred to be ancestrally present at the genus root (11,348/11,644, or 97.5%). Of the 296 non-ancestrally present genes, 252 are either *Sophophora*-specific, or have a complicated pattern of homology requiring more than one gain and/or loss on the phylogeny, and are not discussed further. The remaining 44 proteins include 14 present in the *melanogaster* group, 23 present only in the *melanogaster* subgroup, 3 unique to the *melanogaster* species complex, and 4 found in *D. melanogaster* only. Because we restricted this analysis to unambiguous homologues of high-confidence protein-coding genes in *D. melanogaster*⁸, we are probably undercounting the number of genes that have arisen *de novo* in any particular lineage. However, ancestrally heterochromatic genes that are currently euchromatic in *D. melanogaster* may spuriously seem to be lineage-specific.

The 44 lineage-specific genes (Supplementary Table 11) differ from ancestrally present genes in several ways. They have a shorter median predicted protein length (lineage-specific median 177 amino acids, other median 421 amino acids, MWU, $P = 3.6 \times 10^{-13}$), are more likely to be intronless (Fisher's exact test (FET), $P = 6.2 \times 10^{-6}$), and are more likely to be located in the intron of another gene on the opposite strand (FET, $P = 3.5 \times 10^{-4}$). In addition, 18 of these 44 genes are testis- or accessory-gland-specific in *D. melanogaster*, a significantly greater fraction than is found in the ancestral set (FET, $P = 1.25 \times 10^{-4}$). This is consistent with previous observations that novel genes are often testis-specific in *Drosophila*^{69–73} and

expression studies on seven of the species show that species-restricted genes are more likely to exhibit male-biased expression²⁹. Further, these genes are significantly more tissue-specific in expression (as measured by τ ; ref. 74) ($MWU, P = 9.6 \times 10^{-6}$), and this pattern is not solely driven by genes with testis-specific expression patterns.

Protein-coding gene evolution

Positive selection and selective constraints in *Drosophila* genomes.

To study the molecular evolution of protein-coding genes, we estimated rates of synonymous and non-synonymous substitution in 8,510 single-copy orthologues within the six *melanogaster* group species using PAML⁷⁵ (Supplementary Information section 11.1); synonymous site saturation prevents analysis of more divergent comparisons. We investigate only single-copy orthologues because when paralogues are included, alignments become increasingly problematic. Rates of amino acid divergence for single-copy orthologues in all 12 species were also calculated; these results are largely consistent with the analysis of non-synonymous divergence in the *melanogaster* group, and are not discussed further.

To understand global patterns of divergence and constraint across functional classes of genes, we examined the distributions of ω ($=d_N/d_S$, the ratio of non-synonymous to synonymous divergence) across Gene Ontology categories (GO)⁷⁶, excluding GO

annotations based solely on electronic support (Supplementary Information section 11.2). Most functional categories of genes are strongly constrained, with median estimates of ω much less than one. In general, functionally similar genes are similarly constrained: 31.8% of GO categories have significantly lower variance in ω than expected (q -value true-positive test⁷⁷). Only 11% of GO categories had statistically significantly elevated ω (relative to the median of all genes with GO annotations) at a 5% false-discovery rate (FDR), suggesting either positive selection or a reduction in selective constraint. The GO categories with elevated ω include the biological process terms 'defence response', 'proteolysis', 'DNA metabolic process' and 'response to biotic stimulus'; the molecular function terms 'transcription factor activity', 'peptidase activity', 'receptor binding', 'odorant binding', 'DNA binding', 'receptor activity' and 'G-protein-coupled receptor activity'; and the cellular location term 'extracellular' (Fig. 4, and Supplementary Table 12). Similar results are obtained when d_N is compared across GO categories, suggesting that in most cases differences in ω among GO categories is driven by amino acid rather than synonymous site substitutions. The two exceptions are the molecular function terms 'transcription factor activity' and 'DNA binding activity', for which we observe significantly decelerated d_S (FDR = 7.2×10^{-4} for both; Supplementary Information section 11.2) and no significant differences in d_N .

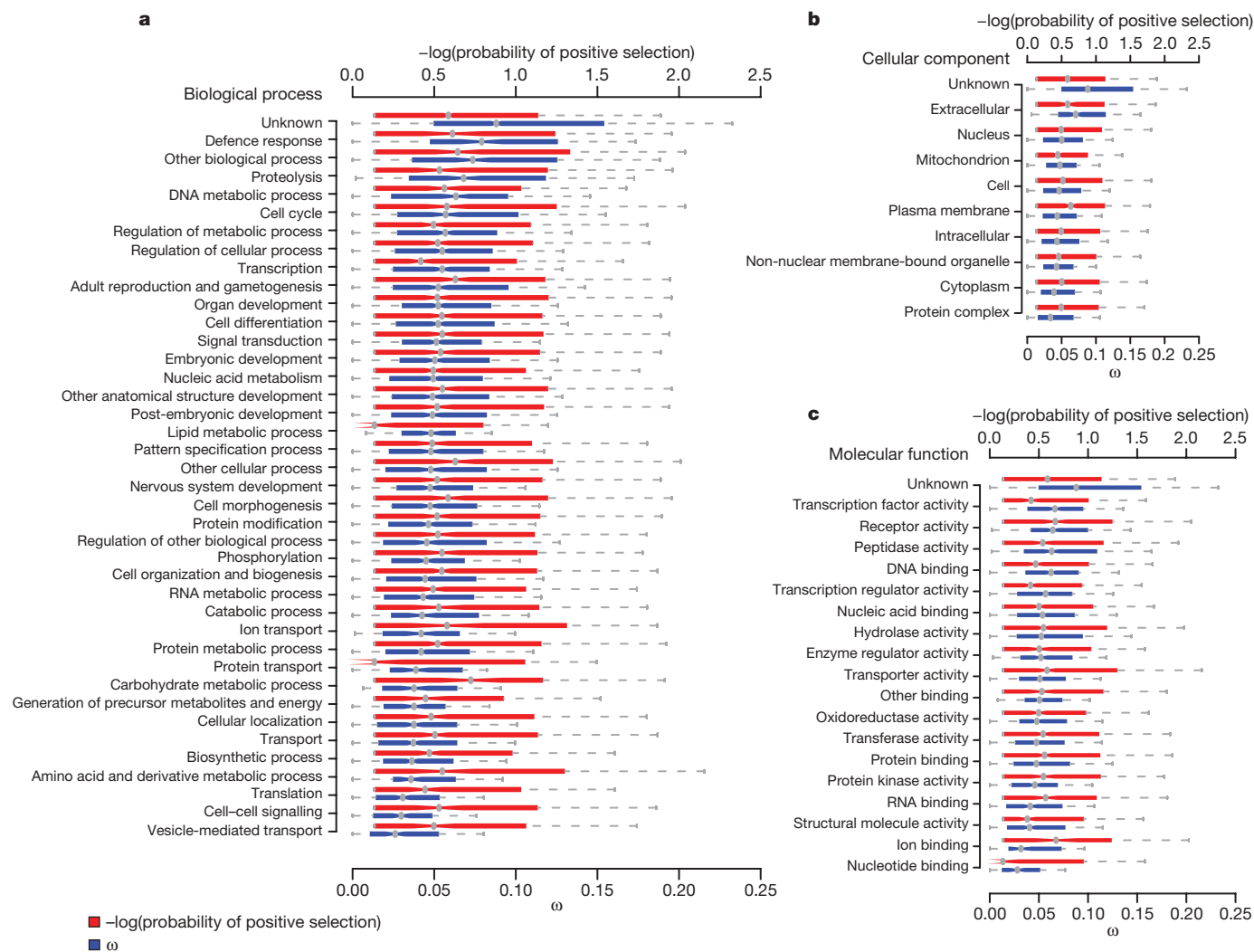


Figure 4 | Patterns of constraint and positive selection among GO terms. Distribution of average ω per gene and the negative \log_{10} of the probability of positive selection (Supplementary Information section 11.2) for genes annotated with: **a**, biological process GO terms; **b**, cellular component GO terms; and **c**, molecular function GO terms. Only GO terms with 200 or more

genes annotated are plotted. See Supplementary Table 12 for median values and significance. Note that most genes evolve under evolutionary constraint at most of their sites, leading to low values of ω ; even genes that experience positive selection do not typically have an average ω across all codons that exceeds one.

To distinguish possible positive selection from relaxed constraint, we tested explicitly for genes that have a subset of codons with signatures of positive selection, using codon-based likelihood models of molecular evolution, implemented in PAML^{78,79} (Supplementary Information section 11.1). Although this test is typically regarded as a conservative test for positive selection, it may be confounded by selection at synonymous sites. However, selection at synonymous sites (that is, codon bias, see below) is quite weak. Moreover, variability in ω presented here tends to reflect variability in d_N . We therefore believe that it is appropriate to treat synonymous sites as nearly neutral and sites with $\omega > 1$ as consistent with positive selection. Despite a number of functional categories with evidence for elevated ω , 'helicase activity' is the only functional category significantly more likely to be positively selected (permutation test, $P = 2 \times 10^{-4}$, FDR = 0.007; Supplementary Table 12); the biological significance of this finding merits further investigation. Furthermore, within each GO class, there is greater dispersion among genes in their probability of positive selection than in their estimate of ω (MWU one-tailed, $P = 0.011$; Supplementary Information section 11.1), suggesting that although functionally similar genes share patterns of constraint, they do not necessarily show similar patterns of positive selection (Fig. 4).

Interestingly, protein-coding genes with no annotated ('unknown') function in the GO database seem to be less constrained (permutation test, $P < 1 \times 10^{-4}$, FDR = 0.006)⁸⁰ and to have on average lower P -values for the test of positive selection than genes with annotated functions (permutation test, $P = 0.001$, FDR = 0.058). It is unlikely that this observation results entirely from an over-representation of mis-annotated or non-protein-coding genes in the 'unknown' functional class, because this finding is robust to the removal of all *D. melanogaster* genes predicted to be non-protein-coding in ref. 8. The bias in the way biological function is ascribed to genes (to laboratory-induced, easily scorable functions) leaves open the possibility that unannotated biological functions may have an important role in evolution. Indeed, genes with characterized mutant alleles in FlyBase evolve significantly more slowly than other genes (median $\omega_{\text{with alleles}} = 0.0525$ and $\omega_{\text{without alleles}} = 0.0701$; MWU, $P < 1 \times 10^{-16}$).

Previous work has suggested that a substantial fraction of non-synonymous substitutions in *Drosophila* were fixed through positive selection^{81–85}. We estimate that 33.1% of single-copy orthologues in the *melanogaster* group have experienced positive selection on at least a subset of codons (q -value true-positive tests⁷⁷) (Supplementary Information section 11.1). This may be an underestimate, because we have only examined single-copy orthologues, owing to difficulties in producing accurate alignments of paralogues by automated methods. On the basis of the 878 genes inferred to have experienced positive selection with high confidence (FDR < 10%), we estimated that an average of 2% of codons in positively selected genes have $\omega > 1$. Thus, several lines of evidence, based on different methodologies, suggest that patterns of amino acid fixation in *Drosophila* genomes have been shaped extensively by positive selection.

The presence of functional domains within a protein may lead to heterogeneity in patterns of constraint and adaptation along its length. Among genes inferred to be evolving by positive selection at a 10% FDR, 63.7% (q -value true-positive tests⁷⁷) show evidence for spatial clustering of positively selected codons (Supplementary Information section 11.2). Spatial heterogeneity in constraint is further supported by contrasting ω for codons inside versus outside defined InterPro domains (genes lacking InterPro domains are treated as 'outside' a defined InterPro domain). Codons within InterPro domains were significantly more conserved than codons outside InterPro domains (median ω : 0.062 InterPro domains, 0.084 outside InterPro domains; MWU, $P < 2.2 \times 10^{-16}$; Supplementary Information section 11.2). Similarly, there were significantly more positively selected codons outside of InterPro domains than inside domains (FET $P < 2.2 \times 10^{-16}$), suggesting that in addition to

being more constrained, codons in protein domains are less likely to be targets of positive selection (Supplementary Fig. 6).

Factors affecting the rate of protein evolution in *Drosophila*. The sequenced genomes of the *melanogaster* group provide unprecedented statistical power to identify factors affecting rates of protein evolution. Previous analyses have suggested that although the level of gene expression consistently seems to be a major determinant of variation in rates of evolution among proteins^{86,87}, other factors probably play a significant, if perhaps minor, part^{88–91}. In *Drosophila*, although highly expressed genes do evolve more slowly, breadth of expression across tissues, gene essentiality and intron number all also independently correlate with rates of protein evolution, suggesting that the additional complexities of multicellular organisms are important factors in modulating rates of protein evolution⁷⁸. The presence of repetitive amino acid sequences has a role as well: non-repeat regions in proteins containing repeats evolve faster and show more evidence for positive selection than genes lacking repeats⁹².

These data also provide a unique opportunity to examine the impact of chromosomal location on evolutionary rates. Population genetic theory predicts that for new recessive mutations, both purifying and positive selection will be more efficient on the X chromosome given its hemizyosity in males⁹³. In contrast, the lack of recombination on the small, mainly heterochromatic dot chromosome^{94,95} is expected to reduce the efficacy of selection⁹⁶. Because codon bias, or the unequal usage of synonymous codons in protein-coding sequences, reflects weak but pervasive selection, it is a sensitive metric for evaluating the efficacy of purifying selection. Consistent with expectation, in all 12 species, we find significantly elevated levels of codon bias on the X chromosome and significantly reduced levels of codon bias on the dot chromosome⁹⁷. Furthermore, X-chromosome-linked genes are marginally over-represented within the set of positively selected genes in the *melanogaster* group (FET, $P = 0.055$), which is consistent with increased rates of adaptive substitution on this chromosome. This analysis suggests that chromosomal context also serves to modulate rates of molecular evolution in protein-coding genes.

To examine further the impact of genomic location on protein evolution, we examined the subset of genes that have moved within or between chromosome arms^{32,98}. Genes inferred to have moved between Muller elements have a significantly higher rate of protein evolution than genes inferred to have moved within a Muller element (MWU, $P = 1.32 \times 10^{-14}$) and genes that have maintained their genomic position (MWU, $P = 0.008$) (Supplementary Fig. 7). Interestingly, genes that move within Muller elements have a significantly lower rate of protein evolution than those for which genomic locations have been maintained (MWU, $P = 3.85 \times 10^{-14}$). It remains unclear whether these differences reflect underlying biases in the types of genes that move inter- versus intra-chromosomally, or whether they are due to *in situ* patterns of evolution in novel genomic contexts.

Codon bias. Codon bias is thought to enhance the efficiency and/or accuracy of translation^{99–101} and seems to be maintained by mutation–selection–drift balance^{101–104}. Across the 12 *Drosophila* genomes, there is more codon bias in the *Sophophora* subgenus than in the *Drosophila* subgenus, and a previously noted^{105–109} striking reduction in codon bias in *D. willistoni*^{110,111} (Fig. 5). However, with only minor exceptions, codon preferences for each amino acid seem to be conserved across 11 of the 12 species. The striking exception is *D. willistoni*, in which codon usage for 6 of 18 redundant amino acids has diverged (Fig. 5). Mutation alone is not sufficient to explain codon-usage bias in *D. willistoni*, which is suggestive of a lineage-specific shift in codon preferences^{111,112}. We found evidence for a lineage-specific genomic reduction in codon bias in *D. melanogaster* (Fig. 5), as has been suggested previously^{113–119}. In addition, maximum-likelihood estimation of the strength of selection on synonymous sites in 8,510 *melanogaster* group single-copy orthologues revealed a marked reduction in the number of genes under selection

for increased codon bias in *D. melanogaster* relative to its sister species *D. sechellia*¹²⁰.

Evolution of genes associated with ecology and reproduction. Given the ecological and environmental diversity encompassed by the 12 *Drosophila* species, we examined the evolution of genes and gene families associated with ecology and reproduction. Specifically, we selected genes with roles in chemoreception, detoxification/metabolism, immunity/defence, and sex/reproduction for more detailed study.

Chemoreception. *Drosophila* species have complex olfactory and gustatory systems used to identify food sources, hazards and mates, which depend on odorant-binding proteins, and olfactory/odorant and gustatory receptors (*Ors* and *Gr*s). The *D. melanogaster* genome has approximately 60 *Ors*, 60 *Gr*s and 50 odorant-binding protein genes. Despite overall conservation of gene number across the 12 species and widespread evidence for purifying selection within the *melanogaster* group, there is evidence that a subset of *Or* and *Gr* genes experiences positive selection^{121–123}. Furthermore, clear lineage-specific differences are detectable between generalist and specialist species within the *melanogaster* subgroup. First, the two independently evolved specialists (*D. sechellia* and *D. erecta*) are losing *Gr* genes approximately five times more rapidly than the generalist species^{121,124}. We believe this result is robust to sequence quality, because all pseudogenes and deletions were verified by direct re-sequencing and synteny-based orthologue searches, respectively. Generalists are expected to encounter the most diverse set of tastants and seem to have maintained the greatest diversity of gustatory receptors. Second, *Or* and *Gr* genes that remain intact in *D. sechellia* and *D. erecta* evolve significantly more rapidly along these two lineages ($\omega = 0.1556$ for *Ors* and 0.1874 for *Gr*s) than along the generalist lineages ($\omega = 0.1049$ for *Ors* and 0.1658 for *Gr*s; paired Wilcoxon, $P = 0.0003$ and 0.003, respectively¹²⁴). There is some evidence that odorant-binding protein genes also evolve significantly faster in specialists compared to generalists¹²². This elevated ω reflects a trend observed throughout the genomes of the two specialists and is likely to result, at least in part, from demographic phenomena. However, the difference between specialist and generalist ω for *Or/Gr* genes (0.0292) is significantly greater than the difference for genes across the genome (0.0091; MWU, $P = 0.0052$)¹²¹, suggesting a change in selective regime. Moreover, the observation that elevated ω as well as accelerated gene loss disproportionately affect groups of *Or* and *Gr* genes that respond to specific chemical ligands and/or are expressed during specific life stages suggests that rapid evolution at *Or/Gr* loci in specialists is related to the ecological shifts these species have sustained¹²¹.

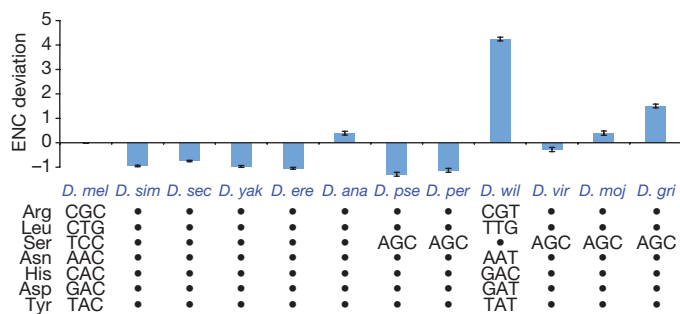


Figure 5 | Deviations in codon bias from *D. melanogaster* in 11 *Drosophila* species. The upper panel depicts differences in ENC (effective number of codons) between *D. melanogaster* and the 11 non-*melanogaster* species, calculated on a gene-by-gene basis. Note that increasing levels of ENC indicates a decrease in codon bias. The *Sophophora* subgenus in general has higher levels of codon bias than the *Drosophila* subgenus with the exception of *D. willistoni*, which shows a dramatic reduction in codon bias. The lower panel shows the 7 codons for which preference changes across the 12 *Drosophila* species. A dot indicates identical codon preference to *D. melanogaster*; otherwise the preferred codon is indicated.

Detoxification/metabolism. The larval food sources for many *Drosophila* species contain a cocktail of toxic compounds, and consequently *Drosophila* genomes encode a wide variety of detoxification proteins. These include members of the cytochrome P450 (P450), carboxyl/choline-esterase (CCE) and glutathione S-transferase (GST) multigene families, all of which also have critical roles in resistance to insecticides^{125–127}. Among the P450s, the five enzymes associated with insecticide resistance are highly dynamic across the phylogeny, with 24 duplication events and 4 loss events since the last common ancestor of the genus, which is in striking contrast to genes with known developmental roles, eight of which are present as a single copy in all 12 species (C. Robin, personal communication). As with chemoreceptors, specialists seem to lose detoxification genes at a faster rate than generalists. For instance, *D. sechellia* has lost the most P450 genes; these 14 losses comprise almost one-third of all P450 loss events (Supplementary Table 13) (C. Robin, personal communication). Positive selection has been implicated in detoxification-gene evolution as well, because a search for positive selection among GSTs identified the parallel evolution of a radical glycine to lysine amino acid change in GSTD1, an enzyme known to degrade DDT¹²⁸. Finally, although metabolic enzymes in general are highly constrained (median $\omega = 0.045$ for enzymes, 0.066 for non-enzymes; MWU, $P = 5.7 \times 10^{-24}$), enzymes involved in xenobiotic metabolism evolve significantly faster than other enzymes (median $\omega = 0.05$ for the xenobiotic group versus $\omega = 0.045$ overall, two-tailed permutation test, $P = 0.0110$; A. J. Greenberg, personal communication).

Metazoans deal with excess selenium in the diet by sequestration in selenoproteins, which incorporate the rare amino acid selenocysteine (Sec) at sites specified by the TGA codon. The recoding of the normally terminating signal TGA as a Sec codon is mediated by the selenocystein insertion sequence (SECIS), a secondary structure in the 3' UTR of selenoprotein messenger RNAs. All animals examined so far have selenoproteins; three have been identified in *D. melanogaster* (SELG, SELM and SPS2^{129,130}). Interestingly, although the three known *melanogaster* selenoproteins are all present in the genomes of the other *Drosophila* species, in *D. willistoni* the TGA Sec codons have been substituted by cysteine codons (TGT/TGC). Consistent with this finding, analysis of the seven genes implicated to date in selenoprotein synthesis including the Sec-specific tRNA suggests that most of these genes are absent in *D. willistoni* (R. Guigo, personal communication). *D. willistoni* thus seems to be the first animal known to lack selenoproteins. If correct, this observation is all the more remarkable given the ubiquity of selenoproteins and the selenoprotein biosynthesis machinery in metazoans, the toxicity of excess selenium, and the protection from oxidative stress mediated by selenoproteins. However, it remains possible that this species encodes selenoproteins in a different way, and this represents an exciting avenue of future research.

Immunity/defence. *Drosophila*, like all insects, possesses an innate immune system with many components analogous to the innate immune pathways of mammals, although it lacks an antibody-mediated adaptive immune system¹³¹. Immune system genes often evolve rapidly and adaptively, driven by selection pressures from pathogens and parasites^{132–134}. The genus *Drosophila* is no exception: immune system genes evolve more rapidly than non-immune genes, showing both high total divergence rates and specific signs of positive selection¹³⁵. In particular, 29% of receptor genes involved in phagocytosis seem to evolve under positive selection, suggesting that molecular co-evolution between *Drosophila* pattern recognition receptors and pathogen antigens is driving adaptation in the immune system¹³⁵. Somewhat surprisingly, genes encoding effector proteins such as antimicrobial peptides are far less likely to exhibit adaptive sequence evolution. Only 5% of effector genes (and no antimicrobial peptides) show evidence of adaptive evolution, compared to 10% of genes genome-wide. Instead, effector genes seem to evolve by rapid duplication and deletion. Whereas 49% of genes genome-wide, 63%

of genes involved in pathogen recognition and 81% of genes implicated in immune-related signal transduction can be found as single-copy orthologues in all 12 species, only 40% of effector genes exist as single-copy orthologues across the genus ($\chi^2 = 41.13$, $P = 2.53 \times 10^{-8}$), suggesting rapid radiation of effector protein classes along particular lineages¹³⁵. Thus, much of the *Drosophila* immune system seems to evolve rapidly, although the mode of evolution varies across immune-gene functional classes.

Sex/reproduction. Genes encoding sex- and reproduction-related proteins are subject to a wide array of selective forces, including sexual conflict, sperm competition and cryptic female choice, and to the extent that these selective forces are of evolutionary consequence, this should lead to rapid evolution in these genes¹³⁶ (for an overview see refs 137, 138). The analysis of 2,505 sex- and reproduction-related genes within the *melanogaster* group indicated that male sex- and reproduction-related genes evolve more rapidly at the protein level than genes not involved in sex or reproduction or than female sex- and reproduction-related genes (Supplementary Fig. 8). Positive selection seems to be at least partially responsible for these patterns, because genes involved in spermatogenesis have significantly stronger evidence for positive selection than do non-spermatogenesis genes (permutation test, $P = 0.0053$). Similarly, genes that encode components of seminal fluid have significantly stronger evidence for positive selection than 'non-sex' genes¹³⁹. Moreover, protein-coding genes involved in male reproduction, especially seminal fluid and testis genes, are particularly likely to be lost or gained across *Drosophila* species^{29,139}.

Evolutionary forces in the mitochondrial genome. Functional elements in mtDNA are strongly conserved, as expected: tRNAs are relatively more conserved than the mtDNA overall (average pairwise nucleotide distance = 0.055 substitutions per site for tRNAs versus 0.125 substitutions per site overall). We observe a deficit of substitutions occurring in the stem regions of the stem-loop structure in tRNAs, consistent with strong selective pressure to maintain RNA secondary structure, and there is a strong signature of purifying selection in protein-coding genes¹³. However, despite their shared role in aerobic respiration, there is marked heterogeneity in the rates of amino acid divergence between the oxidative phosphorylation enzyme complexes across the 12 species (NADH dehydrogenase, $0.059 > \text{ATPase}$, $0.042 > \text{CytB}$, $0.037 > \text{cytochrome oxidase}$, 0.020 ; mean pairwise d_N), which contrasts with the relative homogeneity in synonymous substitution rates. A model with distinct substitution rates for each enzyme complex rather than a single rate provides a significantly better fit to the data ($P < 0.0001$), suggesting complex-specific selective effects of mitochondrial mutations¹³.

Non-coding sequence evolution

ncRNA sequence evolution. The availability of complete sequence from 12 *Drosophila* genomes, combined with the tractability of RNA structure predictions, offers the exciting opportunity to connect patterns of sequence evolution directly with structural and functional constraints at the molecular level. We tested models of RNA evolution focusing on specific ncRNA gene classes in addition to inferring patterns of sequence evolution using more general datasets that are based on predicted intronic RNA structures.

The exquisite simplicity of miRNAs and their shared stem-loop structure makes these ncRNAs particularly amenable to evolutionary analysis. Most miRNAs are highly conserved within the *Drosophila* genus: for the 71 previously described miRNA genes inferred to be present in the common ancestor of these 12 species, mature miRNA sequences are nearly invariant. However, we do find a small number of substitutions and a single deletion in mature miRNA sequences (Supplementary Table 14), which may have functional consequences for miRNA–target interactions and may ultimately help identify targets through sequence covariation. Pre-miRNA sequences are also highly conserved, evolving at about 10% of the rate of synonymous sites.

To link patterns of evolution with structural constraints, we inferred ancestral pre-miRNA sequences and deduced secondary structures at each ancestral node on the phylogeny (Supplementary Information section 12.1). Although conserved miRNA genes show little structural change (little change in free energy), the five *melanogaster* group-specific miRNA genes (*miR-303* and the *mir-310/311/312/313* cluster) have undergone numerous changes across the entire pre-miRNA sequence, including the ordinarily invariant mature miRNA. Patterns of polymorphism and divergence in these lineage-specific miRNA genes, including a high frequency of derived mutations, are suggestive of positive selection¹⁴⁰. Although lineage-specific miRNAs may evolve under less constraint because they have fewer target transcripts in the genome, it is also possible that recent integration into regulatory networks causes accelerated rates of miRNA evolution.

We further investigated patterns of sequence evolution for the subset of 38 conserved pre-miRNAs with mature miRNA sequences at their 3' end by calculating evolutionary rates in distinct site classes (Fig. 6, and Supplementary Information section 12.2). Outside the mature miRNA and its complementary sequence, loops had the highest rate of evolution, followed by unpaired sites, with paired sites having the lowest rate of evolution. Inside the mature miRNA, unpaired sites evolve more slowly than paired sites, whereas the opposite is true for the sequence complementary to the mature miRNA. Surprisingly, a large fraction of unpaired bulges or internal loops in the mature miRNA seem to be conserved—a pattern which may have implications for models of miRNA biogenesis and the degree of mismatch allowed in miRNA–target prediction methods. Overall these results support the qualitative model proposed in ref. 141 for the canonical progression of miRNA evolution, and show that functional constraints on the miRNA itself supersede structural constraints imposed by maintenance of the hairpin-loop.

To assess constraint on stem regions of RNA structures more generally, we compared substitution rates in stems (*S*) to those in nominally unconstrained loop regions (*L*) in a wide variety of ncRNAs (Supplementary Information section 12.3). We estimated substitution rates using a maximum likelihood framework, and compared the observed *L/S* ratio with the average *L/S* ratio estimated from published secondary structures in RFAM, which we normalized to 1.0. *L/S* ratios for *Drosophila* ncRNA families range from a highly constrained 2.57 for the nuclear RNase P family to 0.56 for the 5S ribosomal RNA (Supplementary Table 15).

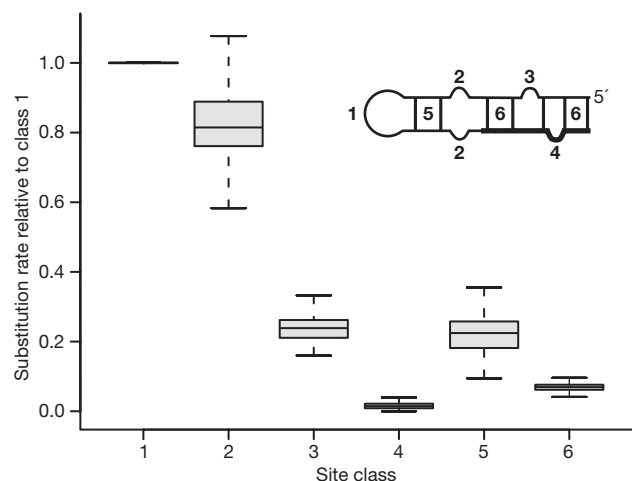


Figure 6 | Substitution rate of site classes within miRNAs. Bootstrap distributions of miRNA substitution rates. Structural alignments of miRNA precursor hairpins were partitioned into six site-classes (inset): (1) hairpin loops; unpaired sites (2) outside, (3) in the complementary region of, and (4) inside the miRNA; and base pairs (5) adjacent to and (6) involving the miRNA. Whiskers show approximate 95% confidence intervals for median differences, boxes show interquartile range.

Finally, we predicted a set of conserved intronic RNA structures and analysed patterns of compensatory nucleotide substitution in *D. melanogaster*, *D. yakuba*, *D. ananassae*, *D. pseudoobscura*, *D. virilis* and *D. mojavensis* (Supplementary Information section 13). Signatures of compensatory evolution in RNA helices are detected as covarying nucleotide sites or 'covariations' (that is, two Watson–Crick bases that interact in species A replaced by a different Watson–Crick pair in species B). The number of covariations (per base pair of a helix) depends on the physical distance between the interacting nucleotides (Supplementary Fig. 9), as has been observed for the RNA helices in the *Drosophila bicoid* 3' UTR region¹⁴². Short-range pairings exhibit a higher average number of covariations with a larger variance among helices than longer-range pairings. The decrease in rate of covariation with increasing distance may be explained by physical properties of a helix, which may impose selective constraints on the evolution of covarying nucleotides within a helix. Alternatively, if individual mutations at each locus are deleterious but compensated by mutations at a second locus, given sufficiently strong selection against the first deleterious mutation these epistatic fitness interactions could generate the observed distance effect¹⁴³.

Evolution of *cis*-regulatory DNAs. Comparative analyses of *cis*-regulatory sequences may provide insights into the evolutionary forces acting on regulatory components of genes, shed light on the constraints of the *cis*-regulatory code and aid in annotation of new regulatory sequences. Here we rely on two recently compiled databases, and present results comparing *cis*-regulatory modules¹⁴⁴ and transcription factor binding sites (derived from DNase I footprints)¹⁴⁵ between *D. melanogaster* and *D. simulans* (Supplementary Information section 8). We estimated mean selective constraint (*C*, the fraction of mutations removed by natural selection) relative to the 'fastest evolving intron' sites at the 5' end of short introns, which represent putatively unconstrained neutral standards (Supplementary Information section 8.2)¹⁴⁶. Note that this approach ignores the contribution of positively selected sites, potentially underestimating the fraction of functionally relevant sites¹⁴⁷.

Consistent with previous findings, *Drosophila cis*-regulatory sequences are highly constrained^{148,149}. Mean constraint within *cis*-regulatory modules is 0.643 (95% bootstrap confidence interval = 0.621–0.662) and within footprints is 0.692 (0.655–0.723), both of which are significantly higher than mean constraint in non-coding DNA overall (0.555 (0.546–0.563)) and significantly lower than constraint at non-degenerate coding sites (0.862 (0.856–0.868)) and ncRNA genes (0.864 (0.846–0.880)) (Supplementary Fig. 10). The high level of constraint in *cis*-regulatory sequences also extends into flanking sequences, only declining to constraint levels typical of non-coding DNA 40 bp away. This is consistent with previous findings that transcription factor binding sites tend to be found in larger blocks of constraint that cluster to form *cis*-regulatory modules¹⁵⁰. To understand selective constraints on nucleotides within *cis*-regulatory sequences that have direct contact with transcription factors, we estimated the selective constraint for the best match to position weight matrices within each footprint¹⁵¹; core motifs in transcription-factor-binding sites have a mean constraint of 0.773 (0.729–0.814), significantly greater than the mean for the footprints as a whole, and approaching the level of constraint found at non-degenerate coding sites and in ncRNA genes (Supplementary Fig. 10).

We next examined the variation in selective constraint across *cis*-regulatory sequences. Surprisingly, we find no evidence that selective constraint is correlated with predicted transcription-factor-binding strength (estimated as the position weight matrix score *P*-value) (Spearman's $r = 0.0681$, $P = 0.0609$). We observe significant variation in constraint both among target genes (Kruskal–Wallis tests, footprints, $P < 0.0001$; and position weight matrix matches within footprints, $P = 0.0023$) and among chromosomes (*cis*-regulatory modules, $P = 0.0186$; footprints, $P = 0.0388$; and position weight

matrix matches within footprints, $P = 0.0108$; Supplementary Table 16).

Discussion and conclusion

Each new genome sequence affords novel opportunities for comparative genomic inference. What makes the analysis of these 12 *Drosophila* genomes special is the ability to place every one of these genomic comparisons on a phylogeny with a taxon separation that is ideal for asking a wealth of questions about evolutionary patterns and processes. It is without question that this phylogenomic approach places additional burdens on bioinformatics efforts, multiplying the amount of data many-fold, requiring extra care in generating multi-species alignments, and accommodating the reality that not all genome sequences have the same degree of sequencing or assembly accuracy. These difficulties notwithstanding, phylogenomics has extraordinary advantages not only for the analyses that are possible, but also for the ability to produce high-quality assemblies and accurate annotations of functional features in a genome by using closely related genomes as guides. The use of multi-species orthology provides especially convincing evidence in support of particular gene models, not only for protein-coding genes, but also for miRNA and other ncRNA genes.

Many attributes of the genomes of *Drosophila* are remarkably conserved across species. Overall genome size, number of genes, distribution of transposable element classes, and patterns of codon usage are all very similar across these 12 genomes, although *D. willistoni* is an exceptional outlier by several criteria, including its unusually skewed codon usage, increased transposable element content and potential lack of selenoproteins. At a finer scale, the number of structural changes and rearrangements is much larger; for example, there are several different rearrangements of genes in the *Hox* cluster found in these *Drosophila* species.

The vast majority of multigene families are found in all 12 genomes, although gene family size seems to be highly dynamic: almost half of all gene families change in size on at least one lineage, and a noticeable fraction shows rapid and lineage-specific expansions and contractions. Particularly notable are cases consistent with adaptive hypotheses, such as the loss of *Gr* genes in ecological specialists and the lineage-specific expansions of antimicrobial peptides and other immune effectors. All species were found to have novel genes not seen in other species. Although lineage-specific genes are challenging to verify computationally, we can confirm at least 44 protein-coding genes unique to the *melanogaster* group, and these proteins have very different properties from ancestral proteins. Similarly, although the relative abundance of transposable element subclasses across these genomes does not differ dramatically, total genomic transposable element content varies substantially among species, and several instances of lineage-specific transposable elements were discovered.

There is considerable variation among protein-coding genes in rates of evolution and patterns of positive selection. Functionally similar proteins tend to evolve at similar rates, although variation in genomic features such as gene expression level, as well as chromosomal location, are also associated with variation in evolutionary rate among proteins. Whereas broad functional classes do not seem to share patterns of positive selection, and although very few GO categories show excesses of positive selection, a number of genes involved in interactions with the environment and in sex and reproduction do show signatures of adaptive evolution. It thus seems likely that adaptation to changing environments, as well as sexual selection, shape the evolution of protein-coding genes.

Annotation of ncRNA genes across all 12 species allows comprehensive analysis of the evolutionary divergence of these genes. MicroRNA genes in particular are more conserved than protein-coding genes with respect to their primary DNA sequence, and the substitutions that do occur often have compensatory changes such that the average estimated free energy of the folding structures remains remarkably constant across the phylogeny. Surprisingly,

mismatches in miRNAs seem to be highly conserved, which may impact models of miRNA biogenesis and target recognition. Lineage-restricted miRNAs, however, have considerably elevated rates of change, suggesting either reduced constraint due to novel miRNAs having fewer targets, or adaptive evolution of evolutionarily young miRNAs.

Virtually any question about the function of genome features in *Drosophila* is now empowered by being embedded in the context of this 12 species phylogeny, allowing an analysis of the ways by which evolution has tuned myriad biological processes across the hundreds of millions of years spanned in total by this phylogeny. The analyses presented herein have generated more questions than they have answered, and these results represent a small fraction of that which is possible. Because much of this rich and extraordinary comparative genomic dataset remains to be explored, we believe that these 12 *Drosophila* genome sequences will serve as a powerful tool for glean-ing further insight into genetic, developmental, regulatory and evolu-tionary processes.

METHODS

The full methods for this paper are described in Supplementary Information. Here, we describe the datasets generated by this project and their availability.

Genomic sequence. Scaffolds and assemblies for all genomic sequence generated by this project are available from GenBank (Supplementary Tables 4 and 5), and FlyBase (ftp://ftp.flybase.net/12_species_analysis/). Genome browsers are available from UCSC (<http://genome.ucsc.edu/cgi-bin/hgGateway?hgid=98180333&clade=insect&org=0&db=0>) and Flybase (<http://flybase.org/cgi-bin/gbrowse/dmel/>). BLAST search of these genomes is available at FlyBase (<http://flybase.org/blast>).

Predicted gene models. Consensus gene predictions for the 11 non-*melanogaster* species, produced by combining several different GLEAN runs that weight homology evidence more or less strongly, are available from FlyBase as GFF files for each species (ftp://ftp.flybase.net/12_species_analysis/). These gene models can also be accessed from the Genome Browser in FlyBase (Gbrowse; <http://flybase.org/cgi-bin/gbrowse/dmel/>). Predictions of non-protein-coding genes are also available in GFF format for each species, from FlyBase (ftp://ftp.flybase.net/12_species_analysis/).

Homology. Multiway homology assignments are available from FlyBase (ftp://ftp.flybase.net/12_species_analysis/), and also in the Genome Browser (Gbrowse).

Alignments. All alignment sets produced are available in FASTA format from FlyBase (ftp://ftp.flybase.net/12_species_analysis/).

PAML parameters. Output from PAML models for the alignments of single copy orthologues in the *melanogaster* group, including the *q*-value for the test for positive selection, are available from FlyBase (ftp://ftp.flybase.net/12_species_analysis/).

Received 19 July; accepted 5 October 2007.

1. Markow, T. A. & O'Grady, P. M. *Drosophila* biology in the genomic age. *Genetics* doi:10.1534/genetics.107.074112 (in the press).
2. Powell, J. R. *Progress and Prospects in Evolutionary Biology: The Drosophila Model* (Oxford Univ. Press, Oxford, 1997).
3. Adams, M. D. *et al.* The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–2195 (2000).
4. Celniker, S. E. *et al.* Finishing a whole-genome shotgun: release 3 of the *Drosophila melanogaster* euchromatic genome sequence. *Genome Biol.* **3**, research0079.1–0079.14 (2002).
5. Richards, S. *et al.* Comparative genome sequencing of *Drosophila pseudoobscura*: chromosomal, gene, and *cis*-element evolution. *Genome Res.* **15**, 1–18 (2005).
6. Myers, E. W. *et al.* A whole-genome assembly of *Drosophila*. *Science* **287**, 2196–2204 (2000).
7. Fleischmann, R. D. *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496–512 (1995).
8. Stark *et al.* Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. *Nature* doi:10.1038/nature06340 (this issue).
9. Begun, D. J. *et al.* Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *PLoS Biol.* **5**, e310, doi:10.1371/journal.pbio.0050310 (2007).
10. Zimin, A. V., Smith, D. R., Sutton, G. & Yorke, J. A. Assembly reconciliation. *Bioinformatics* (in the press).
11. Clary, D. O. & Wolstenholme, D. R. The mitochondrial DNA molecule of *Drosophila yakuba*: nucleotide sequence, gene organization, and genetic code. *J. Mol. Evol.* **22**, 252–271 (1985).

12. Ballard, J. W. When one is not enough: introgression of mitochondrial DNA in *Drosophila*. *Mol. Biol. Evol.* **17**, 1126–1130 (2000).
13. Montooth, K. L., Abt, D. N., Hoffman, J. & Rand, D. M. Evolution of the mitochondrial DNA across twelve species of *Drosophila*. *Mol. Biol. Evol.* (submitted).
14. Salzberg, S. *et al.* Serendipitous discovery of *Wolbachia* genomes in multiple *Drosophila* species. *Genome Biol.* **6**, R23 (2005).
15. Edgar, R. C. & Myers, E. W. PILER: identification and classification of genomic repeats. *Bioinformatics* **21**, i152–i158 (2005).
16. Smith, C. D. *et al.* Improved repeat identification and masking in Dipterans. *Gene* **389**, 1–9 (2007).
17. Li, Q. *et al.* ReAS: Recovery of ancestral sequences for transposable elements from the unassembled reads of a whole shotgun. *PLoS Comput. Biol.* **1**, e43 (2005).
18. Bergman, C. M., Quesneville, H., Anxolabehere, D. & Ashburner, M. Recurrent insertion and duplication generate networks of transposable element sequences in the *Drosophila melanogaster* genome. *Genome Biol.* **7**, R112 (2006).
19. Guigo, R., Knudsen, S., Drake, N. & Smith, T. Prediction of gene structure. *J. Mol. Biol.* **226**, 141–157 (1992).
20. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
21. Gross, S. S. & Brent, M. R. Using multiple alignments to improve gene prediction. *J. Comput. Biol.* **13**, 379–393 (2006).
22. Gross, S. S., Do, C. B. & Batzoglou, S. in *BCATS 2005 Symposium Proc.* **82** (2005).
23. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14**, 988–995 (2004).
24. Slater, G. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).
25. Chatterji, S. & Pachter, L. Reference based annotation with GeneMapper. *Genome Biol.* **7**, R29 (2006).
26. Suvorov, A. *et al.* in *NCBI News Fall/Winter, NIH Publication No. 04-3272* (eds Benson, D & Wheeler, D) (2006).
27. Honeybee Genome Sequencing Consortium. Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature* **443**, 931–949 (2006).
28. Elsik, C. G. *et al.* Creating a honey bee consensus gene set. *Genome Biol.* **8**, R13 (2007).
29. Zhang, Y., Sturgill, D., Parisi, M., Kumar, S. & Oliver, B. Constraint and turnover in sex-biased gene expression in the genus *Drosophila*. *Nature* doi:10.1038/nature06323 (this issue).
30. Manak, J. R. *et al.* Biological function of unannotated transcription during the early development of *Drosophila melanogaster*. *Nature Genet.* **38**, 1151–1158 (2006).
31. Tatusov, R. L., Koonin, E. V. & Lipman, D. J. A genomic perspective on protein families. *Science* **278**, 631–637 (1997).
32. Bhutkar, A., Russo, S., Smith, T. F. & Gelbart, W. M. Techniques for multi-genome synteny analysis to overcome assembly limitations. *Genome Informatics* **17**, 152–161 (2006).
33. Heger, A. & Ponting, C. Evolutionary rate analyses of orthologues and paralogues from twelve *Drosophila* genomes. doi:10.1101/gr6249707 *Genome Res.* (in the press).
34. Notredame, C., Higgins, D. G. & Heringa, J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302**, 205–217 (2000).
35. Rat Genome Sequencing Project Consortium. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**, 493–521 (2004).
36. Waterston, R. H. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
37. Harrison, P. M., Milburn, D., Zhang, Z., Bertone, P. & Gerstein, M. Identification of pseudogenes in the *Drosophila melanogaster* genome. *Nucleic Acids Res.* **31**, 1033–1037 (2003).
38. Bosco, G., Campbell, P., Leiva-Neto, J. & Markow, T. Analysis of *Drosophila* species genome size and satellite DNA content reveals significant differences among strains as well as between species. *Genetics* doi:10.1534/Genetics107.075069 (in the press).
39. Ranz, J. *et al.* Principles of genome evolution in the *Drosophila melanogaster* species group. *PLoS Biol.* **5**, e152, doi:10.1371/journal.pbio.0050152 (2007).
40. Noor, M. A. F., Garfield, D. A., Schaeffer, S. W. & Machado, C. A. Divergence between the *Drosophila pseudoobscura* and *D. persimilis* genome sequences in relation to chromosomal inversions. *Genetics* doi:10.1534/genetics.107.070672 (in the press).
41. Lewis, E. B. A gene complex controlling segmentation in *Drosophila*. *Nature* **276**, 565–570 (1978).
42. Negre, B., Ranz, J. M., Casals, F., Caceres, M. & Ruiz, A. A new split of the *Hox* gene complex in *Drosophila*: relocation and evolution of the gene labial. *Mol. Biol. Evol.* **20**, 2042–2054 (2003).
43. Von Allmen, G. *et al.* Splits in fruitfly *Hox* gene complexes. *Nature* **380**, 116 (1996).
44. Negre, B. & Ruiz, A. HOM-C evolution in *Drosophila*: is there a need for *Hox* gene clustering? *Trends Genet.* **23**, 55–59 (2007).
45. Dowsett, A. P. & Young, M. W. Differing levels of dispersed repetitive DNA among closely related species of *Drosophila*. *Proc. Natl Acad. Sci.* **79**, 4570–4574 (1982).
46. Kapitonov, V. V. & Jurka, J. DNAREP1_DM. (Rebase Update Release 3.4, 1999).
47. Kapitonov, V. V. & Jurka, J. Molecular paleontology of transposable elements in the *Drosophila melanogaster* genome. *Proc. Natl Acad. Sci. USA* **100**, 6569–6574 (2003).

48. Singh, N. D., Arndt, P. F. & Petrov, D. A. Genomic heterogeneity of background substitutional patterns in *Drosophila melanogaster*. *Genetics* **169**, 709–722 (2004).
49. Yang, H.-P., Hung, T.-L., You, T.-L. & Yang, T.-H. Genomewide comparative analysis of the highly abundant transposable element *DINE-1* suggests a recent transpositional burst in *Drosophila yakuba*. *Genetics* **173**, 189–196 (2006).
50. Yang, H.-P. & Barbash, D. Abundant and species-specific miniature inverted-repeat transposable elements in 12 *Drosophila* genomes. *Genome Biol.* (submitted).
51. Wilder, J. & Hollocher, H. Mobile elements and the genesis of microsatellites in dipterans. *Mol. Biol. Evol.* **18**, 384–392 (2001).
52. Marzo, M., Puig, M. & Ruiz, A. The foldback-like element *Galileo* belongs to the P superfamily of DNA transposons and is widespread within the genus *Drosophila*. *Proc. Natl Acad. Sci. USA* (submitted).
53. Casola, C., Lawing, A., Betran, E. & Feschotte, C. PIF-like transposons are common in *Drosophila* and have been repeatedly domesticated to generate new host genes. *Mol. Biol. Evol.* **24**, 1872–1888 (2007).
54. Abad, J. P. et al. Genomic analysis of *Drosophila melanogaster* telomeres: full-length copies of *HeT-A* and *TART* elements at telomeres. *Mol. Biol. Evol.* **21**, 1613–1619 (2004).
55. Abad, J. P. et al. *TAHRE*, a novel telomeric retrotransposon from *Drosophila melanogaster*, reveals the origin of *Drosophila* telomeres. *Mol. Biol. Evol.* **21**, 1620–1624 (2004).
56. Blackburn, E. H. Telomerases. *Annu. Rev. Biochem.* **61**, 113–129 (1992).
57. Villasante, A. et al. *Drosophila* telomeric retrotransposons derived from an ancestral element that as recruited to replace telomerase. *Genome Res.* (in the press).
58. International Chicken Genome Sequencing Consortium. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**, 695–716 (2004).
59. Lander, E. S. et al. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
60. *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282**, 2012–2018 (1998).
61. Mount, S. M., Gotea, V., Lin, C. F., Hernandez, K. & Makalowski, W. Spliceosomal small nuclear RNA genes in 11 insect genomes. *RNA* **13**, 5–14 (2007).
62. Schneider, C., Will, C. L., Brosius, J., Frilander, M. J. & Luhrmann, R. Identification of an evolutionarily divergent U11 small nuclear ribonucleoprotein particle in *Drosophila*. *Proc. Natl Acad. Sci. USA* **101**, 9584–9589 (2004).
63. Deng, X. & Meller, V. H. Non-coding RNA in fly dosage compensation. *Trends Biochem. Sci.* **31**, 526–532 (2006).
64. Amrien, H. & Axel, R. Genes expressed in neurons of adult male *Drosophila*. *Cell* **88**, 459–469 (1997).
65. Park, S.-W. et al. An evolutionarily conserved domain of roX2 RNA is sufficient for induction of H4-Lys16 acetylation on the *Drosophila* X chromosome. *Genetics* (in the press).
66. Stage, D. E. & Eickbush, T. H. Sequence variation within the rRNA gene loci of 12 *Drosophila* species. *Genome Res.* (in the press).
67. Hahn, M. W., De Bie, T., Stajich, J. E., Nguyen, C. & Cristianini, N. Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Res.* **15**, 1153–1160 (2005).
68. Hahn, M. W., Han, M. V. & Han, S.-G. Gene family evolution across 12 *Drosophila* genomes. *PLoS Biol.* **3**, e197 (2007).
69. Levine, M. T., Jones, C. D., Kern, A. D., Lindfors, H. A. & Begun, D. J. Novel genes derived from noncoding DNA in *Drosophila melanogaster* are frequently X-linked and exhibit testis-biased expression. *Proc. Natl Acad. Sci. USA* **103**, 9935–9939 (2006).
70. Ponce, R. & Hartl, D. L. The evolution of the novel *Sdic* gene cluster in *Drosophila melanogaster*. *Gene* **376**, 174–183 (2006).
71. Arguello, J. R., Chen, Y., Tang, S., Wang, W. & Long, M. Originiation of an X-linked testes chimeric gene by illegitimate recombination in *Drosophila*. *PLoS Genet.* **2**, e77 (2006).
72. Begun, D. J., Lindfore, H. A., Thompson, M. E. & Holloway, A. K. Recently evolved genes identified from *Drosophila yakuba* and *D. erecta* accessory gland expressed sequence tags. *Genetics* **172**, 1675–1681 (2006).
73. Betran, E., Thornton, K. & Long, M. Retroposed new genes out of the X in *Drosophila*. *Genome Res.* **12**, 1854–1859 (2002).
74. Yanai, I. et al. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* **21**, 650–659 (2005).
75. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13**, 555–556 (1997).
76. The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nature Genet.* **25**, 25–29 (2000).
77. Storey, J. D. A direct approach to false discovery rates. *J. R. Stat. Soc. B* **64**, 479–498 (2002).
78. Larracuente, A. M. et al. Evolution of protein-coding genes in *Drosophila*. *Trends Genet.* (submitted).
79. Yang, Z., Nielsen, R., Goldman, N. & Pedersen, A. M. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155**, 431–449 (2000).
80. Bergman, C. M. et al. Assessing the impact of comparative genomic sequence data on the functional annotation of the *Drosophila* genome. *Genome Biol.* **3**, research0086.1–0086.20 (2002).
81. Bierne, N. & Eyre Walker, A. C. The genomic rate of adaptive amino acid substitution in *Drosophila*. *Mol. Biol. Evol.* **21**, 1350–1360 (2004).
82. Sawyer, S. A., Kulathinal, R. J., Bustamante, C. D. & Hartl, D. L. Bayesian analysis suggests that most amino acid replacements in *Drosophila* are driven by positive selection. *J. Mol. Evol.* **57** (suppl. 1), S154–S164 (2003).
83. Sawyer, S. A., Parsch, J., Zhang, Z. & Hartl, D. L. Prevalence of positive selection among nearly neutral amino acid replacements in *Drosophila*. *Proc. Natl Acad. Sci. USA* **104**, 6504–6510 (2007).
84. Smith, N. G. & Eyre-Walker, A. Adaptive protein evolution in *Drosophila*. *Nature* **415**, 1022–1024 (2002).
85. Welch, J. J. Estimating the genomewide rate of adaptive protein evolution in *Drosophila*. *Genetics* **173**, 821–837 (2006).
86. Drummond, D. A., Bloom, J. D., Adami, C., Wilke, C. O. & Arnold, F. H. Why highly expressed proteins evolve slowly. *Proc. Natl Acad. Sci. USA* **102**, 14338–14343 (2005).
87. Drummond, D. A., Raval, A. & Wilke, C. O. A single determinant dominates the rate of yeast protein evolution. *Mol. Biol. Evol.* **23**, 327–337 (2006).
88. Pal, C., Papp, B. & Hurst, L. D. Highly expressed genes in yeast evolve slowly. *Genetics* **158**, 927–931 (2001).
89. Pal, C., Papp, B. & Lercher, M. J. An integrated view of protein evolution. *Nature Rev. Genet.* **7**, 337–348 (2006).
90. Wall, D. P. et al. Functional genomic analysis of the rates of protein evolution. *Proc. Natl Acad. Sci. USA* **102**, 5483–5488 (2005).
91. Rocha, E. P. The quest for the universals of protein evolution. *Trends Genet.* **22**, 412–416 (2006).
92. Huntley, M. A. & Clark, A. G. Evolutionary analysis of amino acid repeats across the genomes of 12 *Drosophila* species. *Mol. Biol. Evol.* (in the press).
93. Charlesworth, B., Coyne, J. A. & Barton, N. H. The relative rates of evolution of sex chromosomes and autosomes. *Am. Nat.* **130**, 113–146 (1987).
94. Larsson, J. & Meller, V. H. Dosage compensation, the origin and the afterlife of sex chromosomes. *Chromosome Res.* **14**, 417–431 (2006).
95. Riddle, N. C. & Elgin, S. C. The dot chromosome of *Drosophila*: insights into chromatin states and their change over evolutionary time. *Chromosome Res.* **14**, 405–416 (2006).
96. Gordo, I. & Charlesworth, B. Genetic linkage and molecular evolution. *Curr. Biol.* **11**, R684–R686 (2001).
97. Singh, N. D., Larracuente, A. M. & Clark, A. G. Contrasting the efficacy of selection on the X and autosomes in *Drosophila*. *Mol. Biol. Evol.* (submitted).
98. Bhutkar, A., Russo, S. M., Smith, T. F. & Gelbart, W. M. Genome scale analysis of positionally relocated genes. *Genome Res.* (in the press).
99. Akashi, H. & Eyre-Walker, A. Translational selection and molecular evolution. *Curr. Opin. Genet. Dev.* **8**, 688–693 (1998).
100. Akashi, H., Kliman, R. M. & Eyre-Walker, A. Mutation pressure, natural selection, and the evolution of base composition in *Drosophila*. *Genetica (Dordrecht)* **102–103**, 49–60 (1998).
101. Bulmer, M. The selection–mutation–drift theory of synonymous codon usage. *Genetics* **129**, 897–908 (1991).
102. McVean, G. A. T. & Charlesworth, B. A population genetic model for the evolution of synonymous codon usage: Patterns and predictions. *Genet. Res.* **74**, 145–158 (1999).
103. Sharp, P. M. & Li, W. H. An evolutionary perspective on synonymous codon usage in unicellular organisms. *J. Mol. Evol.* **24**, 28–38 (1986).
104. Akashi, H. & Schaeffer, S. W. Natural selection and the frequency distributions of “silent” DNA polymorphism in *Drosophila*. *Genetics* **146**, 295–307 (1997).
105. Powell, J. R., Sezzi, E., Moriyama, E. N., Gleason, J. M. & Caccione, A. Analysis of a shift in codon usage in *Drosophila*. *J. Mol. Evol.* **57**, S214–S225 (2003).
106. Anderson, C. L., Carew, E. A. & Powell, J. R. Evolution of the *Adh* locus in the *Drosophila willistoni* group: The loss of an intron, and shift in codon usage. *Mol. Biol. Evol.* **10**, 605–618 (1993).
107. Rodriguez-Trelles, F., Tarrío, R. & Ayala, F. J. Switch in codon bias and increased rates of amino acid substitution in the *Drosophila saltans* species group. *Genetics* **153**, 339–350 (1999).
108. Rodriguez-Trelles, F., Tarrío, R. & Ayala, F. J. Evidence for a high ancestral GC content in *Drosophila*. *Mol. Biol. Evol.* **17**, 1710–1717 (2000).
109. Rodriguez-Trelles, F., Tarrío, R. & Ayala, F. J. Fluctuating mutation bias and the evolution of base composition in *Drosophila*. *J. Mol. Evol.* **50**, 1–10 (2000).
110. Heger, A. & Ponting, C. Variable strength of translational selection among twelve *Drosophila* species. *Genetics* (in the press).
111. Vicario, S., Moriyama, E. N. & Powell, J. R. Codon Usage in Twelve Species of *Drosophila*. *BMC Evol. Biol.* (submitted).
112. Singh, N. D., Arndt, P. F. & Petrov, D. A. Minor shift in background substitutional patterns in the *Drosophila saltans* and *willistoni* lineages is insufficient to explain GC content of coding sequences. *BMC Biol.* **4**, 10.1186/1741-7007-4-37 (2006).
113. Akashi, H. Inferring weak selection from patterns of polymorphism and divergence at “silent” sites in *Drosophila* DNA. *Genetics* **139**, 1067–1076 (1995).
114. Akashi, H. Molecular evolution between *Drosophila melanogaster* and *D. simulans*: reduced codon bias, faster rates of amino acid substitution, and larger proteins in *D. melanogaster*. *Genetics* **144**, 1297–1307 (1996).

115. Akashi, H. *et al.* Molecular evolution in the *Drosophila melanogaster* species subgroup: Frequent parameter fluctuations on the timescale of molecular divergence. *Genetics* **172**, 1711–1726 (2006).
116. Bauer DuMont, V., Fay, J. C., Calabrese, P. P. & Aquadro, C. F. DNA variability and divergence at the *Notch* locus in *Drosophila melanogaster* and *D. simulans*: a case of accelerated synonymous site divergence. *Genetics* **167**, 171–185 (2004).
117. McVean, G. A. & Vieira, J. The evolution of codon preferences in *Drosophila*: a maximum-likelihood approach to parameter estimation and hypothesis testing. *J. Mol. Evol.* **49**, 63–75 (1999).
118. Nielsen, R., Bauer DuMont, V., Hubisz, M. J. & Aquadro, C. F. Maximum likelihood estimation of ancestral codon usage bias parameters in *Drosophila*. *Mol. Biol. Evol.* **24**, 228–235 (2007).
119. Begun, D. J. The frequency distribution of nucleotide variation in *Drosophila simulans*. *Mol. Biol. Evol.* **18**, 1343–1352 (2001).
120. Singh, N. S., Bauer DuMont, V. L., Hubisz, M. J., Nielsen, R. & Aquadro, C. F. Patterns of mutation and selection at synonymous sites in *Drosophila*. *Mol. Biol. Evol.* doi:10.1093/mbe/196 (in the press).
121. McBride, C. S. & Arguello, J. R. Five *Drosophila* genomes reveal non-neutral evolution and the signature of host specialization in the chemoreceptor superfamily. *Genetics* (in the press).
122. Vieira, F. G., Sanchez-Gracia, A. & Rozas, J. Comparative genomic analysis of the odorant-binding protein family in 12 *Drosophila* genomes: Purifying selection and birth-and-death evolution. *Genome Biol.* **8**, 235 (2007).
123. Gardiner, A., Barker, D., Butlin, R. K., Jordan, W. C. & Ritchie, M. G. *Drosophila* chemoreceptor evolution: Selection, specialisation and genome size. *Genome Biol.* (submitted).
124. McBride, C. S. Rapid evolution of smell and taste receptor genes during host specialization in *Drosophila sechellia*. *Proc. Natl Acad. Sci. USA* **104**, 4996–5001 (2007).
125. Ranson, H. *et al.* Evolution of supergene families associated with insecticide resistance. *Science* **298**, 179–181 (2002).
126. Tijet, N., Helvig, C. & Feyereisen, R. The cytochrome P450 gene superfamily in *Drosophila melanogaster*. *Gene* **262**, 189–198 (2001).
127. Claudianos, C. *et al.* A deficit of detoxification enzymes: pesticide sensitivity and environmental response in the honeybee. *Insect Mol. Biol.* **15**, 615–636 (2006).
128. Low, W. L. *et al.* Molecular evolution of glutathione S-transferases in the genus *Drosophila*. *Genetics* (in the press).
129. Castellano, S. *et al.* *In silico* identification of novel selenoproteins in the *Drosophila melanogaster* genome. *EMBO Rep.* **2**, 697–702 (2001).
130. Martin-Romero, F. J. *et al.* Selenium metabolism in *Drosophila*: selenoproteins, selenoprotein mRNA expression, fertility, and mortality. *J. Biol. Chem.* **276**, 29798–29804 (2001).
131. Lemaître, B. & Hoffmann, J. The host defense of *Drosophila melanogaster*. *Annu. Rev. Immunol.* **25**, 697–743 (2007).
132. Hughes, A. L. & Nei, M. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* **335**, 167–170 (1988).
133. Murphy, P. M. Molecular mimicry and the generation of host defense protein diversity. *Cell* **72**, 823–826 (1993).
134. Schlenke, T. A. & Begun, D. J. Natural selection drives *Drosophila* immune system evolution. *Genetics* **164**, 1471–1480 (2003).
135. Sackton, T. B. *et al.* The evolution of the innate immune system across *Drosophila*. *Nature Genet.* (submitted).
136. Civetta, A. & Singh, R. S. High divergence of reproductive tract proteins and their association with postzygotic reproductive isolation in *Drosophila melanogaster* and *Drosophila virilis* group species. *J. Mol. Evol.* **41**, 1085–1095 (1995).
137. Civetta, A. Shall we dance or shall we fight? Using DNA sequence data to untangle controversies surrounding sexual selection. *Genome* **46**, 925–929 (2003).
138. Clark, N. L., Aagard, J. E. & Swanson, W. J. Evolution of reproductive proteins from animals and plants. *Reproduction* **131**, 11–22 (2006).
139. Haerty, W. *et al.* Evolution in the fast lane: rapidly evolving sex- and reproduction-related genes in *Drosophila* species. *Genetics* (in the press).
140. Lu, J. *et al.* Adaptive evolution of newly-emerged microRNA genes in *Drosophila*. *Mol. Biol. Evol.* (submitted).
141. Lai, E. C., Tomancak, P., Williams, R. W. & Rubin, G. M. Computational identification of *Drosophila* microRNA genes. *Genome Biol.* **4**, R42 (2003).
142. Parsch, J., Braverman, J. M. & Stephan, W. Comparative sequence analysis and patterns of covariation in RNA secondary structures. *Genetics* **154**, 909–921 (2000).
143. Stephan, W. The rate of compensatory evolution. *Genetics* **144**, 419–426 (1996).
144. Gallo, S. M., Li, L., Hu, Z. & Halfon, M. S. REDfly: a Regulatory Element Database for *Drosophila*. *Bioinformatics* **22**, 381–383 (2006).
145. Bergman, C. M., Carlson, J. W. & Celniker, S. E. *Drosophila* DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, *Drosophila melanogaster*. *Bioinformatics* **21**, 1747–1749 (2005).
146. Halligan, D. L. & Keightley, P. D. Ubiquitous selective constraints in the *Drosophila* genome revealed by a genome-wide interspecies comparison. *Genome Res.* **16**, 875–884 (2006).
147. Andolfatto, P. Adaptive evolution of non-coding DNA in *Drosophila*. *Nature* **437**, 1149–1152 (2005).
148. Bird, C. P., Stranger, B. E. & Dermitzakis, E. T. Functional variation and evolution of non-coding DNA. *Curr. Opin. Genet. Dev.* **16**, 559–564 (2006).
149. Wittkopp, P. J. Evolution of cis-regulatory sequence and function in Diptera. *Heredity* **97**, 139–147 (2006).
150. Ludwig, M. Z., Patel, N. H. & Kreitman, M. Functional analysis of *eve stripe 2* enhancer evolution in *Drosophila*. *Development* **125**, 949–958 (1998).
151. Down, A. T. A., Bergman, C. M., Su, J. & Hubbard, T. J. P. Large scale discovery of promoter motifs in *Drosophila melanogaster*. *PLoS Comput. Biol.* **3**, e7 (2007).
152. Tamura, K., Subramanian, S. & Kumar, S. Temporal patterns of fruit fly (*Drosophila*) evolution revealed by mutation clocks. *Mol. Biol. Evol.* **21**, 36–44 (2004).
153. Kumar, S., Tamura, K. & Nei, M. MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment. *Brief. Bioinform.* **5**, 150–163 (2004).
154. Pollard, D. A., Iyer, V. N., Moses, A. M. & Eisen, M. B. Widespread discordance of gene trees with species tree in *Drosophila*: evidence for incomplete lineage sorting. *PLoS Genet.* **2**, e173 (2006).
155. Bhutkar, A., Gelbart, W. M. & Smith, T. F. Inferring genome-scale rearrangement phylogeny and ancestral gene order: A *Drosophila* case study. *Genome Biol.* (in the press).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements Agencourt Bioscience Corporation, The Broad Institute of MIT and Harvard and the Washington University Genome Sequencing Center were supported by grants and contracts from the National Human Genome Research Institute (NHGRI). T.C. Kaufman acknowledges support from the Indian Genomics Initiative.

Author Contributions The laboratory groups of A. G. Clark (including A. M. Larracuent, T. B. Sackton, and N. D. Singh) and Michael B. Eisen (including V. N. Iyer and D. A. Pollard) played the part of coordinating the primary writing and editing of the manuscript with the considerable help of D. R. Smith, C. M. Bergman, W. M. Gelbart, B. Oliver, T. A. Markow, T. C. Kaufman and M. Kellis. D. R. Smith served as primary coordinator for the assemblies. The remaining authors contributed either through their efforts in sequence production, assembly and annotation, or in the analysis of specific topics that served as the focus of more than 40 companion papers.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to A.G.C. (ac347@cornell.edu), M.B.E. (mbeisen@lbl.gov), D.R.S. (douglas.smith@agencourt.com), C.M.B. (casey.bergman@manchester.ac.uk), W.G. (gelbart@morgan.harvard.edu), B.O. (oliver@helix.nih.gov), T.A.M. (tmarkow@public.arl.arizona.edu), T.C.K. (kaufman@indiana.edu), M.K. (manoli@mit.edu), V.N.I. (venky@berkeley.edu), T.B.S. (tbs7@cornell.edu), A.M.L. (aml69@cornell.edu), D.A.P. (danielpollard@alum.bowdoin.edu), N.D.S. (nds25@cornell.edu), or collectively to 12flies@morgan.harvard.edu.

Drosophila 12 Genomes Consortium

Project Leaders Andrew G. Clark¹, Michael B. Eisen^{2,3}, Douglas R. Smith⁴, Casey M. Bergman⁵, Brian Oliver⁶, Therese A. Markow⁷, Thomas C. Kaufman⁸, Manolis Kellis^{9,10} & William Gelbart^{11,12}

Annotation Coordination Venky N. Iyer¹³ & Daniel A. Pollard¹⁴

Analysis/Writing Coordination Timothy B. Sackton^{1,15}, Amanda M. Larracuent¹ & Nadia D. Singh¹

Sequencing, Assembly, Annotation and Analysis Contributors Jose P. Abad¹⁶, Dawn N. Abt¹⁷, Boris Adryan¹⁸, Montserrat Aguade¹⁹, Hiroshi Akashi²⁰, Wyatt W. Anderson²¹, Charles F. Aquadro¹, David H. Ardell²², Roman Arguello²³, Carlo G. Artieri²⁴, Daniel A. Barbash¹, Daniel Barker²⁵, Paolo Barsanti²⁶, Phil Batterham²⁷, Serafim Batzoglou²⁸, Dave Begun²⁹, Arjun Bhutkar^{11,30}, Enrico Blanco³¹, Stephanie A. Bosak⁴, Robert K. Bradley³², Adrienne D. Brand⁴, Michael R. Brent³³, Angela N. Brooks¹³, Randall H. Brown³³, Roger K. Butlin³⁴, Corrado Caggese²⁶, Brian R. Calvi³⁵, A. Bernardo de Carvalho³⁶, Anat Caspi³², Sergio Castrezana³⁷, Susan E. Celniker², Jean L. Chang¹⁰, Charles Chapple³¹, Sourav Chatterji^{38,39}, Asif Chinwalla⁴⁰, Alberto Civetta⁴¹, Sandra W. Clifton⁴⁰, Josep M. Comeron⁴², James C. Costello⁴³, Jerry A. Coyne²³, Jennifer Daub⁴⁴, Robert G. David⁴, Arthur L. Delcher⁴⁵, Kim Delehaunty⁴⁰, Chuong B. Do²⁸, Heather Ebling⁴, Kevin Edwards⁴⁶, Thomas Eickbush⁴⁷, Jay D. Evans⁴⁸, Alan Filipitski⁴⁹, Sven Findeiß^{49,50}, Eva Freyhult²², Lucinda Fulton⁴⁰, Robert Fulton⁴⁰, Ana C. L. Garcia⁵¹, Anastasia Gardiner²⁵, David A. Garfield⁵², Barry E. Garvin⁴, Greg Gibson⁵³, Don Gilbert⁸, Sante Gnerre¹⁰, Jennifer Godfrey⁴⁰, Robert Good²⁷, Valer Gotea²⁰, Brenton Gravely⁵⁴, Anthony J. Greenberg¹, Sam Griffiths-Jones^{5,44}, Samuel Gross²⁸, Roderic Guigo^{31,55}, Erik A. Gustafson⁴, Wilfried Haerty²⁴, Matthew W. Hahn^{8,43}, Daniel L. Halligan⁵⁶, Aaron L. Halpern⁵⁷, Gillian M. Halter²⁰, Mira V. Han⁴³, Andreas Heger^{58,59}, LaDeana Hillier⁴⁰, Angie S. Hinrichs⁶⁰, Ian Holmes³², Roger A. Hoskins², Melissa J. Hubisz⁶¹, Dan Hultmark⁶², Melanie A. Huntley¹, David B. Jaffe¹⁰, Santosh Jagadeeshan²⁴, William R. Jeck⁶³, Justin Johnson⁵⁷, Corbin D. Jones⁶³, William C. Jordan⁶⁴, Gary H. Karpen^{13,65}, Eiko Kataoka⁶⁶, Peter D. Keightley⁵⁶, Pouya Kheradpour⁹, Ewen F. Kirkness⁵⁷, Leonardo B. Koerich³⁶, Karsten Kristiansen⁶⁷, Dave

Kudrna⁶⁸, Rob J. Kulathinal⁶⁹, Sudhir Kumar^{49,70}, Roberta Kwok⁸, Eric Lander¹⁰, Charles H. Langley²⁹, Richard Lipoent⁷¹, Brian P. Lazzaro⁷², So-Jeong Lee⁶⁸, Lisa Levesque⁴¹, Ruiqiang Li^{67,73}, Chiao-Feng Lin²⁰, Michael F. Lin^{9,10}, Kerstin Lindblad-Toh¹⁰, Ana Llopart⁴², Manyuan Long²³, Lloyd Low²⁷, Elena Lozovsky⁶⁹, Jian Lu²³, Meizhong Luo⁶⁸, Carlos A. Machado⁷, Wojciech Makalowski²⁰, Mar Marzo⁷⁴, Muneo Matsuda⁶⁶, Luciano Matzkin⁷, Bryant McAllister⁴², Carolyn S. McBride²⁹, Brendan McKernan⁷, Kevin McKernan⁴, Maria Mendez-Lago⁶, Patrick Minx⁴⁰, Michael U. Mollenhauer²⁰, Kristi Montooth¹⁷, Stephen M. Mount^{45,75}, Xu Mu²⁰, Eugene Myers⁷⁶, Barbara Negre⁷⁷, Stuart Newfield⁷⁰, Rasmus Nielsen⁷⁸, Mohamed A. F. Noor⁵², Patrick O'Grady⁷¹, Lior Pachter³⁸, Montserrat Papaceit¹⁹, Matthew J. Parisi⁴, Michael Parisi⁶, Leopold Parts⁹, Jakob S. Pedersen^{60,79}, Graziano Pesole⁸⁰, Adam M. Phillippy⁴⁵, Chris P. Ponting^{58,59}, Mihai Pop⁴⁵, Damiano Porcelli²⁶, Jeffrey R. Powell⁸¹, Sonja Prohaska^{49,82}, Kim Pruitt⁸³, Marta Puig⁷⁴, Hadi Quesneville⁸⁴, Kristipati Ravi Ram¹, David Rand¹⁷, Matthew D. Rasmussen⁹, Laura K. Reed⁵³, Robert Reenan⁸⁵, Amy Reily⁴⁰, Karin A. Remington⁵⁷, Tania T. Rieger⁸⁶, Michael G. Ritchie²⁵, Charles Robin²⁷, Yu-Hui Rogers⁵⁷, Claudia Rohde⁸⁷, Julio Rozas¹⁹, Marc J. Rubenfield⁴, Alfredo Ruiz⁷⁴, Susan Russo^{11,12}, Steven L. Salzberg⁴⁵, Alejandro Sanchez-Gracia^{19,88}, David J. Saranga⁴, Hajime Sato⁶⁶, Stephen W. Schaeffer²⁰, Michael C. Schatz⁴⁵, Todd Schlenke⁸⁹, Russell Schwartz²⁰, Carmen Segarra¹⁹, Rama S. Singh²⁴, Laura Sirota¹, Marina Sirota⁹¹, Nicholas B. Sineres⁶⁸, Chris D. Smith^{65,92}, Temple F. Smith³⁰, John Spieth⁴⁰, Deborah E. Stage⁴⁷, Alexander Stark^{9,10}, Wolfgang Stephan⁹³, Robert L. Strausberg⁵⁷, Sebastian Strempel⁹³, David Sturgill⁶, Granger Sutton⁵⁷, Granger G. Sutton⁵⁷, Wei Tao⁴, Sarah Teichmann¹⁸, Yoshiko N. Tobari⁹⁴, Yoshihiko Tomimura⁹⁵, Jason M. Tosol⁴, Vera L. S. Valente⁵¹, Eli Venter⁵⁷, J. Craig Venter⁵⁷, Saverio Vicario⁸¹, Filipe G. Vieira¹⁹, Albert J. Vilella^{19,96}, Alfredo Villasante¹⁶, Brian Walenz⁵⁷, Jun Wang^{67,73}, Marvin Wasserman⁹⁷, Thomas Watts⁷, Derek Wilson¹⁸, Richard K. Wilson⁴⁰, Rod A. Wing⁶⁸, Mariana F. Wolfner¹, Alex Wong¹, Gane Ka-Shu Wong^{73,98}, Chung-I Wu²³, Gabriel Wu³², Daisuke Yamamoto⁹⁹, Hsiao-Pei Yang¹, Shiao-Pyng Yang⁴⁰, James A. Yorke¹⁰⁰, Kiyohito Yoshida¹⁰¹, Evgeny Zdobnov¹⁰², Peili Zhang^{11,12}, Yu Zhang⁶, Aleksey V. Zimin¹⁰⁰, Broad Institute Genome Sequencing Platform* & Broad Institute Whole Genome Assembly Team*

¹Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York 14853, USA. ²Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA. ³Center for Integrative Genomics, Department of Molecular and Cell Biology, University of California at Berkeley, Berkeley, California 94720, USA. ⁴Agencourt Bioscience Corporation, Beverly, Massachusetts 01915, USA. ⁵Faculty of Life Sciences, University of Manchester, Manchester M13 9PT, UK. ⁶Laboratory of Cellular and Developmental Biology, National Institutes of Health, Bethesda, Maryland 20892, USA. ⁷Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721, USA. ⁸Department of Biology, Indiana University, Bloomington, Indiana 47405, USA. ⁹Computer Science and Artificial Intelligence Laboratory, Cambridge, Massachusetts 02139, USA. ¹⁰Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA. ¹¹Department of Molecular and Cellular Biology, Harvard University, Cambridge, Massachusetts 02138, USA. ¹²FlyBase, The Biological Laboratories, Harvard University, Cambridge, Massachusetts 02138, USA. ¹³Department of Molecular and Cell Biology, University of California at Berkeley, Berkeley, California 94720, USA. ¹⁴Biophysics Graduate Group, University of California at Berkeley, Berkeley, California 94720, USA. ¹⁵Field of Ecology and Evolutionary Biology, Cornell University, Ithaca, New York 14853, USA. ¹⁶Centro de Biología Molecular Severo Ochoa, Universidad Autónoma de Madrid, Madrid 28049, Spain. ¹⁷Department of Ecology and Evolutionary Biology, Brown University, Providence, Rhode Island 02912, USA. ¹⁸Structural Studies Division, MRC Laboratory of Molecular Biology, Cambridge CB2 2QH, UK. ¹⁹Departament de Genètica, Universitat de Barcelona, Barcelona 08071, Spain. ²⁰Department of Biology, Pennsylvania State University, University Park, Pennsylvania 16802, USA. ²¹Department of Genetics, University of Georgia, Athens, Georgia 30602, USA. ²²Linnaeus Centre for Bioinformatics, Uppsala Universitet, Uppsala, SE-75124, Sweden. ²³Department of Ecology and Evolution, University of Chicago, Chicago, Illinois 60637, USA. ²⁴Department of Biology, McMaster University, Hamilton, Ontario, L8S 4K1, Canada. ²⁵School of Biology, University of St. Andrews, Fife KY16 9TH, UK. ²⁶Dipartimento di Genetica e Microbiologia dell'Università di Bari, Bari, 70126, Italy. ²⁷Department of Genetics, University of Melbourne, Melbourne 3010, Australia. ²⁸Computer Science Department, Stanford University, Stanford, California 94305, USA. ²⁹Section of Evolution and Ecology and Center for Population Biology, University of California at Davis, Davis, California 95616, USA. ³⁰BioMolecular Engineering Research Center, Boston University, Boston, Massachusetts 02215, USA. ³¹Research Group in Biomedical Informatics, Institut Municipal d'Investigació Mèdica, Universitat Pompeu Fabra, Barcelona 08003, Catalonia, Spain. ³²Department of Bioengineering, University of California at Berkeley, Berkeley, California 94720, USA. ³³Laboratory for Computational Genomics, Washington University, St. Louis, Missouri 63108, USA. ³⁴Animal and Plant Sciences, The University of Sheffield, Sheffield S10 2TN, UK. ³⁵Department of Biology, Syracuse University, Syracuse, New York 13244, USA. ³⁶Departamento de Genética, Universidade Federal do Rio de Janeiro, Rio de Janeiro 21944-970, Brazil. ³⁷Tucson Stock Center, Tucson, Arizona 85721, USA. ³⁸Department of Mathematics, University of California at Berkeley, Berkeley, California 94720, USA. ³⁹Genome Center, University of California at Davis, Davis, California 95616, USA. ⁴⁰Genome Sequencing Center, Washington University School of Medicine, St. Louis, Missouri 63108, USA. ⁴¹Department of Biology, University of Winnipeg, Winnipeg, Manitoba R3B 2E9, Canada. ⁴²Department of Biological Sciences, University of Iowa, Iowa City, Iowa 52242, USA. ⁴³School of Informatics, Indiana University, Bloomington, Indiana 47405, USA. ⁴⁴Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. ⁴⁵Center for Bioinformatics and Computational Biology,

University of Maryland, College Park, Maryland 20742, USA. ⁴⁶Department of Biological Sciences, Illinois State University, Normal, Illinois 61790, USA. ⁴⁷Department of Biology, University of Rochester, Rochester, New York 14627, USA. ⁴⁸Bee Research Lab, USDA-ARS, Beltsville, Maryland 20705, USA. ⁴⁹Center for Evolutionary Functional Genomics, Bionodesign Institute, Arizona State University, Tempe, Arizona 85287, USA. ⁵⁰Department of Computer Science, University of Leipzig, Leipzig 04107, Germany. ⁵¹Departamento de Genética, Universidade Federal do Rio Grande do Sul, Porto Alegre/RS 68011, Brazil. ⁵²Department of Biology, Duke University, Durham, New Carolina 27708, USA. ⁵³Department of Genetics, North Carolina State University, Raleigh, North Carolina 27695, USA. ⁵⁴Health Center, University of Connecticut, Farmington, Connecticut 06030, USA. ⁵⁵Center of Genomic Regulation, Barcelona 8003, Catalonia, Spain. ⁵⁶Institute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3JT, UK. ⁵⁷J. Craig Venter Institute, Rockville, Maryland 20850, USA. ⁵⁸MRC Functional Genetics Unit, University of Oxford, Oxford OX1 3QX, UK. ⁵⁹Department of Physiology, Anatomy and Genetics, University of Oxford, Oxford OX1 3QX, UK. ⁶⁰Center for Biomolecular Science and Engineering, University of California at Santa Cruz, Santa Cruz, California 95064, USA. ⁶¹Department of Human Genetics, University of Chicago, Chicago, Illinois 60637, USA. ⁶²Umeå Center for Molecular Pathogenesis, Umeå University, Umeå SE-90187, Sweden. ⁶³Department of Biology and Carolina Center for Genome Sciences, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. ⁶⁴Institute of Zoology, Regent's Park, London NW1 4RY, UK. ⁶⁵Drosophila Heterochromatin Genome Project, Department of Genome and Computational Biology, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA. ⁶⁶Kyoin University, School of Medicine, Mitaka, Tokyo 181-8611, Japan. ⁶⁷Department of Biochemistry and Molecular Biology, University of Southern Denmark, Odense M DK-5230, Denmark. ⁶⁸Arizona Genomics Institute, Department of Plant Sciences and BIO5, University of Arizona, Tucson, Arizona 85721, USA. ⁶⁹Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138, USA. ⁷⁰School of Life Sciences, Arizona State University, Tempe, Arizona 85287, USA. ⁷¹Department of Environmental Science, Policy and Management, University of California at Berkeley, Berkeley, California 94720, USA. ⁷²Department of Entomology, Cornell University, Ithaca, New York 14853, USA. ⁷³Beijing Genomics Institute at ShenZhen, ShenZhen 518083, China. ⁷⁴Departament Genètica i Microbiologia, Universitat Autònoma de Barcelona, Bellaterra 08193, Spain. ⁷⁵Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, Maryland 20742, USA. ⁷⁶Janelia Farm Research Campus, Howard Hughes Medical Institute, Ashburn, Virginia 20147-2408, USA. ⁷⁷Department of Zoology, University of Cambridge, Cambridge, CB2 3EJ, UK. ⁷⁸Institute of Biology, University of Copenhagen, DK-2100 Copenhagen Ø, Denmark. ⁷⁹Bioinformatics Centre, Department of Molecular Biology, University of Copenhagen, DK-2200 Copenhagen N, Denmark. ⁸⁰Dipartimento di Biochimica e Biologia Molecolare, Università di Bari and Istituto Tecnologie Biomediche del Consiglio Nazionale delle Ricerche, Bari 70126, Italy. ⁸¹Department of Ecology and Evolutionary Biology, Yale University, New Haven, Connecticut 06520, USA. ⁸²Department of Biomedical Informatics, Arizona State University, Tempe, Arizona 85287, USA. ⁸³National Center for Biotechnology Information, National Institutes of Health, Bethesda, Maryland 20894, USA. ⁸⁴Bioinformatics and Genomics Laboratory, Institut Jacques Monod, Paris, 75251, France. ⁸⁵Department of Molecular Biology, Cell Biology and Biochemistry, Brown University, Providence, Rhode Island 02912, USA. ⁸⁶Departamento de Genética, Centro de Ciências Biológicas, Universidade Federal de Pernambuco, Recife/PE 68011, Brazil. ⁸⁷Centro Acadêmico de Vitória, Universidade Federal de Pernambuco, Vitória de Santo Antão/PE, Brazil. ⁸⁸Cajal Institute, CSIC, Madrid 28002, Spain. ⁸⁹Department of Biology, Emory University, Atlanta, Georgia 30322, USA. ⁹⁰Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA. ⁹¹Biomedical Informatics, Stanford University, Stanford, California 94305, USA. ⁹²Department of Biology, San Francisco State University, San Francisco, California 94132, USA. ⁹³Department of Biology, University of Munich, 82152 Planegg-Martinsried, Germany. ⁹⁴Institute of Evolutionary Biology, Setagaya-ku, Tokyo 158-0098, Japan. ⁹⁵Shiba Gakuen, Minato-ku, Tokyo 105-0011, Japan. ⁹⁶European Bioinformatics Institute, Hinxton, CB10 1SD, UK. ⁹⁷Department of Biology, City University of New York at Queens, Flushing, New York 11367, USA. ⁹⁸Department of Biological Sciences and Department of Medicine, University of Alberta, Edmonton, Alberta T6G 2E9, Canada. ⁹⁹Department of Developmental Biology and Neurosciences, Tohoku University, Sendai 980-8578, Japan. ¹⁰⁰Institute for Physical Science and Technology, University of Maryland, College Park, Maryland 20742, USA. ¹⁰¹Hokkaido University, EESBIO, Sapporo, Hokkaido 060-0810, Japan. ¹⁰²Faculty of Medicine, Université de Genève, Geneva CH-1211, Switzerland.

***Broad Institute Genome Sequencing Platform** Jennifer Baldwin¹⁰, Amr Abdouelleil¹⁰, Jamal Abdulkadir¹⁰, Adal Abebe¹⁰, Briki Abera¹⁰, Justin Abreu¹⁰, St Christophe Acer¹⁰, Lynne Aftuck¹⁰, Allen Alexander¹⁰, Peter An¹⁰, Erica Anderson¹⁰, Scott Anderson¹⁰, Harindra Arachi¹⁰, Marc Azer¹⁰, Pasang Bachantsang¹⁰, Andrew Barry¹⁰, Tashi Bayul¹⁰, Aaron Berlin¹⁰, Daniel Bessette¹⁰, Toby Bloom¹⁰, Jason Blye¹⁰, Leonid Boguslavskiy¹⁰, Claude Bonnet¹⁰, Boris Boukhgalter¹⁰, Imane Bourzgui¹⁰, Adam Brown¹⁰, Patrick Cahill¹⁰, Sheridan Channer¹⁰, Yama Cheshatsang¹⁰, Lisa Chuda¹⁰, Mieke Citroen¹⁰, Alville Collymore¹⁰, Patrick Cooke¹⁰, Maura Costello¹⁰, Katie D'Aco¹⁰, Riza Daza¹⁰, Georgius De Haan¹⁰, Stuart DeGray¹⁰, Christina DeMaso¹⁰, Norbu Dhargay¹⁰, Kimberly Dooley¹⁰, Erin Dooley¹⁰, Missole Doricent¹⁰, Passang Dorje¹⁰, Kunsang Dorjee¹⁰, Alan Dupes¹⁰, Richard Elong¹⁰, Jill Falk¹⁰, Abderrahim Farina¹⁰, Susan Faro¹⁰, Diallo Ferguson¹⁰, Sheila Fisher¹⁰, Chelsea D. Foley¹⁰, Alicia Franke¹⁰, Dennis Friedrich¹⁰, Loryn Gadbois¹⁰, Gary Gearin¹⁰, Christina R. Gearin¹⁰, Georgia Giannoukos¹⁰, Tina Goode¹⁰, Joseph Graham¹⁰, Edward Grandbois¹⁰, Sharleen Grewal¹⁰, Kunsang Gyaltzen¹⁰, Nabil Hafez¹⁰, Birhane Hagos¹⁰, Jennifer Hall¹⁰, Charlotte Henson¹⁰, Andrew Hollinger¹⁰, Tracey Honan¹⁰, Monika D. Huard¹⁰, Leanne Hughes¹⁰, Brian Hurhula¹⁰, M Erii Husby¹⁰, Asha Kamat¹⁰, Ben Kanga¹⁰,

Seva Kashin¹⁰, Dmitry Khazanovich¹⁰, Peter Kisner¹⁰, Krista Lance¹⁰, Marcia Lara¹⁰, William Lee¹⁰, Niall Lennon¹⁰, Frances Letendre¹⁰, Rosie LeVine¹⁰, Alex Lipovsky¹⁰, Xiaohong Liu¹⁰, Jinlei Liu¹⁰, Shangtao Liu¹⁰, Tashi Lokyitsang¹⁰, Yeshi Lokyitsang¹⁰, Rakela Lubonja¹⁰, Annie Lui¹⁰, Pen MacDonald¹⁰, Vasilisa Magnisalis¹⁰, Kebede Maru¹⁰, Charles Matthews¹⁰, William McCusker¹⁰, Susan McDonough¹⁰, Teena Mehta¹⁰, James Meldrim¹⁰, Louis Meneus¹⁰, Oana Mihai¹⁰, Atanas Mihalev¹⁰, Tanya Mihova¹⁰, Rachel Mittelman¹⁰, Valentine Mlenga¹⁰, Anna Montmayeur¹⁰, Leonidas Mulrain¹⁰, Adam Navidi¹⁰, Jerome Naylor¹⁰, Tamrat Negash¹⁰, Thu Nguyen¹⁰, Nga Nguyen¹⁰, Robert Nicol¹⁰, Choe Norbu¹⁰, Nyima Norbu¹⁰, Nathaniel Novod¹⁰, Barry O'Neill¹⁰, Sahal Osman¹⁰, Eva Markiewicz¹⁰, Otero L. Oyono¹⁰, Christopher Patti¹⁰, Pema Phunkhang¹⁰, Fritz Pierre¹⁰, Margaret Priest¹⁰, Sujaa Raghuraman¹⁰, Filip Rege¹⁰, Rebecca Reyes¹⁰,

Cecil Rise¹⁰, Peter Rogov¹⁰, Keenan Ross¹⁰, Elizabeth Ryan¹⁰, Sampath Settipalli¹⁰, Terry Shea¹⁰, Ngawang Sherpa¹⁰, Lu Shi¹⁰, Diana Shih¹⁰, Todd Sparrow¹⁰, Jessica Spaulding¹⁰, John Stalker¹⁰, Nicole Stange-Thomann¹⁰, Sharon Stavropoulos¹⁰, Catherine Stone¹⁰, Christopher Strader¹⁰, Senait Tesfaye¹⁰, Talene Thomson¹⁰, Yama Thoulutsang¹⁰, Dawa Thoulutsang¹⁰, Kerri Topham¹⁰, Ira Topping¹⁰, Tsamla Tsamla¹⁰, Helen Vassiliev¹⁰, Andy Vo¹⁰, Tsering Wangchuk¹⁰, Tsering Wangdi¹⁰, Michael Weiland¹⁰, Jane Wilkinson¹⁰, Adam Wilson¹⁰, Shailendra Yadav¹⁰, Geneva Young¹⁰, Qing Yu¹⁰, Lisa Zembek¹⁰, Danni Zhong¹⁰, Andrew Zimmer¹⁰ & Zac Zwirko¹⁰ **Broad Institute Whole Genome Assembly Team** David B. Jaffe¹⁰, Pablo Alvarez¹⁰, Will Brockman¹⁰, Jonathan Butler¹⁰, CheeWhye Chin¹⁰, Sante Gnerre¹⁰, Manfred Grabherr¹⁰, Michael Kleber¹⁰, Evan Mauceli¹⁰ & Iain MacCallum¹⁰

APPENDIX

III

The *Foldback*-like element *Galileo* belongs to the *P* superfamily of DNA transposons and is widespread within the *Drosophila* genus

Mar Marzo, Marta Puig, and Alfredo Ruiz*

Departament de Genètica i de Microbiologia, Universitat Autònoma de Barcelona, Bellaterra, 08193 Barcelona, Spain

Communicated by Margaret G. Kidwell, University of Arizona, Tucson, AZ, December 28, 2007 (received for review August 8, 2007)

Galileo is the only transposable element (TE) known to have generated natural chromosomal inversions in the genus *Drosophila*. It was discovered in *Drosophila buzzatii* and classified as a *Foldback*-like element because of its long, internally repetitive, terminal inverted repeats (TIRs) and lack of coding capacity. Here, we characterized a seemingly complete copy of *Galileo* from the *D. buzzatii* genome. It is 5,406 bp long, possesses 1,229-bp TIRs, and encodes a 912-aa transposase similar to those of the *Drosophila melanogaster* 1360 (Hoppel) and *P* elements. We also searched the recently available genome sequences of 12 *Drosophila* species for elements similar to *DbuzGalileo* by using bioinformatic tools. *Galileo* was found in six species (*ananassae*, *willistoni*, *pseudoobscura*, *persimilis*, *virilis*, and *mojavensis*) from the two main lineages within the *Drosophila* genus. Our observations place *Galileo* within the *P* superfamily of cut-and-paste transposons and extend considerably its phylogenetic distribution. The interspecific distribution of *Galileo* indicates an ancient presence in the genus, but the phylogenetic tree built with the transposase amino acid sequences contrasts significantly with that of the species, indicating lineage sorting and/or horizontal transfer events. Our results also suggest that *Foldback*-like elements such as *Galileo* may evolve from DNA-based transposon ancestors by loss of the transposase gene and disproportionate elongation of TIRs.

class II elements | transposase | terminal inverted repeats | 1360 | inversions

Transposable elements (TEs) are intracellular parasites that populate most eukaryotic genomes and have a huge impact on their evolution (1). Their abundance and diversity are astonishing and a considerable effort is needed to put order in the increasing constellation of families being discovered. So far, two main classes are widely recognized, retrotransposons that transpose by an intermediate RNA molecule and transposons that move by using a single- or double-stranded DNA intermediate (2). Three subclasses of transposons have been defined based on the transposition mechanism: cut-and-paste, rolling-circle, and *Mavericks* (3). Cut-and-paste transposons possess TIRs, usually short, and encode a protein called transposase (TPase) that catalyzes their excision from the original location in the genome and promotes their reinsertion into a new site generating target site duplications (TSDs) in the process (4). The *Drosophila* elements *P* (5) and *mariner* (6) are among the best known families of cut-and-paste transposons but there are many more families classified in ten transposon superfamilies on the basis of similarity among the TPases: *Tc1/mariner*, *hAT*, *P*, *MuDR*, *CACTA*, *PiggyBac*, *PIF/Harbinger*, *Merlin*, *Transib*, and *Banshee* (3). Other elements are still unclassified, seemingly because only defective copies have been found. Defective (nonautonomous) copies coexist and often outnumber the canonical (autonomous) copies, and can move if there is a functional TPase provided by canonical copies present somewhere else in the same genome and if they conserve the signals required for TPase recognition (usually the TIR ends).

Foldback-like elements constitute a group of poorly known TEs with uncertain classification (2, 3). They take their name from the *Foldback* (*FB*) element of *Drosophila melanogaster* (7, 8) and are present in a diverse array of organisms (9–13). The unusual characteristics of *Foldback*-like elements include very long TIRs that make up almost the entire element and are separated by a middle domain with variable length and composition. No coding capacity has been found in many *Foldback*-like elements, and thus, their mechanism of transposition is uncertain. However, a small proportion ($\approx 10\%$) of *FB* copies in *D. melanogaster* is associated with a 4-kb-long sequence called *NOF* encoding a 120-kDa protein of unknown function (14, 15). *FB* has been recently included in the *MuDR* superfamily (3) because of the similarity of the proteins encoded by both *MuDR* and *NOF* to that of *Phantom*, a transposon from *Entamoeba* (16). Besides, some copies of *FARE*, another *Foldback*-like transposon from *Arabidopsis*, harbor a large ORF with weak similarity to the *MuDR* TPase (13). The origin of many other *Foldback*-like elements is still uncertain.

Galileo was discovered in *Drosophila buzzatii* and is the only TE in the genus *Drosophila* that has been shown to have generated chromosomal inversions in nature (17–19). Other TEs, such as *P*, *Hobo*, or *FB* are known to induce chromosomal rearrangements in experimental populations of *D. melanogaster* (20), but there is no direct evidence of their implication in *Drosophila* chromosomal evolution. *Galileo*, together with two closely related elements, *Kepler* and *Newton*, were classified as *Foldback*-like elements because of their long, internally repetitive TIRs (18, 21). All copies of *Galileo*, *Kepler*, and *Newton* isolated so far from the genome of *D. buzzatii* lack any significant protein-coding capacity except for two *Galileo* copies bearing a short segment with weak similarity to the TPase of element 1360 (Hoppel) (21). An experimental search for *Galileo* sequences in other *Drosophila* species suggested that this TE has a rather restricted distribution, being only present in the closest relatives of *D. buzzatii* but not in more distantly related species within the repleta group (21). Here, we take advantage of the recently sequenced genomes of *D. melanogaster* (22), *Drosophila pseudoobscura* (23), and ten additional *Drosophila* species (24) to search for sequences similar to *Galileo* in these genomes by using bioinformatic tools. We found that *Galileo* has a much wider species distribution within the *Drosophila* genus than previously suspected. Furthermore, our results allow us to fully characterize

Author contributions: A.R. designed research; M.M., M.P., and A.R. performed research; M.M., M.P., and A.R. analyzed data; and M.M., M.P., and A.R. wrote the paper.

The authors declare no conflict of interest.

Data deposition: Nucleotide sequences reported in this paper have been deposited in the DBJ/EMBL/GenBank databases [accession nos. EU334682–EU334685 and BK006357–BK006363 (TPA section)].

*To whom correspondence should be addressed. E-mail: Alfredo.Ruiz@uab.es.

This article contains supporting information online at www.pnas.org/cgi/content/full/0712110105/DC1.

© 2008 by The National Academy of Sciences of the USA

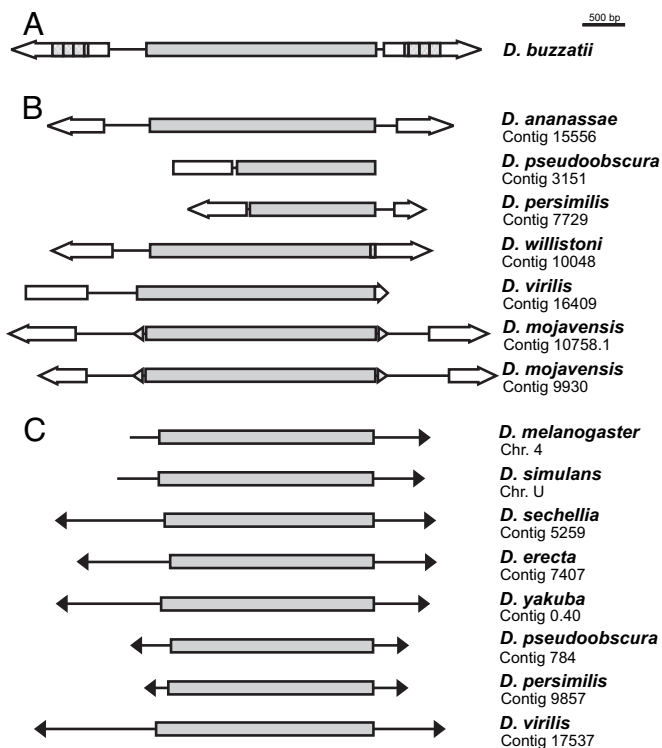


Fig. 1. Most complete copies of *Galileo* and *1360* found in this work. (A) Putative complete *Galileo* copy from the *D. buzzatii* genome. (B) Most complete copies of *Galileo* found in the 12 sequenced genomes. (C) Most complete copies of *1360*. TIRs are represented as arrows and TPases are represented as gray rectangles. The direct repeats of the TIRs in *DbuzGalileo* are indicated by striped patterns. *DmojGalileo* internal inverted repeats are represented as little triangles. In *D. mojavensis* two *Galileo* copies representative of two subfamilies found in this species are depicted. See [SI Table 4](#) for details.

the element *Galileo* and to classify it as a member of the *P* superfamily of cut-and-paste DNA transposons.

Results

Structure of *Galileo* in *D. buzzatii*. By using as a query *Galileo*-3, a defective copy of *DbuzGalileo* (21), we carried out preliminary bioinformatic searches in the genome sequence of *Drosophila mojavensis*, another member of the repleta species group. Some of the hits, on close examination, bounded a protein-coding segment that might be the *Galileo* TPase. Several PCRs were then attempted to isolate longer *Galileo* copies from the *D. buzzatii* genome (see *Methods*). In each of them, one primer was anchored in the known *DbuzGalileo* TIRs and the other in the possible *DmojGalileo* TPase. A putatively complete copy of *DbuzGalileo* could be assembled in this way (Fig. 1A). This copy is 5,406 bp long, possesses 1,229-bp TIRs and an intronless 2,738-bp ORF (nt 1348–4087) encoding a 912-aa protein (after fixing two STOP codons, and a 1-bp deletion that causes a frameshift mutation).

A search using BLASTX revealed significant similarity of the *DbuzGalileo* TPase to those of the related *D. melanogaster 1360* and *P* elements (25, 26) [AAN39288, E-value = $1e-95$; Q7M3K2, E-value = $3e-25$]. The *DbuzGalileo* TPase includes a THAP domain near the N terminus (amino acids 27–104) similar to the DNA binding domain of *P* element TPase (27–30). A copy of *1360* located in chromosome 4 of *D. melanogaster* (31) encodes a TPase (854 aa) longer than that in the National Center for Biotechnology Information database (25), including a THAP domain near the N terminus (after curation of a 1-bp frameshift mutation). A global alignment of the *DbuzGalileo* TPase with

those of *Dme1360* and *Dme1P* yielded 34.5% and 27.6% identity, respectively. No significant similarity was found between the *DbuzGalileo* TPase and the proteins encoded by *Dme1FB* (14, 15).

Distribution of *Galileo* and *1360* in the 12 Sequenced *Drosophila* Genomes. Systematic bioinformatic searches using as queries the TPases and TIRs of *DbuzGalileo* and *Dme1360* were carried out (see *Methods*). The results [[supporting information \(SI\) Tables 1–3](#)] suggested that elements similar to *Galileo* are present in *D. ananassae*, *D. pseudoobscura*, *D. persimilis*, *D. willistoni*, *D. virilis*, and *D. mojavensis*, whereas elements similar to *1360* are present in the five melanogaster subgroup species (*melanogaster*, *simulans*, *sechellia*, *yakuba*, and *erecta*) plus *D. pseudoobscura*, *D. persimilis*, and *D. virilis*. Therefore, none of the two TEs is seemingly present in *D. grimshawi* but both are found in *D. pseudoobscura*, *D. persimilis*, and *D. virilis*.

Characterization of *Galileo* Copies. We characterized 46 relatively long copies of *Galileo* containing segments encoding a partial or full TPase from the six genomes where this TE is present ([SI Table 4](#)). All of them possess one or two long TIRs with similarity to those of *DbuzGalileo* (see below) and nine are flanked by perfect 7-bp TSDs. The structure of the longest, presumably most complete, copy in each species is depicted in Fig. 1B. These *Galileo* copies are 4,386 bp (*D. willistoni*) to 5,989 bp long (*D. mojavensis*) and exhibit TIRs of 684 bp (*D. ananassae*) to 813 bp (*D. mojavensis*). However, none of them contains a single ORF encoding a fully functional TPase (all bear STOP codons, frameshift mutations, and/or deletions). In *D. mojavensis* 16 long copies were characterized. Many of them include nearly complete TPase-coding segments and all but three contain one or more insertions of other TEs ([SI Table 4](#)). These 16 copies belong to two groups with distinctive structures (see Fig. 1B for representative copies) and encoding somewhat different TPases (see below).

We also searched each of the six *Drosophila* genomes for short nonautonomous *Galileo* copies by using BLASTN and the most complete copy already found in the same genome (Fig. 1B) as query (see *Methods*). *Galileo* was rather abundant in the six genomes, the number of significant hits being >100 in all cases with a maximum of 495 in *D. willistoni* ([SI Table 1](#)). We identified and isolated 109 *Galileo* copies from the contigs producing significant hits in the six species. All of them possess two long TIRs separated by a relatively short middle segment and 97 show perfect 7-bp TSDs ([SI Table 5](#)). Thus, these copies are structurally similar to the copies of *Galileo*, *Kepler*, and *Newton* previously found in *D. buzzatii* (21). A summary of the characteristics of these relatively short nonautonomous copies is given in [SI Table 6](#).

TSDs. In *D. buzzatii*, *Galileo* generates on insertion 7-bp TSDs with the consensus GTAGTAC (21). Likewise, in the six *Drosophila* genomes analyzed here, 106 *Galileo* copies were flanked by identical 7-bp sequences ([SI Tables 4 and 5](#)). We calculated the frequency of the four nucleotides in each of the seven sites for each species separately. The frequency pattern observed in the six species was similar to that of *DbuzGalileo* and the 106 sequences were combined. All positions but the fourth show a significant departure from randomness, and the consensus is the palindrome GTANTAC.

Divergence Between *Galileo* Copies. To estimate the time since the most recent transpositional activity of *Galileo*, we measured the average pairwise divergence between the short nonautonomous copies within each species (see *Methods* and [SI Table 6](#)). In *D. ananassae*, the average pairwise divergence among 20 copies was 2.8%, which implies a divergence time of ≈ 1.8 myr. However,

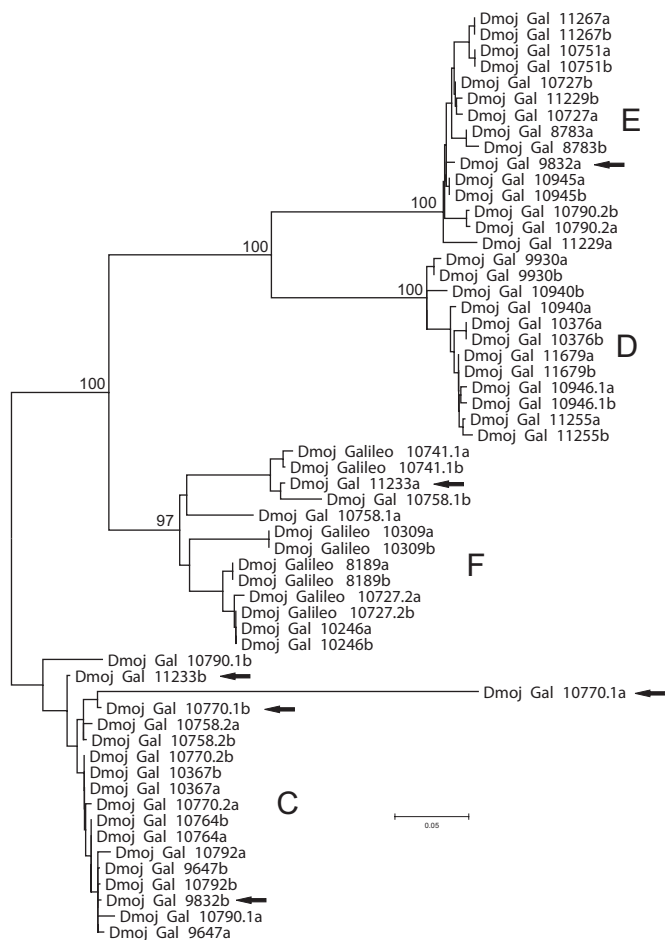


Fig. 2. Neighbor-joining phylogenetic tree inferred from the analysis of 29 *Galileo* copies found in the *D. mojavensis* genome. The two TIRs of each copy were included in the tree as separate sequences to allow their comparison within and between copies. TIRa is the TIR located at 5' from the TPase or the first TIR that appears in the contig if the copy could not be oriented. The complete deletion option was used leaving 269 informative sites. Bootstrap values at main nodes are shown. The average pairwise divergence between groups D and E is $\approx 25\%$, indicating a divergence time of ≈ 8 myr, and the average pairwise divergence between these two groups and groups C and F is $\approx 32\%$, implying a divergence time of ≈ 10 myr. The putative chimeric elements with highly divergent TIRs are marked with an arrow. Details of these *Galileo* copies are given in SI Tables 4 and 5.

evidence for more recent transpositional events was found because a subgroup of 13 copies shows an average divergence of 0.36% equivalent to a divergence time of only 0.225 myr. Similar observations were made in *D. pseudoobscura*, *D. persimilis*, and *D. willistoni* (SI Table 6). In each case, subgroups with $\approx 1\%$ average divergence (implying divergence times ≈ 0.6 myr) were found. In *D. virilis*, analysis of 13 short nonautonomous copies uncovered two highly divergent groups that we named A and B (SI Fig. 5). Copies within each group were aligned and analyzed separately (SI Table 6). The average pairwise divergence within groups A and B was 4.6 and 5.7%, implying divergence times of 2.9 and 3.6 myr, respectively. Inclusion in the analysis of the longest copy found in the species (contig 16409) indicated unequivocally that it is a member of group A (SI Fig. 5). In *D. mojavensis*, analysis of 20 short nonautonomous copies revealed the presence of four well defined groups, here named C–F. We included in the analysis nine of the long copies containing the two TIRs and generated a phylogenetic tree with the 29 copies (Fig. 2). Groups C and D correspond to the two groups

previously detected when the long, nearly complete, copies were analyzed. Copies within each group were separately aligned and analyzed. Average pairwise divergences within groups C through F were 2.2%, 2.3%, 2.4%, and 8.9%, respectively, indicating divergence times ranging from 1.4 to 5.5 myr (SI Table 6). The two and four *Galileo* groups or subfamilies found in *D. virilis* and *D. mojavensis*, respectively, seemingly represent relatively old transposition bursts in these genomes. We suggest that the *Newton* and *Kepler* elements previously found in the *D. buzzatii* genome (18, 21) should likewise be considered only as different groups or subfamilies of *Galileo* in this species.

One copy in *D. pseudoobscura* (contig 4355), one copy in *D. willistoni* (contig 10422), and three copies in *D. mojavensis* (contigs 11233, 10770.1, and 9832) are likely chimeric because they are flanked by dissimilar 7-bp sequences and show increased levels of divergence between the two TIRs (see for instance Fig. 2).

Characterization of 1360 Copies. The longest and complete or nearly complete copies of element 1360 found in the eight genomes are shown in Fig. 1C (see also SI Table 7). The eight copies possess TPase-coding segments 2,428 bp (*D. erecta*) to 2,565 bp long (*D. melanogaster*), although only *D. yakuba* includes three different copies with 2,562-bp ORFs encoding a fully functional TPase. All of them bear 31- or 32-bp-long TIRs and total size for seemingly complete copies varies between 2,985 bp (*D. persimilis*) and 4,702 bp (*D. virilis*). The longest copies found in each species (Fig. 1C) were used as queries to interrogate the eight genomes by using BLASTN. The results showed that 1360 is very abundant in all genomes with a maximum number of 690 significant hits in *D. sechellia* (SI Table 1).

Comparison of Galileo, 1360, and P Element TIRs. With the exception of *D. pseudoobscura* and *D. persimilis*, the long *Galileo* TIRs show little similarity between the different species either in length or sequence composition. Conservation seems to be restricted to the terminus as revealed by the alignment of the first 40 bp of *Galileo* in *D. buzzatii* (including *Kepler* and *Newton*) and the six species analyzed here (including *D. virilis* groups A and B and *D. mojavensis* groups C–F). A total of 17 of the 40 terminal bp are conserved in the 13 sequences (Fig. 3A). Likewise, alignment of the 31 bp of 1360 TIRs in the longest copies described earlier (Fig. 1C) revealed 14 conserved bp (Fig. 3B). We generated the consensus sequences of the element terminus in *Galileo* and 1360 in the different species. Fifteen of 31 bp are identical, which provides further evidence of the evolutionary relationship between both TEs. In addition, the consensus *Galileo* terminus shares 17 bp with the 31-bp TIRs of *Dme*ΔP (Fig. 3C).

Comparison of Galileo, 1360, and P Element TPases. We generated consensus amino acid sequences for the *Galileo* and 1360 TPases within each species (see Methods). For *Dmoj*/*Galileo*, the consensus sequences of the TPases encoded by copies in groups C and D are 937 and 936 aa long, respectively, and when aligned alone show a 87.2% identity and a 96.4% similarity.

A multiple alignment of the eight consensus *Galileo* TPases, the eight consensus 1360 TPases, and five TPases of representative *P* elements was carried out (SI Fig. 6). Besides, the human *P*-like THAP9 protein (32) was included in the analysis as outgroup. The *Galileo* TPases are 30–35% identical to those of 1360 and 20–25% identical to those of *P* elements (SI Table 8). Within the *Galileo* TPases, identity varies between 97.2% in the closely related pair *D. pseudoobscura*–*D. persimilis*, and 39.3% between *D. persimilis* and *D. virilis*. In addition, we examined the multiple alignment for conservation of several functional domains and motifs that have been identified in the *Dme*ΔP TPase (5). The THAP domain is a zinc-dependent DNA binding domain evolutionarily conserved in an array of different proteins including the *P* TPase, cell-cycle regulators, proapoptotic fac-

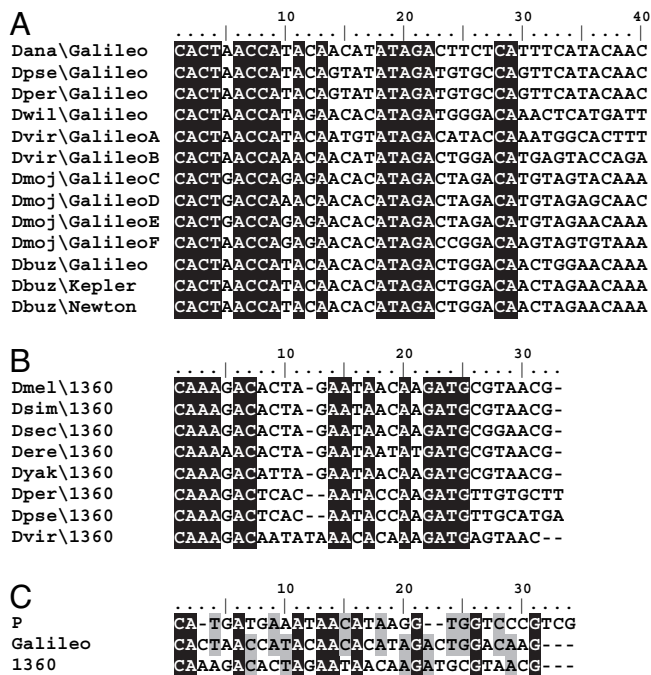


Fig. 3. Comparison of TIR ends. (A) Alignment of 40 bp of the TIR end of *Galileo*. A consensus sequence was constructed for *Galileo* TIRs in each TE subfamily and species. (B) Alignment of the 31-bp TIR of *1360*. A representative TIR from a single copy of the TE is included. (C) Comparison of the *Galileo* TIR end with the TIRs of elements *1360* and *P*. Identical positions in all sequences are shown in black. Sites identical between *Galileo* and *1360* or *P* are shown in gray.

tors, transcriptional repressors, and chromatin-associated proteins (28–30). It includes a metal-coordinating C2CH signature plus four other residues (P, W, F, and P) that are also required for DNA binding. These eight residues are fully conserved (with one exception) in positions C29, C34, P53, W63, C89, H92, F93, and P119 of the multiple alignment (SI Fig. 6). A leucine zipper coiled-coil motif involved in protein dimerization is located after the DNA binding domain (5). We predicted *in silico* a similar 22-aa-long coiled-coil motif after the THAP domain in the *Galileo* and *1360* TPases (SI Fig. 6). Finally, although the *Dmel*/P TPase does not contain the characteristic catalytic motif DD(35)E shared by many other TPases and integrases (4), the C-terminal portion of this protein contains numerous aspartic (D) or glutamic (E) residues and four of them seem to be critical for TPase function: D(83)D(2)E(13)D (see ref. 5). The first 3 aa are fully conserved in positions D677, D774, and E777 of the multiple alignment with one exception (SI Fig. 6), thus supporting this model (5). The conservation of the fourth amino acid is unclear.

A phylogenetic tree was generated with the 21 *Galileo*, *1360*, and *P* TPases and the human THAP9 protein (see Methods). The tree (Fig. 4) shows three clades corresponding to the *Galileo*, *1360*, and *P* elements. Therefore, the three TEs seem monophyletic, although only the *Galileo* and *P* clades have very high statistical support. *Galileo* and *1360* are more closely related to each other than to the *P* element, which is connected to the other two by a deeper branch.

Discussion

We characterized a seemingly complete copy of *Galileo* from the genome of *D. buzzatii* that contains a 2,738-bp ORF encoding a TPase. Three observations indicate that this is the true *Galileo* TPase instead of that of another TE accidentally associated with

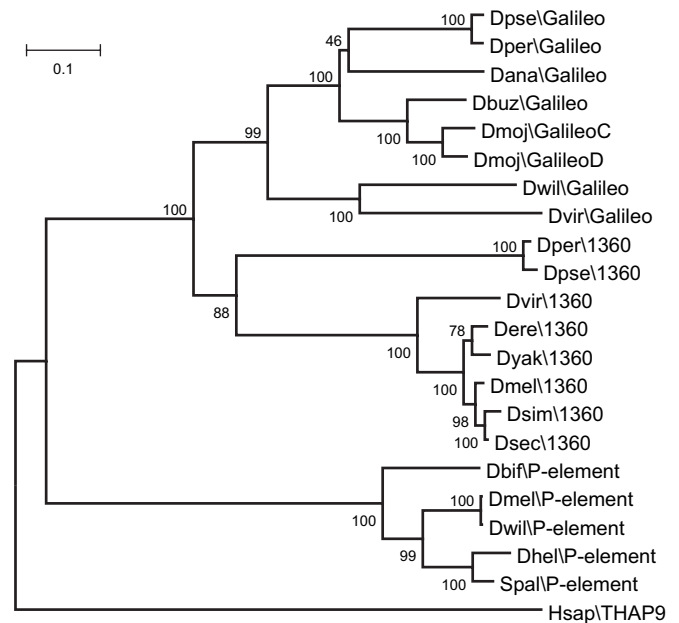


Fig. 4. Neighbor-joining phylogenetic tree constructed with the eight consensus *Galileo* TPases, eight consensus *1360* TPases, and five TPases from representative *P* elements. The human *P*-like THAP9 protein is included as an outgroup. The complete alignment without Gblocks filtering is shown in SI Fig. 6. The tree topology was identical when using maximum likelihood and parsimony methods.

the long *Galileo* TIRs. (i) Two previously isolated *Galileo* copies bear a 141-bp portion of the same ORF in the right position and orientation (21), suggesting that all previously isolated *Galileo* copies are defective versions of the complete structure reported here. (ii) Our bioinformatic searches uncovered TEs structurally similar to *Galileo* in the genomes of six phylogenetically distant *Drosophila* species. These searches were carried out by using as queries the *Dbuz*/*Galileo* and *Dmel*/*1360* TPases, and a careful scrutiny of the contigs producing significant hits led to the finding of the TIRs associated with the TPase segment and the characterization of the elements as either *Galileo* or *1360*. No other TIRs besides those of these two TEs were found flanking the hits (but note that in *Dmoj*/*Galileo* 160-bp internal inverted repeats bound the TPase; Fig. 1B). The persistent association (over tens of myr) of this TPase with the same type of TIRs renders the possibility of an accidental association extremely unlikely. (iii) The presence of multiple *Galileo* copies comprising both TIRs and TPase-coding segments in seven *Drosophila* genomes suggests that these are integral components of the same elements, and these elements are (or have been) able to replicate and transpose within these genomes.

Further evidence leads us to infer that *Galileo*, previously considered a *Foldback*-like element, is in fact a transposon related to the *D. melanogaster* *1360* and *P* elements, and thus, it is probably a TE moving by a cut-and-paste reaction (3, 4). (iv) The *Galileo* TPase is 30–35% and 20–25% identical to those of *1360* and *P* elements, respectively, and the three proteins harbor similar functional domains such as a DNA binding THAP domain, a coiled-coil motif for protein dimerization, and a catalytic domain (5, 27–30). (v) Despite their dramatically different size (several hundred base pairs vs. 31 bp), the *Galileo* terminus includes sequences clearly related to the *1360* and *P* TIRs. Specifically, the consensus *Galileo* terminus shares 15 bp with the *1360* consensus TIR and 17 bp with the *Dmel*/P TIR. The three elements share identical 5'-CA...TG-3' termini. (vi) Both *Galileo* and *1360* generate on insertion 7-bp TSDs that, in

the case of *Galileo*, match the consensus sequence GTANTAC, a palindrome. The TSDs of *DmelP* are 8 bp long and the consensus also corresponds to a palindrome, GTCCGGAC, a fact related to the dimerization of the *P* TPase (5). This suggests that the functional *Galileo* TPase is also a dimer. We conclude that *Galileo* belongs to the *P* superfamily of cut-and-paste transposons.

A parsimonious interpretation of the phylogenetic tree relating *Galileo* with the *1360* and *P* elements (Fig. 4) suggests that *Galileo* arose from an ancestor with much shorter TIRs. *Galileo* long TIRs are variable in size both between and within species, suggesting a remarkable structural dynamism. For instance, in *D. willistoni*, the longest and putatively complete copy (contig 10048) has 765-bp TIRs, but another copy (contig 9452) has 959-bp-long TIRs. Similarly, TIRs of *Galileo* copies in *D. mojavensis* are 458 bp (contig 10940) to 1,260 bp (contig 10757.2) long. TIRs may accidentally shorten (e.g., by deletion) but very likely they may also be elongated by internal duplication, unequal recombination, and/or other mechanisms, such as long-tract gene conversion (33) or single-strand break and synthesis-repair (see figure 5B in ref. 34). We suggest that different *Foldback*-like elements might have originated from independent transposon lineages in a similar manner as the *Drosophila* element *Galileo*. In other words, TIR length and structure is not a reliable criterion for TE classification, and *Foldback*-like elements do not constitute a monophyletic group.

The phylogeny of the *Galileo* elements in the seven *Drosophila* species (Fig. 4) is clearly inconsistent with that of the species (cf. figure 1 in ref. 24). The elements of *D. willistoni* and *D. virilis*, pertaining to different subgenera (*Sophophora* and *Drosophila*, respectively) are each other's closest relative. Similarly, the *Galileo* elements of *D. mojavensis* and *D. buzzatii* (*Drosophila* subgenus) are more closely related to those of *D. ananassae*, *D. pseudoobscura*, and *D. persimilis* (*Sophophora* subgenus) than to those of *D. virilis*, a species from the same subgenus. Equally inconsistent with the species relationships is the phylogeny of the *1360* element (Fig. 4). There are two possible explanations for these topological disparities: lineage sorting and horizontal transfer (35). Lineage sorting refers to the vertical diversification of TE lineages and their differential loss along the branches of the species tree. Horizontal transfer is the process of invasion of a new genome by a TE, which is common for transposons and is considered as an integral phase of the transposon life cycle that allows long-term survival (6, 36). The strongest evidence for horizontal transfer is probably the detection of elements with a high degree of similarity in very divergent taxa, such as in the *P* element colonization of the *D. melanogaster* genome within the last century from the distantly related species *D. willistoni* (37). Many more events of horizontal transfer have occurred during the evolution of *P* elements in the genus *Drosophila* based on the available evidence (38). However, despite their close evolutionary relationship to *P*, the available evidence for horizontal transfer in *Galileo* and *1360* (Fig. 4) is not compelling and lineage sorting should be considered, at this time, as an equally likely explanation.

The origin of the numerous chromosomal inversions in *Drosophila* and other Dipterans is still an open question and very few species have been investigated in this regard. Strong evidence implicating TE-mediated ectopic exchange has been found in four polymorphic inversions only, including the two *D. buzzatii* inversions generated by *Galileo* (39). In *D. melanogaster* and its close relatives, no TEs have been involved in the origin of three polymorphic inversions and only 2 of 29 fixed inversions contain repetitive sequences inverted with respect to each other at both breakpoints, pointing to a completely different mechanism for inversion generation (39). The fact that *Galileo* generated two independent inversions in *D. buzzatii* suggests that *Galileo* is not a passive substrate where ectopic recombination operates but

may be actively generating inversions as a byproduct of its transposition mechanism. If this is correct, to create inversions, *Galileo* has to be active in a genome and a recent transpositional activity would be a necessary condition for *Galileo* to have any role in the generation of current inversions. We have not found any functional TPase in any of these species but only one genome was sequenced in each case, so they could still exist in unsequenced genomic regions, other genomes, and/or other natural populations. However, we have provided evidence of recent (<1 myr) transpositional activity of *Galileo* in *D. ananassae*, *D. persimilis*, *D. pseudoobscura*, and *D. willistoni*. These four are among the most polymorphic species of the genus with 24, 28, 13, and 50 inversions, respectively (40). In *D. mojavensis*, with fewer inversions (41), the most recent transpositional activity of *Galileo* seems somewhat older (≈ 1.5 myr). Finally, *D. virilis* with the oldest *Galileo* activity (≈ 3 myr) is chromosomally monomorphic (40). Therefore, there is a qualitative correlation between the number of inversions and the time of the most recent activity of *Galileo* in this small group of species. This correlation is suggestive but might be only coincidental. However, the detection of chimerical copies that may be the result of chromosomal rearrangements (19) indicates that, indeed, *Galileo* might have been involved in the origin of inversions, at least in some other species besides *D. buzzatii*.

Methods

PCR Amplification and DNA Sequencing. Genomic DNA from *D. buzzatii* (strain st-1) and *D. mojavensis* (strain 15081-1352.22, Tucson *Drosophila* Stock Center) (as control) was used as template for PCR amplification of *Galileo* copies. Primers located in the TIRs were designed based on *D. buzzatii* known incomplete copies of *Galileo* (21), whereas primers inside the TPase were designed on the *D. mojavensis* putative complete TPases found in a preliminary bioinformatic search (SI Fig. 7). Primers in the TIRs were always used in combination with primers anchored in the TPase to avoid multiple bands generated by the highly repetitive primer alone or the amplification of defective copies without TPase. PCRs were carried out in a total volume of 25 μ l including 100–200 ng of genomic DNA, 20 pmol of each primer, 200 μ M dNTPs, 1.5 mM MgCl₂, and 1–1.5 units of Taq DNA polymerase. PCR products were gel-purified by using QIAquick Gel Extraction kit (Qiagen) and sequenced directly with the amplification primers and sequencing primers designed over the end sequences to close gaps (SI Fig. 7). Sequences were aligned and assembled by using multialign software MUSCLE 3.6 (42).

Bioinformatic Searches. BLAST searches were performed on the chromosome assemblies of *D. melanogaster* and *D. simulans* and the contig CAF1 assemblies of the other ten publicly available *Drosophila* genomes (<http://rana.lbl.gov/drosophila>). We used BLAST algorithm version 2.2.2 (43) implemented in the *Drosophila* Polymorphism Database server (<http://bioinformatica.uab.es/dpdb>) with default parameters. TBLASTN searches in the different species were performed by using as queries the TPases of *DbuzGalileo* and *Dmel1360* (SI Table 1). Hits with an E-value $\leq 10^{-20}$ (which in the conditions of our searches amounts roughly to $\approx 30\%$ identity over a stretch of 200 aa) were considered significant. BLASTN searches were also carried out with the 40 terminal bp of *DbuzGalileo* and the 31 bp of the *Dmel1360* TIR (SI Table 1). The cutoff in this case was an E-value $\leq 10^{-3}$ (that requires ≈ 21 –22 consecutive identical base pairs).

Contigs producing significant hits with the *DbuzGalileo* and *Dmel1360* TPases in each species were scrutinized to characterize the different copies of both TEs. TIRs and TSDs were searched around the putative TPases by using Dotlet 1.5 (44) to define the boundaries of each copy. Insertions of other TEs inside *Galileo* were identified by aligning the different *Galileo* copies found in the same species and further analyzing the sequences present in only one of them. Significant contigs <1 kb long and those that were found to contain complex clusters of several TE insertions (likely of heterochromatic origin) were not further investigated.

Nonautonomous Copies. BLASTN searches were carried out with the longest copies of *Galileo* and *1360* (Fig. 1 B and C) to estimate the abundance of the two TEs within each species (SI Table 1). Significant hits were those with E-value $\leq 10^{-20}$ (equivalent to $\approx 80\%$ identity over a stretch of 200 bp). The number of significant contigs in these searches provides usually a minimum estimate for the number of TE copies because the searched databases were the

CAF1 contig assemblies in most cases and each contig contains at least one copy but may actually contain two or more. For similarity analyses, only the TIRs were used as they produced the most reliable alignments. The two TIRs of each TE copy were analyzed separately to estimate the divergence between the two TIRs within each copy as well as the pairwise divergence between copies.

Consensus Sequences. The consensus sequences for *Galileo* and 1360 TPases and *Galileo* TIRs were generated by using BioEdit 7.0.5 (45) after aligning the respective nucleotide sequences (SI Table 9) with MUSCLE 3.6 software (42). In the case of TPases, this consensus sequence was then translated into protein to allow the comparison among different species (SI Fig. 6). Conserved protein domains were detected by using InterProScan (46) and Conserved Domain Search (47). Coiled-coil regions were predicted by using the Coils server (48).

Phylogenetic Analyses. TPase sequences were aligned with MUSCLE 3.6 (42) and the alignment was filtered with Gblocks version 0.91b (49) to remove the poorly aligned and highly divergent segments. Gblocks was used with the default parameters except for the maximum number of contiguous nonconserved positions = 15, the minimum length of a block = 6, and allowed gap position = half. These parameters were fixed so that the conserved THAP domain was included in the filtered alignment. All phylogenetic trees were constructed with MEGA 3.1 (50) by using the neighbor-joining method with

complete deletion and 500 replicates to generate bootstrap values. Poisson correction and Kimura 2 parameters were used as substitution models for amino acid and nucleotide sequences, respectively. We dated the most recent transposition events within each species by dividing the average pairwise divergence between the elements in the same group or subgroup by the *Drosophila* synonymous substitution rate, 0.016 substitutions per nucleotide/myr (21). To date the divergence between different groups or subfamilies we calibrated the tree with the same substitution rate by using the appropriate option in MEGA (50). Time estimates for TEs should be taken with caution; if the synonymous substitution rate were an underestimate of the true mutation rate for TEs, our time estimates would provide an upper bound for the true values.

ACKNOWLEDGMENTS. We thank Margaret Kidwell, Cedric Feschotte, Dmitri Petrov, Mario Cáceres, Josefa González, and two anonymous referees for many constructive comments and Diana Garzón for help with the initial bioinformatic searches in *D. mojavensis*. This work was completed while A.R. was on sabbatical leave at Stanford University; he thanks Dmitri Petrov, Josefa González, James Cai, Yael Salzman, Ruth Hershberg, and Mike Macpherson for their warm hospitality and personal help. This work was supported by a Formación de Personal Investigador doctoral fellowship (to M.M.) and Secretaría de Estado de Universidades e Investigación (Ministerio de Educación y Ciencia, Spain) Grant BFU2005-022379 and mobility grant PR2006-0329 (to A.R.).

- Kidwell MG, Lisch DR (2002) in *Mobile DNA II*, eds Craig NL, Craigie R, Gellert M, Lambowitz AM (American Society for Microbiology, Washington, DC), pp 59–89.
- Capy P, Bazin C, Higuert D, Langin T (1998) *Dynamics and Evolution of Transposable Elements* (Springer, Heidelberg).
- Feschotte C, Pritham EJ (2007) DNA transposons and the evolution of eukaryotic genomes. *Annu Rev Genet* 41:331–368.
- Haren L, Ton-Hoang B, Chandler M (1999) Integrating DNA: transposases and retroviral integrases. *Annu Rev Microbiol* 53:245–281.
- Rio DC (2002) in *Mobile DNA II*, eds Craig NL, Craigie R, Gellert M, Lambowitz AM (American Society for Microbiology, Washington, DC), pp 484–518.
- Hartl DL, Lohse AR, Lozovskaya ER (1997) Modern thoughts on an ancient mariner: Function, evolution, regulation. *Annu Rev Genet* 31:337–358.
- Potter S, Truett M, Phillips M, Maher A (1980) Eucaryotic transposable genetic elements with inverted terminal repeats. *Cell* 20:639–647.
- Truett MA, Jones RS, Potter SS (1981) Unusual structure of the FB family of transposable elements in *Drosophila*. *Cell* 24:753–763.
- Liebermann D, et al. (1983) An unusual transposon with long terminal inverted repeats in the sea urchin *Strongylocentrotus purpuratus*. *Nature* 306:342–347.
- Rebatchouk D, Narita JO (1997) Foldback transposable elements in plants. *Plant Mol Biol* 34:831–835.
- Ade J, Belzile FJ (1999) Hairpin elements, the first family of foldback transposons (FTs) in *Arabidopsis thaliana*. *Plant J* 19:591–597.
- Simmen MW, Bird A (2000) Sequence analysis of transposable elements in the sea squirt, *Ciona intestinalis*. *Mol Biol Evol* 17:1685–1694.
- Windsor AJ, Waddell CS (2000) FARE, a new family of foldback transposons in *Arabidopsis*. *Genetics* 156:1983–1995.
- Templeton NS, Potter SS (1989) Complete foldback transposable elements encode a novel protein found in *Drosophila melanogaster*. *EMBO J* 8:1887–1894.
- Harden N, Ashburner M (1990) Characterization of the FB-NOF transposable element of *Drosophila melanogaster*. *Genetics* 126:387–400.
- Pritham EJ, Feschotte C, Wessler SR (2005) Unexpected diversity and differential success of DNA transposons in four species of entamoeba protozoans. *Mol Biol Evol* 22:1751–1763.
- Cáceres M, Ranz JM, Barbadilla A, Long M, Ruiz A (1999) Generation of a widespread *Drosophila* inversion by a transposable element. *Science* 285:415–418.
- Cáceres M, Puig M, Ruiz A (2001) Molecular characterization of two natural hotspots in the *Drosophila buzzatii* genome induced by transposon insertions. *Genome Res* 11:1353–1364.
- Casals F, Cáceres M, Ruiz A (2003) The foldback-like transposon *Galileo* is involved in the generation of two different natural chromosomal inversions of *Drosophila buzzatii*. *Mol Biol Evol* 20:674–685.
- Lim JK, Simmons MJ (1994) Gross chromosome rearrangements mediated by transposable elements in *Drosophila melanogaster*. *BioEssays* 16:269–275.
- Casals F, Cáceres M, Manfrin MH, Gonzalez J, Ruiz A (2005) Molecular characterization and chromosomal distribution of *Galileo*, Kepler and Newton, three foldback transposable elements of the *Drosophila buzzatii* species complex. *Genetics* 169:2047–2059.
- Adams MD, et al. (2000) The genome sequence of *Drosophila melanogaster*. *Science* 287:2185–2195.
- Richards S, et al. (2005) Comparative genome sequencing of *Drosophila pseudoobscura*: Chromosomal, gene, and cis-element evolution. *Genome Res* 15:1–18.
- Drosophila* 12 Genomes Consortium (2007) Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* 448:203–218.
- Reiss D, Quesneville H, Nouaud D, Andrieu O, Anxolabehere D (2003) Hoppel, a P-like element without introns: a P-element ancestral structure or a retrotranscription derivative? *Mol Biol Evol* 20:869–879.
- Laski FA, Rio DC, Rubin GM (1986) Tissue specificity of *Drosophila* P element transposition is regulated at the level of mRNA splicing. *Cell* 44:7–19.
- Lee CC, Beall EL, Rio DC (1998) DNA binding by the KP repressor protein inhibits P-element transposase activity in vitro. *EMBO J* 17:4166–4174.
- Clouaire T, et al. (2005) The THAP domain of THAP1 is a large C2CH module with zinc-dependent sequence-specific DNA-binding activity. *Proc Natl Acad Sci USA* 102:6907–6912.
- Roussigne M, et al. (2003) The THAP domain: A novel protein motif with similarity to the DNA-binding domain of P element transposase. *Trends Biochem Sci* 28:66–69.
- Quesneville H, Nouaud D, Anxolabehere D (2005) Recurrent recruitment of the THAP DNA-binding domain and molecular domestication of the P-transposable element. *Mol Biol Evol* 22:741–746.
- Kapitonov VV, Jurka J (2003) Molecular paleontology of transposable elements in the *Drosophila melanogaster* genome. *Proc Natl Acad Sci USA* 100:6569–6574.
- Hagemann S, Pinsky W (2001) *Drosophila* P transposons in the human genome? *Mol Biol Evol* 18:1979–1982.
- Richardson C, Moynahan ME, Jasin M (1998) Double-strand break repair by interchromosomal recombination: suppression of chromosomal translocations. *Genes Dev* 12:3831–3842.
- Kapitonov VV, Jurka J (2006) Self-synthesizing DNA transposons in eukaryotes. *Proc Natl Acad Sci USA* 103:4540–4545.
- Page RD, Charleston MA (1998) Trees within trees: Phylogeny and historical associations. *Trends Ecol Evol* 13:356–359.
- Silva JC, Loreto EL, Clark JB (2004) Factors that affect the horizontal transfer of transposable elements. *Curr Issues Mol Biol* 6:57–71.
- Clark JB, Kidwell MG (1997) A phylogenetic perspective on P transposable element evolution in *Drosophila*. *Proc Natl Acad Sci USA* 94:11428–11433.
- Silva JC, Kidwell MG (2000) Horizontal transfer and selection in the evolution of P elements. *Mol Biol Evol* 17:1542–1557.
- Ranz JM, et al. (2007) Principles of genome evolution in the *Drosophila melanogaster* species group. *PLoS Biol* 5:e152.
- Sperlich D, Pfrim P (1986) in *The Genetics and Biology of Drosophila*, eds Ashburner M, Carson HL, Thompson JNJ (Academic, London), pp 257–309.
- Ruiz A, Heed WB, Wasserman M (1990) Evolution of the *mojavensis* cluster of cactophilic *Drosophila* with descriptions of two new species. *J Hered* 81:30–42.
- Edgar RC (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792–1797.
- Altschul SF, et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402.
- Junier T, Pagni M (2000) Dotlet: Diagonal plots in a web browser. *Bioinformatics* 16:178–179.
- Hall TA (1999) BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp Ser* 41:95–98.
- Zdobnov EM, Apweiler R (2001) InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* 17:847–848.
- Marchler-Bauer A, et al. (2005) CDD: A Conserved Domain Database for protein classification. *Nucleic Acids Res* 33:D192–D196.
- Lupas A, Van Dyke M, Stock J (1991) Predicting coiled coils from protein sequences. *Science* 252:1162–1164.
- Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* 17:540–552.
- Kumar S, Tamura K, Nei M (2004) MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment. *Brief Bioinform* 5:150–163.

APPENDIX

IV

The Transposon *Galileo* Generates Natural Chromosomal Inversions in *Drosophila* by Ectopic Recombination

Alejandra Delprat, Bàrbara Negre[✉], Marta Puig[✉], Alfredo Ruiz*

Departament de Genètica i de Microbiologia, Universitat Autònoma de Barcelona, Bellaterra (Barcelona), Spain

Abstract

Background: Transposable elements (TEs) are responsible for the generation of chromosomal inversions in several groups of organisms. However, in *Drosophila* and other Diptera, where inversions are abundant both as intraspecific polymorphisms and interspecific fixed differences, the evidence for a role of TEs is scarce. Previous work revealed that the transposon *Galileo* was involved in the generation of two polymorphic inversions of *Drosophila buzzatii*.

Methodology/Principal Findings: To assess the impact of TEs in *Drosophila* chromosomal evolution and shed light on the mechanism involved, we isolated and sequenced the two breakpoints of another widespread polymorphic inversion from *D. buzzatii*, *Zz*³. In the non inverted chromosome, the *Zz*³ distal breakpoint was located between genes *CG2046* and *CG10326* whereas the proximal breakpoint lies between two novel genes that we have named *Dlh* and *Mdp*. In the inverted chromosome, the analysis of the breakpoint sequences revealed relatively large insertions (2,870-bp and 4,786-bp long) including two copies of the transposon *Galileo* (subfamily *Newton*), one at each breakpoint, plus several other TEs. The two *Galileo* copies: (i) are inserted in opposite orientation; (ii) present exchanged target site duplications; and (iii) are both chimeric.

Conclusions/Significance: Our observations provide the best evidence gathered so far for the role of TEs in the generation of *Drosophila* inversions. In addition, they show unequivocally that ectopic recombination is the causative mechanism. The fact that the three polymorphic *D. buzzatii* inversions investigated so far were generated by the same transposon family is remarkable and is conceivably due to *Galileo*'s unusual structure and current (or recent) transpositional activity.

Citation: Delprat A, Negre B, Puig M, Ruiz A (2009) The Transposon *Galileo* Generates Natural Chromosomal Inversions in *Drosophila* by Ectopic Recombination. PLoS ONE 4(11): e7883. doi:10.1371/journal.pone.0007883

Editor: Robert DeSalle, American Museum of Natural History, United States of America

Received: June 18, 2009; **Accepted:** October 1, 2009; **Published:** November 18, 2009

Copyright: © 2009 Delprat et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported by grant BFU2005-02237 from the Ministerio de Educación y Ciencia (MEC, Spain) and grant BFU2008-04988 from the Ministerio de Ciencia e Innovación (MICINN, Spain) awarded to A.R. and by a post-doctoral fellowship from the Fundación Carolina (Spain) awarded to A.D. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: Alfredo.Ruiz@uab.es

✉ Current address: Department of Zoology, University of Cambridge, Cambridge, United Kingdom

✉ These authors contributed equally to this work.

Introduction

A sizable portion of eukaryotic and prokaryotic genomes is composed of transposable elements (TEs) with the potential to cause chromosomal rearrangements such as inversions, translocations and duplications [1–3]. These rearrangements however may be generated also by other processes that do not involve TEs (see below). Thus, the actual contribution of TEs to the evolutionary reorganization of genomes is unclear. One of the most frequent and widespread types of chromosomal rearrangements during evolution are inversions, which alter gene order often without changing total gene content [4]. Inversions are remarkably abundant in the genus *Drosophila*, both as intraspecific polymorphisms and as interspecific fixed differences [5,6] and increasing evidence point to their prevalence in many other species, e.g. humans [7–11].

TEs can generate chromosomal inversions by intrachromosomal homologous recombination between two copies of the same TE family arranged in opposite orientation [12]. This mechanism is

known as TE-mediated ectopic recombination or nonallelic homologous recombination (NAHR). TEs can also induce inversions as well as other types of rearrangements when two ends coming from different TE copies participate together in an aberrant transposition event. The outcome depends on the location and orientation of the two cooperating TE copies in the parental chromosome and the chromosomal site where they insert (Figure S1). If the two copies are located in sister chromatids or homologous chromosomes, the process is referred to as hybrid element insertion [13–15]. When the two copies are located at neighboring sites on the same chromatid, the mechanism is known as reversed ends transposition [16,17]. Inversions can be also generated by two other mechanisms not involving TEs. One such mechanism is chromosomal breakage and repair by non-homologous end-joining (NHEJ). Double strand breaks (DSBs) are produced in many ways in all cells and the machinery to deal with these lesions is conserved from yeasts to vertebrates [18,19]. When two or more DSBs occur simultaneously, repair by NHEJ may produce gross rearrangements if the joining takes place

between previously unlinked DNA molecules [20]. Finally, inversions may result from ectopic recombination between other repeated sequences besides TEs, such as tRNA genes [21] or segmental duplications (SDs) [7,8].

TE-mediated ectopic recombination has generated natural chromosomal inversions in bacteria [22–27] and some lineages have experienced an striking degree of rearrangement caused by this process [28–31]. Likewise, $T\gamma$ -recombination mediated deletions, duplications, inversions and translocations have been found to occur in yeast [12,32–34]. In mammals, long and short interspersed elements (LINEs and SINEs, respectively) have been implicated in the generation by ectopic recombination of 50 inversions fixed between humans and chimpanzees [35,36]. In *Drosophila*, the evidence for the implication of TEs in the generation of inversions is limited. Two *D. buzzatii* polymorphic inversions, $2j$ and $2q^7$, were seemingly generated by ectopic recombination between copies of the transposon *Galileo* [37,38]. In *D. pseudoobscura*, the polymorphic inversion Arrowhead and a number of fixed inversions have been also generated by ectopic recombination between 128-bp and 315-bp repeats, yet the nature of these repeats is obscure [39]. Inversion *In(4)a* of *D. americana* has been found to be flanked by copies of a new transposon and was likely generated by an intrachromosomal exchange between these repeats [40]. TEs have been found also at the breakpoints of two *Anopheles gambiae* inversions, $2Rd'$ and $2La$, but the implication of these TEs in the origin of the inversions is circumstantial [41,42].

Chromosomal breakage and repair by NHEJ is also a common mechanism for the generation of chromosomal inversions. This process may generate duplications flanking the inverted segment when one or both DSBs occur in a staggered manner [43]. In *Drosophila*, this process has been responsible for most of the inversions fixed between *D. melanogaster* and *D. yakuba* [43] as well as three *D. melanogaster* polymorphic inversions [44–46]. In addition, this mechanism likely generated several inversions fixed in other lineages where TEs were not detected at the breakpoints or when present were not involved in the origin of the inversion [47–50]. SDs represent a significant fraction of mammalian genomes and ectopic recombination between SDs seems to be a common mechanism inducing chromosomal inversions in these genomes. Six of the nine large pericentric inversion differences between the human and chimpanzee genomes have been associated with SDs [51] and there is a significant SD enrichment at the sites of breakpoints which occurred during primate evolution [52–58] although it is not clear whether ectopic recombination is always the cause for the co-location of SDs and breakpoints. Ectopic recombination between SDs is also responsible for the generation of chromosomal inversions in other groups, e.g insects [59].

The transposon *Galileo* was discovered in *D. buzzatii* and tentatively classified (along with two related elements named *Newton* and *Kepler*) as a *Foldback*-like element because of its long, internally repetitive, terminal inverted repeats (TIRs) and lack of coding capacity [60,61]. We have recently shown that *Galileo* is a cut-and-paste transposon belonging to the *P* superfamily that is present in six of the 12 recently sequenced *Drosophila* genomes [62]. *Galileo*, *Newton* and *Kepler* show a high degree of nucleotide similarity (including the most terminal 40 bp that are almost identical) and produce 7-bp target site duplications (TSDs) with the same consensus sequence, GTAGTAC, which suggests that they are mobilized by the same transposase [61]. They should be considered only as different subfamilies of *Galileo* in the genome of *D. buzzatii* and will be denoted hereafter as *GalileoG*, *GalileoN* and *GalileoK*, respectively.

In order to increase our understanding of the mechanisms underlying the generation of *Drosophila* inversions in nature and test for an implication of transposable elements, here we isolated and characterized the breakpoints of another *D. buzzatii* polymorphic inversion, $2z^3$. This inversion arose on a chromosome carrying the $2j$ inversion, giving rise to arrangement $2jz^3$. The $2z^3$ segment encompasses about one third of chromosome 2 (~11 Mb) and overlaps the $2j$ segment so that the two inversions can not be separated by recombination [63]. Thus, three chromosome 2 arrangements are commonly found in *D. buzzatii* natural populations, 2 standard ($2st$), $2j$ and $2jz^3$. Arrangement $2jz^3$ has a wide geographical distribution being present in natural populations of Argentina, Southern Brazil, Chile and the Old World [64,65]. In 18 Argentina populations where arrangement $2jz^3$ is present, its relative frequencies range from 0.5 to 31.5% with an average of ~8% [65]. We choose to study this inversion in part because its proximal breakpoint was located at chromosomal band 2F1c [63] very near the site (2F1c-e) where the *proboscipedia-Ultrabithorax* portion of the Hox gene complex has been localized [66,67]. We seek to determine the precise distance from the inversion breakpoint to the Hox genes and find out whether these genes were affected in any way by the inversion. The results show that copies of the transposon *GalileoN* are located at both inversion $2z^3$ breakpoints. The arrangement of TSDs and the chimeric nature of both *GalileoN* copies provide unequivocal evidence that this transposon generated inversion $2z^3$ by ectopic recombination. The $2z^3$ proximal breakpoint lies ~24 kb downstream of the *proboscipedia* gene in a poorly annotated region where two novel genes, *Dlh* and *Mdp*, have been discovered.

Results

Physical Mapping of the $2z^3$ Inversion Breakpoints in the *D. buzzatii* Genome

Previous cytological observations in *D. buzzatii* located the distal and proximal breakpoints of inversion $2z^3$ near chromosome 2 bands 2E4c and 2F1c, respectively [63,68]. We used the BAC-based physical map of the *D. buzzatii* genome [69] and the available genome sequence of the related species *D. mojavensis* [70] to pinpoint the $2z^3$ distal breakpoint in the intergenic region between *CG2046* and *CG10326* (see Figure 1 left and Materials and Methods for details). A detailed physical map of the *D. buzzatii* chromosomal region encompassing the $2z^3$ proximal breakpoint had been constructed in a previous study [67] and one of the four BAC clones bearing the breakpoint (BAC 40C11) was already fully sequenced and annotated. We mapped the proximal breakpoint within the gene *lodestar* (*lds*) that had been tentatively annotated in that region of BAC 40C11 (see Figure 1 right and Materials and Methods). This annotation was put into question by the subsequent annotation of the *D. mojavensis* genome [70] and a close scrutiny of the region (see below) revealed the presence of two novel genes that we have named *Dlh* and *Mdp*. The $2z^3$ proximal breakpoint falls in the intergenic space between them.

Breakpoint Sequences in the Non-Inverted Chromosomes

Following previous sequence analyses of inversion breakpoints [37,44], the distal and proximal breakpoint regions of $2z^3$ were designated as AB and CD in the non-inverted chromosomes ($2st$ or $2j$) and as AC and BD in the inverted chromosome ($2jz^3$). Using primers designed in the *D. mojavensis* genome, we amplified and sequenced 1,022 bp of the distal breakpoint region (AB) between genes *CG2046* and *CG10326* in three $2st$ lines and five $2j$ lines from diverse geographic origins. In line st-1, the AB sequence comprises

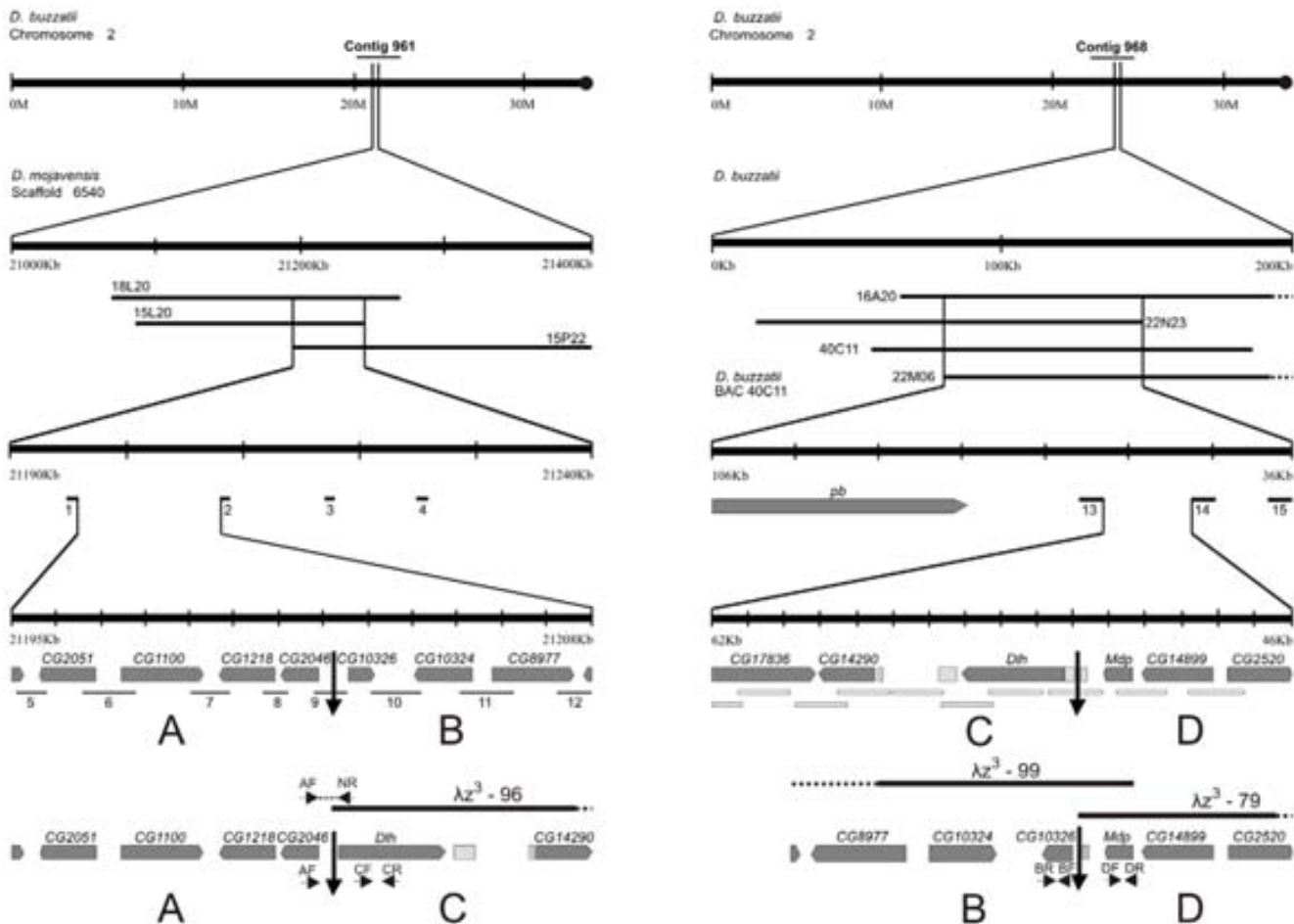


Figure 1. Experimental strategy used for mapping the distal (left) and proximal (right) breakpoints of inversion $2z^3$. The segments depicted in each column are ordered from top to bottom in four successive steps with increasing mapping resolution. The distance between consecutive bars stands for 10 Mb, 100 kb, 10 kb and 1 kb, in the four steps, respectively. Line 1: Relative position of the contigs on the physical map of *D. buzzatii* standard chromosome 2. Line 2: Relative position of the BAC clones encompassing the distal breakpoint (left) and the proximal breakpoint (right). Line 3: Position of the PCR probes used to pinpoint the breakpoints within the overlapping segment of BAC clones. Line 4: Genes located in the breakpoint regions of the non-inverted chromosome (designated as AB and CD) are represented by dark grey rectangles with a pointed end indicating the direction of transcription and TEs by light grey rectangles. Short numbered segments under the genes in the distal breakpoint region (left) correspond to plasmid subclones of BAC 40C11. Line 5: Genes located in the breakpoint region of the inverted chromosome (designated as AC and BD). Thick lines above the inverted chromosome represent the lambda clones isolated during the cloning of the $2z^3$ breakpoints. Small horizontal arrows represent PCR primers (e.g. AF, NR, ...). Vertical arrows mark the location of the breakpoints. Note that there is a reversal of orientation between lines 1 and 2 in the distal breakpoint (left). The reason is inversion $2z^3$ took place in a $2j$ chromosome and not in the standard chromosome 2 represented in line 1. See Materials and Methods for details.
doi:10.1371/journal.pone.0007883.g001

281 bp of gene *CG2046*, 163 bp of gene *CG10326* and the 578-bp intergenic region (Figure 2) including an (AT)₂₃ microsatellite (272 bp away from the start codon of *CG2046*). No structural variation was found in the AB region between the eight non-inverted lines except for the number of repeats in the microsatellite (between 16 and 24).

The proximal breakpoint (CD) was localized in the *Dlh* - *Mdp* intergenic region (Figure 2). In line st-1, the intergenic region between these two genes is 1,102-bp long and includes two TE fragments: a 296-bp fragment of *GalileoN* (element *Galileo*, subfamily *Newton*), and a 202-bp fragment of *BuT5* (an unclassified *D. buzzatii* transposon [60]). The CD region was amplified by PCR and sequenced in seven non-inverted lines besides st-1. The CD sequence (1,771 bp) includes 238 bp of gene *Dlh* and 337 bp of gene *Mdp*. All seven lines contained the *BuT5* fragment but only one (j-19) contained the *GalileoN* fragment.

Levels of nucleotide variation in the $2z^3$ breakpoint regions were estimated from the AB and CD sequences of the eight lines without the inversion (Figure 3 and Table S1). Overall, 2,422 bp were analyzed comprising 719 bp of coding sequence, 1,501 bp of non-coding sequence (introns and intergenic segments) and 202 bp of the *BuT5* insertion. Coding and non-coding sequences were analyzed separately. Both the (AT)₁₆₋₂₄ microsatellite and the polymorphic *GalileoN* insertion were excluded from the analysis. Besides this *GalileoN* insertion, one small insertion of 4 bp and 9 deletions (ranging in size from 1 to 64 nucleotides) were observed in the set of eight lines. Non-coding sequences contain 33 segregating sites (10 in AB and 23 in CD), coding sequences 12 and the *BuT5* insertion six (Figure 3). Nucleotide diversity [71] values in the different regions are given in Table S1 and a neighbour-joining phylogenetic tree built with the non-coding sequences of the single-copy breakpoint regions (ABCD) is shown in Figure 4.

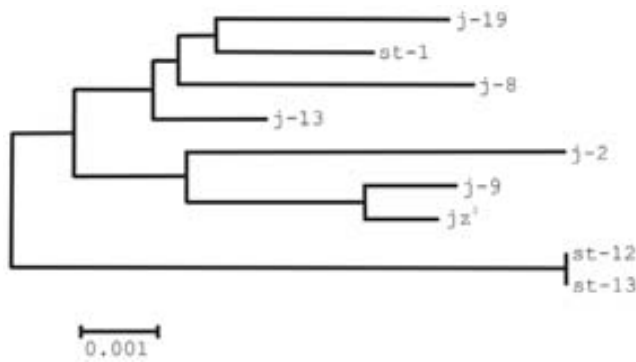


Figure 4. Neighbour-joining phylogenetic tree of the ABCD breakpoint sequences excluding the (AT)₁₆₋₂₄ microsatellite and the TE insertions.
doi:10.1371/journal.pone.0007883.g004

do not span the complete distal breakpoint region (AC) and also that they contain repetitive sequences. Clone λz³-96 was selected for subcloning because its insert reached further away in direction to the breakpoint, and subclones containing the fragments located closest to the breakpoint were sequenced (Figure 1). This provided the sequence of region C and also repetitive sequences inserted at the breakpoint junction but not region A. The rest of the AC region was isolated by PCR using two primers, NR located at the end of the λz³-96 clone and AF anchored in gene *CG2046* from region A (Figure 1). The resulting PCR product was sequenced (1,072 bp) and assembled together with the insert of clone λz³-96 to complete the sequence of the distal breakpoint AC.

Three positive clones were isolated with probe D. These clones produced an intense signal at the proximal breakpoint when hybridized to 2j chromosomes but also weak additional signals in multiple locations. This indicates that these clones bear sequences from region D but do not span the complete proximal breakpoint region (BD) and also that they contain repetitive sequences. Clone λz³-79 containing the longest insert was subcloned and subclones of interest sequenced confirming that it did not contain sequences from region B (Figure 1). Moreover, this time the remaining part of the sequence could not be amplified by PCR, so we screened the two 2z³ lambda libraries with a probe from region B. Three additional lambda clones were isolated and tested by PCR for the presence of the genes at each side of the breakpoint. Clone λz³-99 was chosen as it contained the genes *CG10326* and *Mdp*, located in regions B and D, respectively. Southern blot analysis revealed that in λz³-99 clone these markers were separated by ~5 kb, therefore it was completely sequenced and the sequence of the proximal breakpoint (BD) was determined.

In total, we sequenced 4,067 bp and 12,520 bp from the distal (AC) and proximal (BD) breakpoint regions in a chromosome with the 2z³ inversion. Comparison of these sequences with the breakpoint regions in non-inverted chromosomes (AB and CD) allowed us to locate the precise site of the breakpoint junctions within the intergenic regions (Figure 2). This comparison also revealed that there are no fixed nucleotide substitutions between inverted and non-inverted chromosomes (Figure 3). In the phylogenetic tree the 2jz³ chromosome does not form a separate lineage and appears to be closest to the j-9 line (Figure 4), with which it shares the premature stop codon in *Dlh* exon 2 (Figure 3). Relatively large insertions were found at the AC (2,870 bp) and BD (4,786 bp) junctions that were not present in non-inverted chromosomes (Figure 2). These insertions are composed of several TE insertions, most of them similar to elements previously

characterized in *D. buzzatii* [38,60]. The detailed TE content of the breakpoint insertions is summarized in Table 1.

The 2,870-bp insertion in the AC junction comprises a copy of *GalileoN* (*GalileoN-4*) with two nested insertions: a copy of *BuT5* (*BuT5-7*) flanked by 8-bp TSDs and a 261-bp copy of a *LINE*-like element (Figure 2). The latter copy has no apparent ORF and no significant sequence homology with described elements. We have classified this insertion as a partial *LINE*-like element because it shows a 41-bp long polyA tail and two flanking 13-bp TSDs. The 4,786-bp insertion in the BD junction comprises also a copy of *GalileoN* (*GalileoN-5*) with two other nested TE insertions (Figure 2): a copy of *BuT4* (*BuT4-3*) flanked by 8-bp TSDs and a copy of *BuT3* (*BuT3-7*) flanked also by 8-bp TSDs. *BuT4* was previously classified as a Class II element of the *hAT* superfamily [60]. This is corroborated by the 87% nucleotide identity observed between this copy of *BuT4* and *Homo7*, a *hAT* element recently described in *D. mojavensis* [73]. This copy of *BuT4* includes a 1774-bp segment with a 87.7% identity to *Homo7* transposase-encoding ORF.

The two *GalileoN* copies inserted at the breakpoint junctions (*GalileoN-4* and *GalileoN-5*) have relatively long TIRs (Table 1) and are very similar to copies of the subfamily *Newton* previously described in *D. buzzatii* [60]. Upon insertion, *Galileo* generates 7-bp TSDs with the consensus sequence GTAGTAC [61,62]. The 7-bp sequence flanking *GalileoN-4* in region C (GTAGTAC) is the reverse and complementary version of the 7-bp sequence flanking *GalileoN-5* in region D (GTACTAC). Likewise, the 7-bp sequence flanking *GalileoN-4* in region A (GTACTAT) is the inverted and complementary version of that flanking *GalileoN-5* in region B (ATAGTAC). Only one single copy of the 7-bp sequence GTACTAT is present at the distal breakpoint (AB) and one copy of the target sequence GTACTAC is found at the proximal breakpoint (CD) in the non-inverted chromosomes. This pattern of exchanged TSDs is consistent with ectopic recombination as the mechanism that generated the 2z³ inversion (see Discussion).

Two Novel Drosophila Genes

The proximal breakpoint of inversion 2z³ was located within BAC 40C11 in the genomic region between genes *CG14899* and *CG14290*. This *D. buzzatii* chromosome 2 region had been tentatively annotated as containing a single five-exon gene orthologous to *D. melanogaster lds* [67]. However, only three of the five exons of the *D. buzzatii* gene model showed significant homology with *Dmel\lds*. We failed to corroborate the structure of the putative *D. buzzatii lds* gene by RT-PCR using primers anchored in exons 1 and 5. In addition, the sequencing and

Table 1. Transposable elements found at the breakpoint regions of inversion 2z³ in *D. buzzatii*.

Breakpoint Region	Family-copy	Size (bp)	TIR (bp)	TSD (bp)
AC	<i>GalileoN-4</i>	1541	575/610	7
AC	<i>BuT5-7</i>	1039	3/3	8
AC	<i>LINE-like</i>	261	-	13
BD	<i>GalileoN-5</i>	1533	606/580	7
BD	<i>BuT4-3</i>	2441	23/24	8
BD	<i>BuT3-7</i>	795	23/23	8
C	<i>GalileoN-6</i>	296	10/10	7
D	<i>BuT5-8</i>	202	3	-

doi:10.1371/journal.pone.0007883.t001

annotation of 12 *Drosophila* genomes [70] revealed that in *D. mojavensis*, the closest species to *D. buzzatii*, the *lds* ortholog is located in a distant chromosome 2 region casting doubts on the *D. buzzatii* annotation. These observations prompted a detailed comparative analysis of the 7.5-kb *D. buzzatii* region between genes *CG14899* and *CG14290* with the homologous regions in *D. mojavensis* and *D. virilis* and a search for RNA expression by RT-PCR (see Materials and Methods).

The results lead us to discard the *lds* annotation and discover two novel *Drosophila* genes, whose main characteristics are described in Table S2. In *D. buzzatii*, the gene that we have named *MADF domain protein (Mdp)* is composed of three exons and two introns with a total length of 794 bp (Table S2). The coding sequence is 651-bp long and encodes a 216-aa protein with a MADF domain (Figure S2). *Mdp* has been found also in *D. mojavensis* and *D. virilis* with a similar structure, although a somewhat longer coding sequence in *D. virilis* and a stop codon in position 142 of the third exon in *D. mojavensis*. As expected from the phylogenetic relationships, nucleotide identity and amino acid identity were higher with *D. mojavensis* (82.5% and 76.3%, respectively) than with *D. virilis* (70.4% and 60.9%, respectively). The overall codon-based Z-test of purifying selection shows highly significant results ($Z = -10.15$, $P < 10^{-6}$) and the ratio of synonymous to non-synonymous substitutions ($Ka/Ks = 0.22$) shows a moderate degree of functional constraint. The second gene has been named *DEAD-like helicase (Dlh)* and in *D. buzzatii* it comprises four exons and three introns with a total length of 2,826 bp. The coding sequence is 1,554-bp long and encodes a 517-aa protein with a SNF2-related or DEAD-like helicase N-terminal domain and a DNA/RNA helicase C-terminal domain (Figure S3). This gene is also present in *D. mojavensis* with a similar structure, but could not be found in *D. virilis* (Table S2). Nucleotide identity of the coding sequence (76.8%) and amino acid identity of the protein (64.5%) support orthology. The estimated ratio Ka/Ks was relatively high (0.48), but significantly lower than 1 ($Z = -5.56$, $P = 2 \times 10^{-7}$) suggesting that this is a relatively fast evolving gene.

Discussion

Inversion $2z^3$ Was Generated by Ectopic Recombination between *Galileo* Copies

Many studies have shown the potential of TEs to induce chromosomal rearrangements in experimental *Drosophila* populations implicating retrotransposons (e.g. *BEL*, *roo*, *Doc*, and *I*) as well as transposons (e.g. *P*, *hobo*, and *FB*) [74]. In contrast, the evidence for the involvement of TEs in the generation of natural *Drosophila* inversions, i.e. those effectively contributing to adaptation and/or evolution of natural populations, is scarce (see Introduction). We have previously found that the cut-and-paste transposon *Galileo* was involved in the generation of two polymorphic inversions of *D. buzzatii*, $2j$ and $2q^7$ [37,38]. Here we have isolated and sequenced the breakpoints of another polymorphic inversion of *D. buzzatii*, $2z^3$. Our results provide the most compelling evidence for the participation of *Galileo* in the generation of *Drosophila* inversions and for ectopic recombination as the responsible mechanism.

Several TE insertions were found at the breakpoint regions in the chromosome with the $2z^3$ inversion that were not present in non-inverted chromosomes (Table 1). Remarkably, only *GalileoN* was present at the two breakpoint junctions. This fact and the evidence presented below indicate that *GalileoN* is the element responsible for the generation of the $2z^3$ inversion. Two other TE insertions, *BuT5* and *LINE*-like, were found nested within the

GalileoN copy in the distal breakpoint and another two, *BuT3* and *BuT4*, within the *GalileoN* copy in the proximal breakpoint. These four TE insertions are present at a single breakpoint junction only and each of them is flanked by identical direct TSDs. Thus, they are unlikely to be responsible for the generation of the inversion and are best interpreted as secondary colonizers of the breakpoint regions (see below). Another two TE fragments (*BuT5* and *GalileoN*) are present in the proximal breakpoint region (but not in the junction) of non-inverted chromosomes and thus can not be involved in the generation of the inversion either.

Two processes can explain the induction of chromosomal inversions by TEs: ectopic recombination [12,74] and aberrant transposition [13–17]. Ectopic recombination requires the presence in the parental chromosome of two homologous TE copies inserted in opposite orientation at different sites. After the inversion is generated, two chimeric TE copies are expected to be found flanking the inverted segment with their TSDs exchanged. On the other hand, two transposon copies may participate in an aberrant transposition event, by which a hybrid element formed by the 5' end of one copy and the 3' end of the other copy transposes to a new chromosomal site. The outcome of this process is an inversion flanked by two transposon copies in opposite orientation accompanied by deletions or duplications when the original copies were inserted at separate chromosomal sites (Figure S1). The lack of any deletions or duplications and the pattern of TSDs in the $2z^3$ breakpoints allow us to reject this latter possibility. However, we must consider the possibility of an aberrant transposition with the two original transposon copies located at the same chromosomal site (hybrid insertion model). The outcome in this case (Figure S1 A) is strikingly similar to that of ectopic recombination except for the fact that the two TE copies flanking the inversion are identical under the hybrid element insertion model but chimeric under the ectopic recombination [38].

The two *GalileoN* copies found in the $2z^3$ breakpoints (named *GalileoN-4* and *GalileoN-5*) have similar sizes and structures, with relatively long TIRs and a middle segment oriented in opposite direction in the two copies, and show a high similarity with two other copies previously described (*GalileoN-1* and *GalileoN-2*) [60,61]. Each of the latter two copies was flanked by perfect 7-bp TSDs generated upon insertion. By contrast, the 7-bp duplications flanking the *GalileoN* copies at the $2z^3$ breakpoints are exchanged (Figure 2 and Results). In the non-inverted chromosomes, only one copy of the corresponding 7-bp target sequence is detected at each breakpoint (Figure 2). These observations are consistent with the presence of two *GalileoN* insertions in the parental chromosome and the generation of the $2z^3$ inversion by ectopic recombination between them, but does not rule out the hybrid element insertion model (see above). Further evidence was revealed by comparing the nucleotide sequence of the TIRs within and between *GalileoN* copies. *GalileoN-1* and *GalileoN-2* possess TIRs >99% identical within each copy but ~7% divergent between copies (Table 2). In contrast, *GalileoN-4* and *GalileoN-5* show TIRs that are ~6% divergent within each copy but >99% identical between copies (Table 2). These results suggest that both *GalileoN-4* and *GalileoN-5* are chimeric. A closer scrutiny of the four *GalileoN* copies revealed a striking pattern and led to the same conclusion (Figure 5). In 33 variable sites, from position 1 through 824, the nucleotide present in *GalileoN-4* is identical to that in *GalileoN-1* and the nucleotide present in *GalileoN-5* is identical to that in *GalileoN-2* (Figure 5 top). The situation is completely reversed for 20 variable sites from position 966 to the end of the element where the nucleotide present in *GalileoN-4* is identical to that in *GalileoN-2* while that in *GalileoN-5* is

Table 2. Nucleotide divergence between the TIRs of four *GalileoN* copies.

TIR	1R	2L	2R	4L	4R	5L	5R
1L	0.0018	0.0670	0.0670	0.0326	0.0708	0.0552	0.0727
1R		0.0690	0.0690	0.0345	0.0728	0.0271	0.0747
2L			0.0071	0.0234	0.0591	0.0514	0.0234
2R				0.0591	0.0234	0.0514	0.0234
4L					0.0628	0.0071	0.0647
4R						0.0552	0.0036
5L							0.0570

TIR number indicates the *GalileoN* copy, and L and R correspond to 5' TIR and 3' TIR, respectively; values in boldface correspond to the comparisons between TIRs of the same copy.

doi:10.1371/journal.pone.0007883.t002

identical to that in *GalileoN*-1 (Figure 5 top). Phylogenetic analyses of the four sequences carried out separately for the two portions of the element (Figure 5 bottom) and the maximum chi-square method ($\chi^2 = 53.00$, $df = 1$, $P < 1 \times 10^{-7}$) [75,76] corroborated the chimeric structure of *GalileoN*-4 and *GalileoN*-5. These observations provide strong support for the ectopic recombination model and suggest that the recombination event that gave rise to the $2z^3$ inversion took place within 141-bp of the middle segment between positions 825 and 965 of *GalileoN* (Figure 5). The absence of *GalileoN* insertions in the analyzed non-inverted chromosomes should be no surprise because insertions of actively transposing families are expected to be present at low population frequencies under transposition-selection balance [77,78] and we sampled just a few non-inverted chromosomes.

We can conclude that the three polymorphic inversions of *D. buzzatii* studied so far, $2j$, $2q^7$ and $2z^3$, have been generated by the same TE family, *Galileo*, and very likely by the same molecular mechanism, ectopic recombination. In all three cases, after the generation of the inversion, many TE copies have accumulated at the breakpoint regions, which became hotspots for secondary TE insertions (Table 3). This accumulation is probably a consequence of the reduction of recombination in these regions [79,80] that

protects TE copies from being eliminated by deleterious ectopic exchanges [77,78]. It is intriguing though that the 40 TE copies associated with inversion breakpoints in *D. buzzatii* belong to a limited set of nine TE families (Table 3). All of them but one (the *LLNE*-like element in the distal breakpoint of inversion $2z^3$) are Class II elements: *ISBu* elements are Helitrons [81] and the remaining elements are cut-and-paste transposons [82]. This enrichment of breakpoint regions in specific TE families may be due (1) to the fact that these TE families were among the most transpositionally active elements in the *D. buzzatii* genome when the opportunity window for insertion was open, and/or (2) to insertional preference [83].

Because many different TE families are able to induce chromosomal rearrangements in *Drosophila* [74], the question arises as to why the three polymorphic *D. buzzatii* inversions should be generated by the same TE family, namely *Galileo*. The frequency of ectopic recombination should increase with copy number and length, and this prediction is borne out by the data ([78], D. Petrov, personal communication). In the *D. melanogaster* genome, at least 121 TE families are present [84,85]. A total of 996 copies from 81 families were annotated in the euchromatin of the sequenced genome (excluding the proximal 2 Mb where TEs regularly accumulate) and copy number per family varied between 1 and 124 with an average of 12.3 [84]. Although no detailed inventory of the TE families in the *D. buzzatii* genome is yet available, there is no ground for assuming a smaller number of families than in *D. melanogaster*. *Galileo* copy number per genome was estimated as 11.7 in the euchromatic distal-central region of chromosomes (i.e. excluding the dot and pericentromeric regions) [61]. The analogous figure for *BuT5* is 11.4 copies per genome and lower values were estimated for another five *D. buzzatii* transposons [83]. In summary, *Galileo* copy number does not seem particularly high in the *D. buzzatii* genome, although more data is needed. Length of *Galileo* copies is not unusual either. The canonical copy is ~5.4 kb long [62] but most copies are non-autonomous and much shorter. Average length (\pm SD) of a combined sample of 23 non-autonomous copies of *GalileoG*, *GalileoN* and *GalileoK* is 953 bp (\pm 640 bp) [61]. In *D. melanogaster*, the average length of the TE copies annotated by [84] was 2.9 kb.

Two characteristics of *Galileo* can explain its primary role in the generation of rearrangements by ectopic recombination: (1) its

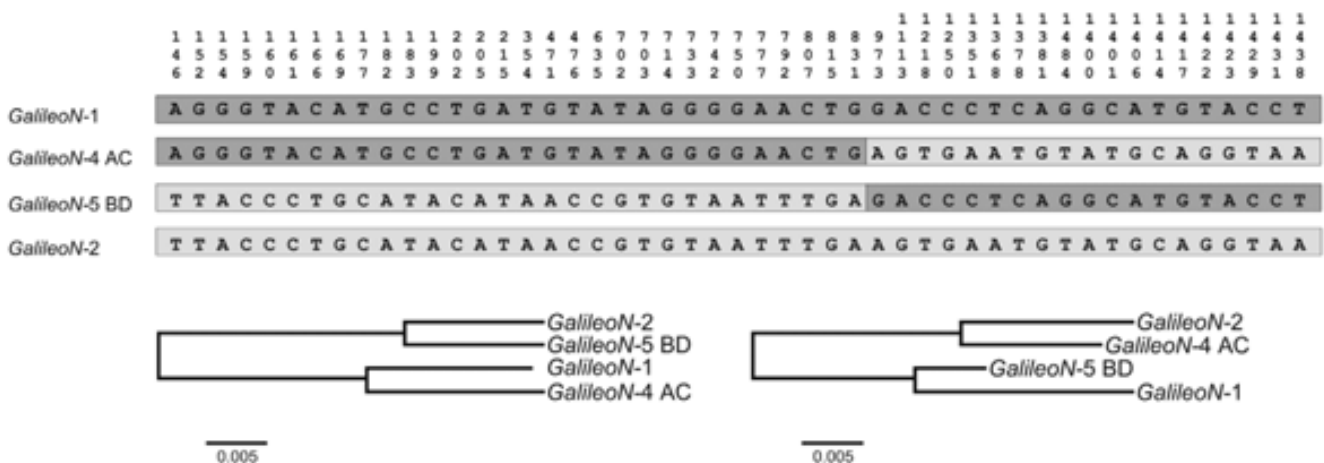


Figure 5. Chimeric structure of the two *GalileoN* copies (*GalileoN*-4 and *GalileoN*-5) observed at the breakpoints of inversion $2z^3$. *GalileoN*-1 and *GalileoN*-2 were found in a previous study [60]. Top: Nucleotides present in the four *GalileoN* copies at 53 variable sites are shown. Bottom: Neighbor-joining phylogenetic trees of the *GalileoN* sequences built separately for the two portions of the sequence: sites 1–824 (left) and sites 966–1567 (right).

doi:10.1371/journal.pone.0007883.g005

Table 3. Number of TE copies found in the breakpoint regions of three *D. buzzatii* polymorphic inversions.

Inversion	$2j$		$2q^7$		$2z^3$		TOTAL
	D	P	D	P	D	P	
Breakpoint region ^a	D	P	D	P	D	P	D+P
<i>Galileo</i>	4	5	1	3	1	1	16
<i>BuT1</i>		1					1
<i>BuT2</i>		1					1
<i>BuT3</i>	2	3		1		1	7
<i>BuT4</i>	1					1	2
<i>BuT5</i>		1	1	4	1	1	8
<i>BuT6</i>			1				1
<i>ISBu</i>	3	1					4
<i>LINE-like</i>					1		1
TOTAL	10	12	3	8	3	4	40

Number of chromosomal lines investigated: 30 for $2j$, 6 for $2q^7$ and 1 for $2z^3$.

Data from [38,60] and this work.

^aD = distal; P = proximal.

doi:10.1371/journal.pone.0007883.t003

transpositional activity; and (2) its unusual structure. *Galileo* belongs to the *P* superfamily of TIR transposons and is likely to transpose by a cut-and-paste mechanism similar to that of the *D. melanogaster* *P* element [86,87]. This transposition mechanism involves the binding of the transposase to the element TIRs and the excision of the element generating a DSB at the donor site followed by the integration of the element into a different chromosomal site. Hence DSBs produced during normal or aberrant transposition events may provide the initial step for ectopic recombination events. The accumulation of *Galileo* copies after the generation of inversions $2j$ and $2q^7$ (Table 3) indicates that *Galileo* is (or has been recently) active in the genome of *D. buzzatii*. Nevertheless, unless *Galileo* has an unusually high transposition rate, this explanation is insufficient because *Galileo* is not the only TE family transpositionally active in the *D. buzzatii* genome (at least another eight TE families must be active; Table 3).

The participation of *Galileo* in the generation of inversions may be also related to its unusual structure with up to 1.2-kb long TIRs [61,62]. The two *GalileoN* copies involved in the generation of the $2z^3$ inversion have ~575 bp long TIRs separated by a ~350 bp long middle segment (Table 1). This kind of spaced inverted repeat sequences is well known to form stem-loop structures in single-stranded DNA or cruciform structures in double-stranded DNA and induce DSBs and rearrangements in a wide variety of organisms [88–93]. Generation of DSBs by these secondary structures may be due to the fact that they are substrates for nuclease cleavage or because they interrupt replication fork progression [94,95]. In *D. melanogaster*, *Foldback* (FB) elements, which also present very long TIRs and induce secondary structures, are known to cause rearrangements at a high rate in the laboratory [96,97]. We propose that the long TIRs of *Galileo* induce the formation of secondary structures and DSBs at high rate and this contributes to its unique capacity to generate chromosomal inversions. The fact that the recombination event that generated inversion $2z^3$ took place in the middle segment of *GalileoN* seems consistent with nuclease cleavage at the loop.

Functional Consequences of the $2z^3$ Inversion

Inversion $2z^3$ seems to have a recent origin as no fixed nucleotide substitution was observed in the breakpoint regions

between non-inverted and inverted chromosomes (Figure 3). This is in clear contrast with the ~1 Myr and ~0.5 Myr old inversions $2j$ and $2q^7$ where 17 and 14 fixed nucleotide substitutions were observed, respectively [38,60]. The monomorphism of the α -*esterase5* gene in $2jz^3$ chromosomes is also consistent with a recent origin of inversion $2z^3$ [98]. In spite of being a very young inversion, $2z^3$ exhibits a widespread distribution in natural populations (see Introduction), suggesting that it must have a considerable selective value. In Argentina, the frequency of $2jz^3$ is significantly correlated with latitude, a putatively selective pattern [65]. Furthermore, selection component analyses and biometrical studies they all have detected significant effects of $2jz^3$ chromosomes [99–102]. One possible explanation for its adaptive advantage is provided by the position effect hypothesis, which proposes that the localization of the inversion breakpoints near or inside genes could affect their function or expression profile by disrupting their coding regions or causing changes in the promoter and regulatory regions [103,104]. Another factor that could affect the expression of genes adjacent to the breakpoints is the presence of TEs in these regions as they have been shown to alter gene expression in different ways [103,105,106].

The $2z^3$ proximal breakpoint lies in a region previously sequenced where a gene named *lodestar* (*lds*) had been tentatively annotated [67]. A comparative analysis with other *Drosophila* genomes and expression experiments by RT-PCR discarded the *lds* annotation and has unveiled two novel genes flanking the inversion breakpoint, *Dlh* in region C and *Mdp* in region D. Three observations suggest that these two genes are fully functional. (i) In *D. buzzatii*, both genes are expressed throughout the whole life cycle, although they present slightly different expression patterns (results not shown). (ii) Their overall structure and encoded protein sequence are conserved in at least another *Drosophila* species (Table S2). (iii) Both genes are evolving under purifying selection with Ka/Ks ratios significantly different from 1 (strict neutrality). The relatively short intergenic region (796 bp) and the close proximity of the proximal breakpoint to the initiation codon of *Dlh* (118 bp) suggest that the inversion might be affecting the expression of *Dlh* and/or *Mdp*, a question that deserves further work.

In *D. buzzatii*, the Hox gene complex is split in three portions: *proboscipedia* (*pb*)-*Ultrabithorax* (*Ubx*), *abdominalA* (*abdA*)-*AbdominalB* (*AbdB*) and *labial* (*lab*) [66,67]. We analyzed the breakpoints of inversion $2z^3$ in part because of the cytological vicinity of the $2z^3$ proximal breakpoint to the *pb-Ubx* portion of the Hox gene complex [63,66,67]. Our results show that the $2z^3$ proximal breakpoint lies outside of the Hox gene complex ~23.7-kb downstream of *pb*. The segment that separates the $2z^3$ proximal breakpoint from *pb* contains three genes, *CG17836*, *CG14290* and *Dlh*. It seems unlikely that the $2z^3$ proximal breakpoint altered the regulatory sequences or the expression pattern of *pb* because the *lab-pb* split that took place much nearer the 3' end of *pb* in the ancestor of the repleta group did not [67]. Nevertheless, the *pb-Ubx* portion of the Hox gene complex is located within the inverted segment and thus the $2z^3$ inversion relocates these genes to a much more distal region within chromosome 2. Whether this change in the chromatin environment has had any effect on the expression of Hox genes remains an open question.

Materials and Methods

Drosophila Stocks

Nine lines of *D. buzzatii* homokaryotypic for one of three different chromosome 2 arrangements (*2st*, $2j$ and $2jz^3$) were used. These lines were isolated from natural populations with different

geographical origin: st-1, Carboneras (Spain); st-12, Trinkey (Australia); st-13, Mazán (Argentina); j-2, Carboneras (Spain); j-8, San Luis (Argentina); j-9, Quilmes (Argentina); j-13, Guaritas (Brazil); j-19, Ticucho (Argentina); and jz³-2, Carboneras (Spain). The stock of *D. mojavensis* (15081-1352.22, UC San Diego Drosophila Species Stock Center) comes from Santa Catalina Island (California) and is the stock used to sequence the *D. mojavensis* genome [70].

Probes and *In Situ* Hybridization

DNA from BAC and plasmid clones was extracted by alkaline lysis following standard protocols and used as probes for *in situ* hybridization. All remaining probes were produced by polymerase chain reaction (PCR) amplification of *D. buzzatii* or *D. mojavensis* genomic DNA with different primer pairs. Probes were labelled with biotin-16-dUTP (Roche) by random priming and hybridization to the larval salivary gland polytene chromosomes was carried out according to the procedure described [107]. Intraspecific *in situ* hybridizations with *D. buzzatii* lines and probes were carried out at 37°C while interspecific hybridizations of *D. mojavensis* probes to *D. buzzatii* polytene chromosomes were carried out at 25°C. Hybridization results were recorded as digital images captured with phase contrast Nikon Optiphot-2 microscope at 600× magnification and a Nikon Coolpix 4500 camera. Cytological localization of the hybridization signal was determined using the cytological maps of *D. buzzatii* [63,69].

Physical Mapping of the Inversion Breakpoints

We searched the BAC-based physical map of the *D. buzzatii* genome [69] for clones located near the cytological breakpoints and selected eight clones from contig 961 mapping near the distal breakpoint, and seven clones from contig 968 mapping near the proximal breakpoint (Table S3). The fifteen BAC clones were hybridized to the salivary gland chromosomes of one line with the inversion (jz³-2) and one line without the inversion (j-9) to identify those clones containing a breakpoint (that should produce two hybridization signals in the first case and a single hybridization signal in the second). Three BAC clones from contig 961 (18L15, 15P22 and 15L20) were found to include the distal breakpoint (Figure S4A), and four clones from contig 968 (22N23, 22M06, 16A20, and 40C11) were found to contain the proximal breakpoint (Figure S4E).

Both ends of each BAC clone bearing the distal breakpoint were sequenced and the sequences mapped onto the genome sequence of *D. mojavensis* using BLASTN (Figure 1 left). The distal breakpoint was located in the overlapping region between the three *D. buzzatii* BAC clones, a segment ~50-kb long of *D. mojavensis* scaffold_6540 that corresponds to chromosome 2 [108]. To narrow down the position of the breakpoint we chose four genes within this segment (*CG1193*, *CG14906*, *Adk3* and *CG4674*) and used them as probes for *in situ* hybridization to 2jz³ chromosomes (Table S4). The *CG1193* probe (marker 1 in Figure 1 left) mapped at the distal breakpoint, outside the inversion, while the other three probes (markers 2, 3 and 4 in Figure 1) hybridized at the proximal breakpoint, indicating that they are located inside the inverted segment. As a result, we located the distal breakpoint in the 13-kb segment between genes *CG1193* and *CG14906* (markers 1 and 2 in Figure 1). Seven genes had been annotated in this segment of *D. mojavensis* chromosome 2 and we designed primers to amplify the intergenic region between each pair of genes in this species, as well as in *D. buzzatii* strains with and without inversion 2z³. Our rationale was that the intergenic region containing the distal breakpoint would amplify in *D. mojavensis* and in the line with the non-inverted chromosome,

but not in the line carrying the inversion. In fact, all the intergenic segments were amplified in the three lines, except that between *CG2046* and *CG10326* (segment 9 in Figure 1 left) which failed to amplify in the line carrying the inversion. To corroborate this observation, PCR products amplified using the primers 8F-8R, 9F-9R and 10F-10R were used as *in situ* hybridization probes to chromosomes with the inversion, and they produced the expected results (Figure S4B, C and D). Therefore, the distal breakpoint of inversion 2z³ was located in the ~600-bp region between genes *CG2046* and *CG10326* of *D. mojavensis*.

One of the four BAC clones bearing the 2z³ proximal breakpoint (BAC 40C11) was already fully sequenced and annotated and a physical map of the region was built using sequence tagged sites (STSs) [67]. This map allowed us to locate the proximal breakpoint in the ~70-kb region of overlap between the four clones (Figure 1 right). Three STS markers generated in this region were amplified and hybridized to 2jz³ chromosomes, in order to further delimit the region which contains the proximal breakpoint (Figure 2 right). One marker (number 13 in Figure 1 right) hybridized to the distal breakpoint and therefore was located inside the inversion, whereas the other two (markers 14 and 15 in Figure 2 right) mapped on the region of the proximal breakpoint, indicating that they are located outside the inverted segment. As a result, the proximal breakpoint could be narrowed down to a 16-kb segment between genes *CG17836* and *CG2520* (markers 13 and 14 in Figure 1 right). Ten plasmid subclones from BAC 40C11 which cover this segment were also used for hybridization to inverted chromosomes (Figure S4F, G and H and Table S4), allowing us to locate the proximal breakpoint more precisely in the ~0.8-kb intergenic region between genes *Dlh* and *Mdp* (Figure 1 right).

Southern Blot and Screening of Genomic Libraries

Southern hybridization and library screenings were carried out by standard methods [109]. Three different probes amplified from *D. buzzatii* DNA: DF-DR (800 bp), CF-CR (337 bp) and BF-BR (505 bp) were used (Table S5). Probes were labelled by random priming with digoxigenin-11-dUTP under the conditions specified by the supplier (Roche). Hybridization was carried out overnight at 42°C in a standard hybridization solution (Roche). Stringency washes were performed with 0.5x SSC 0.1% SDS solution at 65°C. Two lambda genomic libraries were screened. One library was constructed with DNA derived from *D. buzzatii* line jz³-2 using the LambdaGEM-11 vector following manufacturer's instructions (Promega). The second lambda library was derived previously from *D. buzzatii* line jz³-4 [60] and was amplified using standard methods [109]. Two positive clones (λ z³-91 and λ z³-96) were recovered from the first library with probe CF-CR and six positive clones were recovered from the second library, three with probe DF-DR (λ z³-77, λ z³-79 and λ z³-98) and three with probe BF-BR (λ z³-99, λ z³-102 and λ z³-104). The span of each clone was determined through a combination of PCR, restriction mapping and Southern blotting. DNA fragments of interest from positive phages were subcloned into pBluescript II SK vector (Stratagene).

PCR Amplification

Polymerase chain reaction was carried out in a volume of 25 μ l, including 50–100 ng of genomic DNA, 10 pmol of each primer, 100 μ M dNTPs, 1x buffer and 1–1.5 units of Taq DNA polymerase. Temperature cycling conditions were 30 rounds of 30 s at 94°C; 30 s at the annealing temperature, and 30–60 s at 72°C, with annealing temperatures varying from 55 to 60°C depending on the primer pair. Sequences of oligonucleotide primers are given in Table S5.

RNA Extraction and RT-PCR Amplification

Total RNA was isolated from embryos, larvae, pupae, and adults of the *D. buzzatii* st-1 line using TRIzol (Invitrogen). Total RNA was treated with 1 unit of DNase I (Ambion) for 30 min at 37°C to eliminate DNA contamination. cDNA was synthesized from 1 µg of DNase I-treated RNA by using an oligo(dT) primer (Transcriptor First Strand cDNA Synthesis kit for RT-PCR, Roche). PCR reactions were performed as describe above. To differentiate the size of amplification products, both cDNA and st-1 genomic DNA were used as templates. RT-PCR products were sequenced and their sequences compared with those of genomic DNA to determine exon-intron boundaries (Figures S2 and S3).

DNA Sequencing and Sequence Analysis

Sequencing was performed in the Servei de Genòmica of the Universitat Autònoma de Barcelona, Macrogen Inc. (Seoul, Korea) and GATC Biotech (Konstanz, Germany). Fragments cloned into pBluescript II SK were sequenced with the M13 universal and reverse primers. PCR products were gel purified using QIAquick Gel Extraction Kit (Qiagen), and sequenced directly with the same primers used for amplification.

Sequences from different lines were aligned with MUSCLE 3.2 [110] and similarity searches in the GenBank/EMBL, Assembly/Alignment/Annotation of 12 related *Drosophila* species (<http://rana.lbl.gov/drosophila/>) and FlyBase databases were carried out using BLASTN [111]. Nucleotide variability was estimated by means of the number of segregating sites (S), and the nucleotide diversity (π , average number of pairwise differences per site) using DnaSP (version 4.50.3) software [112]. This software was also used to test for differences in nucleotide variability by means of computer simulations based on the coalescent process. Simulations were carried out given the number of segregating sites and analysing the nucleotide diversity (π) on the genealogy, fixing the options of no recombination to AB region and free recombination to CD region, because AB region mapped inside the 2j inversion. Interspecific nucleotide and amino acid similarities were estimated with MEGA 4 [113]. The ratios of non-synonymous to synonymous nucleotide substitutions (Ka/Ks) were estimated using Nei-Gojobori method and Jukes-Cantor distance. The null hypothesis that Ka/Ks = 1 was tested by means of the Z-test of selection. Phylogenetic analyses were also conducted using MEGA 4.

Sequence data from this article have been deposited in the GenBank/EMBL Database Libraries under accession nos. GU132438-GU132454.

Supporting Information

Figure S1 Chromosomal inversions may be generated by transposons when two ends that are not part of the same transposon participate in an aberrant transposition event to a new site [13–17]. Target site duplications (TSD) are indicated by ○ or □ (cooperating TE copies) and Δ (new insertion site). (A) The two TE copies are located at the same site of sister chromatids or homologous chromosomes and share the same TSD (○). The result of the aberrant transposition is an inversion (segment BC) flanked by two TE copies. (B) The two TE copies are inserted at separate sites in the two homologous chromosomes and each has its own TSD (indicated by ○ and □). The aberrant transposition event produces an inversion (segment BC) and a deletion (segment D). (C) The two TE copies are arranged as in (B) but two different element ends are involved. The resulting chromosome carries an inversion (segment BC) and a duplication (segment D). (D) The

two TE copies are inserted at separate sites on the same chromatid and each has its own TSD (indicated by ○ and □). The resulting chromosome has an inversion (segment BC) and a deletion (segment D).

Found at: doi:10.1371/journal.pone.0007883.s001 (0.02 MB PDF)

Figure S2 Alignment of gene *Mdp* sequences in three *Drosophila* species. The aligned sequences are: positions 50294–51354 from *D. buzzatii* BAC clone 40C11 (accession number AY900632), positions 6137692–6136590 from *D. mojavensis* scaffold_6540 and positions 5807143–5806092 from *D. virilis* scaffold_12855. Yellow boxes indicate exons with the initial methionine and the final stop codon colored in orange and red, respectively. The premature stop codon found in the *D. mojavensis* sequence is also shown as a red box. Note that there are some parts of the sequence upstream of the coding region that are conserved in the different species suggesting that they may be part of the 5' UTR or the regulatory regions of the gene. A putative polyA signal determined only on the basis of sequence conservation in the different species is included in a purple rectangle. The blue bar below the alignment indicates the 763-bp fragment amplified by RT-PCR and sequenced in *D. buzzatii* with primer pair DF-DR. The protein sequence encoded by the *D. buzzatii* gene is shown above the alignment. The residues enclosed in a green box correspond to the MADF domain found using InterProScan (<http://www.ebi.ac.uk/Tools/InterProScan/>).

Found at: doi:10.1371/journal.pone.0007883.s002 (0.01 MB PDF)

Figure S3 Alignment of gene *Dlh* sequences in two *Drosophila* species. The aligned sequences are: positions 52175–55219 from *D. buzzatii* BAC clone 40C11 (accession number AY900632) and positions 6136143–6133352 from *D. mojavensis* scaffold_6540. This gene could not be found in the *D. virilis* genome sequence. Yellow boxes indicate exons with the initial methionine and the final stop codon colored in orange and red, respectively. Enclosed in a purple rectangle is the codon in the second exon of the gene that becomes a polymorphic premature stop codon in lines j-9 and jz3-1 by changing from TCA to TAA. No further upstream non-coding sequence could be included in the alignment because of the presence of a polymorphic *GalileoN* insertion in the st-1 line, from which the *D. buzzatii* BAC clone is derived. Bars below the alignment in different shades of blue indicate the three overlapping fragments amplified by RT-PCR and sequenced in *D. buzzatii* with primer pairs CF-CR (278 bp), CF-RT1R (609 bp) and RT2F-RT2R (1,011 bp). The protein sequence encoded by the *D. buzzatii* gene is shown above the alignment. The residues enclosed in a dark green box correspond to a SNF2-related or a DEAD-like helicase N-terminal domain and the aminoacids in a light green box correspond to a DNA/RNA helicase C-terminal domain. The protein domains have been analyzed using InterProScan (<http://www.ebi.ac.uk/Tools/InterProScan/>).

Found at: doi:10.1371/journal.pone.0007883.s003 (0.02 MB PDF)

Figure S4 *In situ* hybridization to *D. buzzatii* chromosomes carrying inversion $2z^3$ of BAC clones, plasmid clones and PCR probes coming from the distal breakpoint (A-D) and the proximal breakpoint (E-H). A: BAC clone 18L15; B: PCR fragment 10F-10R; C: PCR fragment 9F-9R; D: PCR fragment 8F-8R. E: BAC clone 40C11; F: plasmid clone 9F01; G: plasmid clone 8H04; H: plasmid clone 8D03. Arrows indicate hybridization signals.

Found at: doi:10.1371/journal.pone.0007883.s004 (4.84 MB TIF)

Table S1 Nucleotide variability in non-inverted chromosomes. N = number of chromosomal lines; m = number of compared nucleotides.

Found at: doi:10.1371/journal.pone.0007883.s005 (0.05 MB PDF)

Table S2 Structure and similarities of two novel *Drosophila* genes: *MADF domain protein (Mdp)* and *DEAD-like helicase (Dlh)*. NT = nucleotide; AA = amino acid.

Found at: doi:10.1371/journal.pone.0007883.s006 (0.01 MB PDF)

Table S3 BAC clones used for in situ hybridization.

Found at: doi:10.1371/journal.pone.0007883.s007 (0.01 MB PDF)

Table S4 Plasmid clones used as probes for *in situ* hybridization to map the proximal breakpoint of the 2 ζ ³ inversion.

References

- Finnegan DJ (1989) Eukaryotic transposable elements and genome evolution. *Trends Genet* 5: 103–107.
- McDonald JF (1993) Evolution and consequences of transposable elements. *Curr Opin Genet Dev* 3: 855–864.
- Kidwell MG, Lisch D (2002) Transposable Elements as Sources of Genomic Variation. In: Craig NL, ed. *Mobile DNA II*. Washington, D.C.: ASM Press. pp 59–90.
- Coghlan A, Eichler EE, Oliver SG, Paterson AH, Stein L (2005) Chromosome evolution in eukaryotes: a multi-kingdom perspective. *Trends Genet* 21: 673–682.
- Powell JR (1997) *Progress and prospects in evolutionary biology: The Drosophila model*. Oxford: Oxford University Press.
- Bhutkar A, Schaeffer SW, Russo SM, Xu M, Smith TF, et al. (2008) Chromosomal rearrangement inferred from comparisons of 12 *Drosophila* genomes. *Genetics* 179: 1657–1680.
- Shaffer LG, Lupski JR (2000) Molecular mechanisms for constitutional chromosomal rearrangements in humans. *Annu Rev Genet* 34: 297–329.
- Cáceres M, Sullivan RT, Thomas JW (2007) A recurrent inversion on the eutherian X chromosome. *Proc Natl Acad Sci U S A* 104: 18571–18576.
- Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, et al. (2007) Paired-end mapping reveals extensive structural variation in the human genome. *Science* 318: 420–426.
- Kidd JM, Cooper GM, Donahue WF, Hayden HS, Samps N, et al. (2008) Mapping and sequencing of structural variation from eight human genomes. *Nature* 453: 56–64.
- Antonacci F, Kidd JM, Marques-Bonet T, Ventura M, Siswara P, et al. (2009) Characterization of six human disease-associated inversion polymorphisms. *Hum Mol Genet* 18: 2555–2566.
- Petes TD, Hill CW (1988) Recombination between repeated genes in microorganisms. *Annu Rev Genet* 22: 147–168.
- Svoboda YH, Robson MK, Sved JA (1995) P-element-induced male recombination can be produced in *Drosophila melanogaster* by combining end-deficient elements in trans. *Genetics* 139: 1601–1610.
- Gray YH, Tanaka MM, Sved JA (1996) P-element-induced recombination in *Drosophila melanogaster*: hybrid element insertion. *Genetics* 144: 1601–1610.
- Gray YH (2000) It takes two transposons to tango: transposable-element-mediated chromosomal rearrangements. *Trends Genet* 16: 461–468.
- Huang JT, Dooner HK (2008) Macrotransposition and other complex chromosomal restructuring in maize by closely linked transposons in direct orientation. *Plant Cell* 20: 2019–2032.
- Zhang J, Yu C, Pulletikurti V, Lamb J, Danilova T, et al. (2009) Alternative Ac/Ds transposition induces major chromosomal rearrangements in maize. *Genes Dev* 23: 755–765.
- Pastink A, Ecken JC, Lohman PH (2001) Genomic integrity and the repair of double-strand DNA breaks. *Mutat Res* 480–481: 37–50.
- Sonoda E, Hohegger H, Saberi A, Taniguchi Y, Takeda S (2006) Differential usage of non-homologous end-joining and homologous recombination in double strand break repair. *DNA Repair (Amst)* 5: 1021–1029.
- Hefferin ML, Tomkinson AE (2005) Mechanism of DNA double-strand break repair by non-homologous end joining. *DNA Repair (Amst)* 4: 639–648.
- Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423: 241–254.
- Daveran-Mingot ML, Campo N, Ritzenthaler P, Le Bourgeois P (1998) A natural large chromosomal inversion in *Lactococcus lactis* is mediated by homologous recombination between two insertion sequences. *J Bacteriol* 180: 4834–4842.
- Parkhill J, Wren BW, Thomson NR, Titball RW, Holden MT, et al. (2001) Genome sequence of *Yersinia pestis*, the causative agent of plague. *Nature* 413: 523–527.
- Deng W, Burland V, Plunkett G 3rd, Boutin A, Mayhew GF, et al. (2002) Genome sequence of *Yersinia pestis* KIM. *J Bacteriol* 184: 4601–4611.
- Brinig MM, Cummings CA, Sanden GN, Stefanelli P, Lawrence A, et al. (2006) Significant gene order and expression differences in *Bordetella pertussis* despite limited gene content variation. *J Bacteriol* 188: 2375–2382.
- Redder P, Garrett RA (2006) Mutations and rearrangements in the genome of *Sulfolobus solfataricus* P2. *J Bacteriol* 188: 4198–4206.
- Beare PA, Unsworth N, Andoh M, Voth DE, Omsland A, et al. (2009) Comparative genomics reveal extensive transposon-mediated genomic plasticity and diversity among potential effector proteins within the genus *Coxiella*. *Infect Immun* 77: 642–656.
- Chain PS, Carniel E, Larimer FW, Lamerdin J, Stoutland PO, et al. (2004) Insights into the evolution of *Yersinia pestis* through whole-genome comparison with *Yersinia pseudotuberculosis*. *Proc Natl Acad Sci U S A* 101: 13826–13831.
- Parkhill J, Sebahia M, Preston A, Murphy LD, Thomson N, et al. (2003) Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica*. *Nat Genet* 35: 32–40.
- Cho NH, Kim HR, Lee JH, Kim SY, Kim J, et al. (2007) The *Orientia tsutsugamushi* genome reveals massive proliferation of conjugative type IV secretion system and host-cell interaction genes. *Proc Natl Acad Sci U S A* 104: 7981–7986.
- Reith ME, Singh RK, Curtis B, Boyd JM, Bouevitch A, et al. (2008) The genome of *Aeromonas salmonicida* subsp. *salmonicida* A449: insights into the evolution of a fish pathogen. *BMC Genomics* 9: 427.
- Roeder GS (1983) Unequal crossing-over between yeast transposable elements. *Mol Gen Genet* 190: 117–121.
- Picologlou S, Dicig ME, Kovarik P, Liebman SW (1988) The same configuration of Ty elements promotes different types and frequencies of rearrangements in different yeast strains. *Mol Gen Genet* 211: 272–281.
- Kupiec M, Petes TD (1988) Meiotic recombination between repeated transposable elements in *Saccharomyces cerevisiae*. *Mol Cell Biol* 8: 2942–2954.
- Schwartz A, Chan DC, Brown LG, Alagappan R, Pettay D, et al. (1998) Reconstructing hominid Y evolution: X-homologous block, created by X-Y transposition, was disrupted by Yp inversion through LINE-LINE recombination. *Hum Mol Genet* 7: 1–11.
- Lee J, Han K, Meyer TJ, Kim HS, Batzer MA (2008) Chromosomal inversions between human and chimpanzee lineages caused by retrotransposons. *PLoS ONE* 3: e4047.
- Cáceres M, Ranz JM, Barbadilla A, Long M, Ruiz A (1999) Generation of a widespread *Drosophila* inversion by a transposable element. *Science* 285: 415–418.
- Casals F, Cáceres M, Ruiz A (2003) The foldback-like transposon Galileo is involved in the generation of two different natural chromosomal inversions of *Drosophila buzzatii*. *Mol Biol Evol* 20: 674–685.
- Richards S, Liu Y, Bettencourt BR, Hradecky P, Letovsky S, et al. (2005) Comparative genome sequencing of *Drosophila pseudoobscura*: chromosomal, gene, and cis-element evolution. *Genome Res* 15: 1–18.
- Evans AL, Mena PA, McAllister BF (2007) Positive selection near an inversion breakpoint on the neo-X chromosome of *Drosophila americana*. *Genetics* 177: 1303–1319.
- Mathiopoulos KD, della Torre A, Predazzi V, Petrarca V, Coluzzi M (1998) Cloning of inversion breakpoints in the *Anopheles gambiae* complex traces a

- transposable element at the inversion junction. *Proc Natl Acad Sci U S A* 95: 12444–12449.
42. Sharakhov IV, White BJ, Sharakhova MV, Kayondo J, Lobo NF, et al. (2006) Breakpoint structure reveals the unique origin of an interspecific chromosomal inversion (2La) in the *Anopheles gambiae* complex. *Proc Natl Acad Sci U S A* 103: 6258–6262.
 43. Ranz JM, Maurin D, Chan YS, von Grotthuss M, Hillier LW, et al. (2007) Principles of genome evolution in the *Drosophila melanogaster* species group. *PLoS Biol* 5: e152.
 44. Wesley CS, Eanes WF (1994) Isolation and analysis of the breakpoint sequences of chromosome inversion In(3L)Payne in *Drosophila melanogaster*. *Proc Natl Acad Sci U S A* 91: 3132–3136.
 45. Andolfatto P, Kreitman M (2000) Molecular variation at the In(2L)t proximal breakpoint site in natural populations of *Drosophila melanogaster* and *D. simulans*. *Genetics* 154: 1681–1691.
 46. Matzkin LM, Merritt TJ, Zhu CT, Eanes WF (2005) The structure and population genetics of the breakpoints associated with the cosmopolitan chromosomal inversion In(3R)Payne in *Drosophila melanogaster*. *Genetics* 170: 1143–1152.
 47. Cirera S, Martin-Campos JM, Segarra C, Aguade M (1995) Molecular characterization of the breakpoints of an inversion fixed between *Drosophila melanogaster* and *D. subobscura*. *Genetics* 139: 321–326.
 48. Bergman CM, Pfeiffer BD, Rincon-Limas DE, Hoskins RA, Gnrke A, et al. (2002) Assessing the impact of comparative genomic sequence data on the functional annotation of the *Drosophila* genome. *Genome Biol* 3: RESEARCH0086.
 49. Runcie DE, Noor MA (2009) Sequence signatures of a recent chromosomal rearrangement in *Drosophila mojavensis*. *Genetica* 136: 5–11.
 50. Prazeres da Costa O, Gonzalez J, Ruiz A (2009) Cloning and sequencing of the breakpoint regions of inversion 5g fixed in *Drosophila buzzatii*. *Chromosoma* 118: 349–360.
 51. Kehrer-Sawatzki H, Cooper DN (2008) Molecular mechanisms of chromosomal rearrangement during primate evolution. *Chromosome Res* 16: 41–56.
 52. Armengol L, Pujana MA, Cheung J, Scherer SW, Estivill X (2003) Enrichment of segmental duplications in regions of breaks of synteny between the human and mouse genomes suggest their involvement in evolutionary rearrangements. *Hum Mol Genet* 12: 2201–2208.
 53. Bailey JA, Baertsch R, Kent WJ, Haussler D, Eichler EE (2004) Hotspots of mammalian chromosomal evolution. *Genome Biol* 5: R23.
 54. Murphy WJ, Agarwala R, Schaffer AA, Stephens R, Smith C Jr, et al. (2005) A rhesus macaque radiation hybrid map and comparative analysis with the human genome. *Genomics* 86: 383–395.
 55. Feuk L, MacDonald JR, Tang T, Carson AR, Li M, et al. (2005) Discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee DNA sequence assemblies. *PLoS Genet* 1: e56.
 56. Bailey JA, Eichler EE (2006) Primate segmental duplications: crucibles of evolution, diversity and disease. *Nat Rev Genet* 7: 552–564.
 57. Zody MC, Garber M, Adams DJ, Sharpe T, Harrow J, et al. (2006) DNA sequence of human chromosome 17 and analysis of rearrangement in the human lineage. *Nature* 440: 1045–1049.
 58. Ji X, Zhao S (2008) DA and Xiao-two giant and composite LTR-retrotransposon-like elements identified in the human genome. *Genomics* 91: 249–258.
 59. Coulibaly MB, Lobo NF, Fitzpatrick MC, Kern M, Grushko O, et al. (2007) Segmental duplication implicated in the genesis of inversion 2Rj of *Anopheles gambiae*. *PLoS ONE* 2: e849.
 60. Cáceres M, Puig M, Ruiz A (2001) Molecular characterization of two natural hotspots in the *Drosophila buzzatii* genome induced by transposon insertions. *Genome Res* 11: 1353–1364.
 61. Casals F, Cáceres M, Manfrin MH, Gonzalez J, Ruiz A (2005) Molecular characterization and chromosomal distribution of Galileo, Kepler and Newton, three foldback transposable elements of the *Drosophila buzzatii* species complex. *Genetics* 169: 2047–2059.
 62. Marzo M, Puig M, Ruiz A (2008) The Foldback-like element Galileo belongs to the P superfamily of DNA transposons and is widespread within the *Drosophila* genus. *Proc Natl Acad Sci U S A* 105: 2957–2962.
 63. Ruiz A, Wasserman M (1993) Evolutionary cytogenetics of the *Drosophila buzzatii* species complex. *Heredity* 70 (Pt 6): 582–596.
 64. Ruiz A, Naveira H, Fontdevila A (1984) La historia evolutiva de *Drosophila buzzatii*. IV. Aspectos citogenéticos de su polimorfismo cromosómico. *Genét Ibér* 36: 13–35.
 65. Hasson E, Rodriguez C, Fanara JJ, Naveira H, Reig OA, et al. (1995) The evolutionary history of *Drosophila buzzatii*. XXVI. Macrogeographic patterns of inversion polymorphism in New World populations. *Journal of Evolutionary Biology* 8: 369–384.
 66. Negre B, Ranz JM, Casals F, Cáceres M, Ruiz A (2003) A new split of the Hox gene complex in *Drosophila*: relocation and evolution of the gene labial. *Mol Biol Evol* 20: 2042–2054.
 67. Negre B, Casillas S, Suzanne M, Sanchez-Herrero E, Akam M, et al. (2005) Conservation of regulatory sequences and gene expression patterns in the disintegrating *Drosophila* Hox gene complex. *Genome Res* 15: 692–700.
 68. Laayouni H, Santos M, Fontdevila A (2000) Toward a physical map of *Drosophila buzzatii*. Use of randomly amplified polymorphic dna polymorphisms and sequence-tagged site landmarks. *Genetics* 156: 1797–1816.
 69. González J, Nefedov M, Bosdet I, Casals F, Calvete O, et al. (2005) A BAC-based physical map of the *Drosophila buzzatii* genome. *Genome Res* 15: 885–892.
 70. Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, et al. (2007) Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* 450: 203–218.
 71. Nei M (1987) *Molecular evolutionary genetics*. New York: Columbia University Press.
 72. Kreitman M, Hudson RR (1991) Inferring the evolutionary histories of the Adh and Adh-dup loci in *Drosophila melanogaster* from patterns of polymorphism and divergence. *Genetics* 127: 565–582.
 73. de Freitas Ortiz M, Loreto EL (2009) Characterization of new hAT transposable elements in 12 *Drosophila* genomes. *Genetica* 135: 67–75.
 74. Lim JK, Simmons MJ (1994) Gross chromosome rearrangements mediated by transposable elements in *Drosophila melanogaster*. *Bioessays* 16: 269–275.
 75. Smith JM (1992) Analyzing the mosaic structure of genes. *Journal of Molecular Evolution* 34: 126–129.
 76. Jordan IK, McDonald JF (1998) Evidence for the role of recombination in the regulatory evolution of *Saccharomyces cerevisiae* Ty elements. *J Mol Evol* 47: 14–20.
 77. Charlesworth B, Sniegowski P, Stephan W (1994) The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature* 371: 215–220.
 78. Petrov DA, Aminetzach YT, Davis JC, Bensasson D, Hirsh AE (2003) Size matters: non-LTR retrotransposable elements and ectopic recombination in *Drosophila*. *Mol Biol Evol* 20: 880–892.
 79. Navarro A, Betran E, Barbadiella A, Ruiz A (1997) Recombination and gene flux caused by gene conversion and crossing over in inversion heterokaryotypes. *Genetics* 146: 695–709.
 80. Andolfatto P, Depaulis F, Navarro A (2001) Inversion polymorphisms and nucleotide variability in *Drosophila*. *Genet Res* 77: 1–8.
 81. Yang HP, Barbash DA (2008) Abundant and species-specific DINE-1 transposable elements in 12 *Drosophila* genomes. *Genome Biol* 9: R39.
 82. Feschotte C, Pritham EJ (2007) DNA transposons and the evolution of eukaryotic genomes. *Annu Rev Genet* 41: 331–368.
 83. Casals F, Gonzalez J, Ruiz A (2006) Abundance and chromosomal distribution of six *Drosophila buzzatii* transposons: BuT1, BuT2, BuT3, BuT4, BuT5, and BuT6. *Chromosoma* 115: 403–412.
 84. Kaminker JS, Bergman CM, Kronmiller B, Carlson J, Svirskaas R, et al. (2002) The transposable elements of the *Drosophila melanogaster* euchromatin: a genomics perspective. *Genome Biol* 3: RESEARCH0084.
 85. Bergman CM, Quesneville H, Anxolabehere D, Ashburner M (2006) Recurrent insertion and duplication generate networks of transposable element sequences in the *Drosophila melanogaster* genome. *Genome Biol* 7: R112.
 86. Beall EL, Rio DC (1997) *Drosophila* P-element transposase is a novel site-specific endonuclease. *Genes Dev* 11: 2137–2151.
 87. Tang M, Cecconi C, Bustamante C, Rio DC (2007) Analysis of P element transposase protein-DNA interactions during the early stages of transposition. *J Biol Chem* 282: 29002–29012.
 88. Lobachev KS, Stenger JE, Kozyreva OG, Jurka J, Gordenin DA, et al. (2000) Inverted Alu repeats unstable in yeast are excluded from the human genome. *Embo J* 19: 3822–3830.
 89. Nag DK, Kurst A (1997) A 140-bp-long palindromic sequence induces double-strand breaks during meiosis in the yeast *Saccharomyces cerevisiae*. *Genetics* 146: 835–847.
 90. Lobachev KS, Shor BM, Tran HT, Taylor W, Keen JD, et al. (1998) Factors affecting inverted repeat stimulation of recombination and deletion in *Saccharomyces cerevisiae*. *Genetics* 148: 1507–1524.
 91. Zhou ZH, Akgun E, Jasin M (2001) Repeat expansion by homologous recombination in the mouse germ line at palindromic sequences. *Proc Natl Acad Sci U S A* 98: 8326–8333.
 92. VanHulle K, Lemoine FJ, Narayanan V, Downing B, Hull K, et al. (2007) Inverted DNA repeats channel repair of distant double-strand breaks into chromatid fusions and chromosomal rearrangements. *Mol Cell Biol* 27: 2601–2614.
 93. Lewis SM, Cote AG (2006) Palindromes and genomic stress fractures: bracing and repairing the damage. *DNA Repair (Amst)* 5: 1146–1160.
 94. Eykelenboom JK, Blackwood JK, Okely E, Leach DR (2008) SbcCD causes a double-strand break at a DNA palindrome in the *Escherichia coli* chromosome. *Mol Cell* 29: 644–651.
 95. Voineagu I, Narayanan V, Lobachev KS, Mirkin SM (2008) Replication stalling at unstable inverted repeats: interplay between DNA hairpins and fork stabilizing proteins. *Proc Natl Acad Sci U S A* 105: 9936–9941.
 96. Levis R, Collins M, Rubin GM (1982) FB elements are the common basis for the instability of the wDZL and wC *Drosophila* mutations. *Cell* 30: 551–565.
 97. Smith PA, Corces VG (1991) *Drosophila* transposable elements: mechanisms of mutagenesis and interactions with the host genome. *Adv Genet* 29: 229–300.
 98. Piccinalli RV, Mascord IJ, Barker JS, Oakshott JG, Hasson E (2007) Molecular population genetics of the alpha-esterase5 gene locus in original and colonized populations of *Drosophila buzzatii* and its sibling *Drosophila koepferae*. *J Mol Evol* 64: 158–170.
 99. Rodriguez C, Fanara JJ, Hasson E (1999) Inversion polymorphism, longevity, and body size in a natural population of *Drosophila buzzatii*. *Evolution* 53: 612–620.

100. Ruiz A, Fontdevila A, Santos M, Seoane M, Torroja E (1986) The evolutionary history of *Drosophila buzzatii*. VIII. Evidence for endocyclic selection acting on the inversion polymorphism in a natural population. *Evolution* 40: 740–755.
101. Hasson E, Vilardi JC, Naveira H, Fanara JJ, Rodriguez C, et al. (1991) The evolutionary history of *Drosophila buzzatii*. XVI. Fitness component analysis in a natural population from Argentina. *J Evol Biol* 4: 209–225.
102. Fernandez Iriarte PJ, Norry FM, Hasson ER (2003) Chromosomal inversions effect body size and shape in different breeding resources in *Drosophila buzzatii*. *Heredity* 91: 51–59.
103. Puig M, Cáceres M, Ruiz A (2004) Silencing of a gene adjacent to the breakpoint of a widespread *Drosophila* inversion by a transposon-induced antisense RNA. *Proc Natl Acad Sci U S A* 101: 9013–9018.
104. Hurles ME, Dermizakis ET, Tyler-Smith C (2008) The functional impact of structural variation in humans. *Trends Genet* 24: 238–245.
105. Feschotte C (2008) Transposable elements and the evolution of regulatory networks. *Nat Rev Genet* 9: 397–405.
106. Pereira V, Enard D, Eyre-Walker A (2009) The effect of transposable element insertions on gene expression evolution in rodents. *PLoS ONE* 4: e4321.
107. Montgomery E, Charlesworth B, Langley CH (1987) A test for the role of natural selection in the stabilization of transposable element copy number in a population of *Drosophila melanogaster*. *Genet Res* 49: 31–41.
108. Schaeffer SW, Bhutkar A, McAllister BF, Matsuda M, Matzkin LM, et al. (2008) Polytene chromosomal maps of 11 *Drosophila* species: the order of genomic scaffolds inferred from genetic and physical maps. *Genetics* 179: 1601–1655.
109. Sambrook J, Fritsch EF, Maniatis T (1989) *Molecular Cloning. A laboratory manual*: Cold Spring Harbor Laboratory Press.
110. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32: 1792–1797.
111. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215: 403–410.
112. Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* 19: 2496–2497.
113. Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol* 24: 1596–1599.

Bibliography

- ABRAHAM, J.M., FREITAG, C.S., CLEMENTS, J.R. and EISENSTEIN, B.I. (1985) An invertible element of DNA controls phase variation of type 1 fimbriae of *Escherichia coli*. *Proc Natl Acad Sci U S A* **82**: 5724-5727.
- ACKERMAN, H., UDALOVA, I., HULL, J. and KWIATKOWSKI, D. (2002) Evolution of a polymorphic regulatory element in interferon-gamma through transposition and mutation. *Mol Biol Evol* **19**: 884-890.
- ADAMS, M.D., CELNIKER, S.E., HOLT, R.A., EVANS, C.A., GOCAYNE, J.D., AMANATIDES, P.G., SCHERER, S.E., LI, P.W., HOSKINS, R.A., GALLE, R.F., *et al.* (2000) The genome sequence of *Drosophila melanogaster*. *Science* **287**: 2185-2195.
- AKBARI, O.S., BAE, E., JOHNSEN, H., VILLALUZ, A., WONG, D. and DREWELL, R.A. (2008) A novel promoter-tethering element regulates enhancer-driven gene expression at the bithorax complex in the *Drosophila* embryo. *Development* **135**: 123-131.
- AMINETZACH, Y.T., MACPHERSON, J.M. and PETROV, D.A. (2005) Pesticide resistance via transposition-mediated adaptive gene truncation in *Drosophila*. *Science* **309**: 764-767.
- ANDERSEN, A.A. and PANNING, B. (2003) Epigenetic gene regulation by noncoding RNAs. *Curr Opin Cell Biol* **15**: 281-289.
- ANDERSON, A.R., HOFFMANN, A.A., MCKECHNIE, S.W., UMINA, P.A. and WEEKS, A.R. (2005) The latitudinal cline in the *In(3R)Payne* inversion polymorphism has shifted in the last 20 years in Australian *Drosophila melanogaster* populations. *Mol Ecol* **14**: 851-858.
- ANDOLFATTO, P., WALL, J.D. and KREITMAN, M. (1999) Unusual haplotype structure at the proximal breakpoint of *In(2L)t* in a natural population of *Drosophila melanogaster*. *Genetics* **153**: 1297-1311.
- ARAVIN, A.A., NAUMOVA, N.M., TULIN, A.V., VAGIN, V.V., ROZOVSKY, Y.M. and GVOZDEV, V.A. (2001) Double-stranded RNA-mediated silencing of genomic tandem repeats and transposable elements in the *D. melanogaster* germline. *Curr Biol* **11**: 1017-1027.
- ARAVIN, A.A., KLENOV, M.S., VAGIN, V.V., BANTIGNIES, F., CAVALLI, G. and GVOZDEV, V.A. (2004) Dissection of a natural RNA silencing process in the *Drosophila melanogaster* germ line. *Mol Cell Biol* **24**: 6742-6750.
- ARIKAWA, E., SUN, Y., WANG, J., ZHOU, Q., NING, B., DIAL, S.L., GUO, L. and YANG, J. (2008) Cross-platform comparison of SYBR Green real-time PCR with TaqMan PCR, microarrays and other gene expression measurement technologies evaluated in the MicroArray Quality Control (MAQC) study. *BMC Genomics* **9**: 328.
- ARST, H.N., JR., TOLLERVEY, D., DOWZER, C.E. and KELLY, J.M. (1990) An inversion truncating the *creA* gene of *Aspergillus nidulans* results in carbon catabolite derepression. *Mol Microbiol* **4**: 851-854.
- AULARD, S., MONTI, L., CHAMINADE, N. and LEMEUNIER, F. (2004) Mitotic and polytene chromosomes: comparisons between *Drosophila melanogaster* and *Drosophila simulans*. *Genetica* **120**: 137-150.
- AVEROF, M. and PATEL, N.H. (1997) Crustacean appendage evolution associated with changes in *Hox* gene expression. *Nature* **388**: 682-686.
- AYROLES, J.F., CARBONE, M.A., STONE, E.A., JORDAN, K.W., LYMAN, R.F., MAGWIRE, M.M., ROLLMANN, S.M., DUNCAN, L.H., LAWRENCE, F., ANHOLT, R.R., *et al.* (2009) Systems genetics of complex traits in *Drosophila melanogaster*. *Nat Genet* **41**: 299-307.
- BALANYÀ, J., OLLER, J.M., HUEY, R.B., GILCHRIST, G.W. and SERRA, L. (2006) Global genetic change tracks global climate warming in *Drosophila subobscura*. *Science* **313**: 1773-1775.
- BANSAL, V., BASHIR, A. and BAFNA, V. (2007) Evidence for large inversion polymorphisms in the human genome from HapMap data. *Genome Res* **17**: 219-230.
- BARKER, J.S.F. (1982) Population genetics of *Opuntia* breeding *Drosophila* in Australia. In *Ecological genetics and evolution. The cactus-yeast-Drosophila model system*. (eds. Barker, J.S.F. and Starmer, W.T.), pp. 209-224. Academic Press, Sydney.

- BARKER, J.S.F., SENE, F.M., EAST, P.D. and PEREIRA, M.A.Q.R. (1985) Allozyme and chromosomal polymorphism of *Drosophila buzzatii* in Brazil and Argentina. *Genetica* **67**: 161-170.
- BEJERANO, G., LOWE, C.B., AHITUV, N., KING, B., SIEPEL, A., SALAMA, S.R., RUBIN, E.M., KENT, W.J. and HAUSSLER, D. (2006) A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature* **441**: 87-90.
- BEKPEN, C., MARQUES-BONET, T., ALKAN, C., ANTONACCI, F., LEOGRANDE, M.B., VENTURA, M., KIDD, J.M., SISWARA, P., HOWARD, J.C. and EICHLER, E.E. (2009) Death and resurrection of the human *IRGM* gene. *PLoS Genet* **5**: e1000403.
- BELIZARIO, J.E., ALVES, J., GARAY-MALPARTIDA, M. and OCCHIUCCHI, J.M. (2008) Coupling caspase cleavage and proteasomal degradation of proteins carrying PEST motif. *Curr Protein Pept Sci* **9**: 210-220.
- BELLEN, H.J., LEVIS, R.W., LIAO, G., HE, Y., CARLSON, J.W., TSANG, G., EVANS-HOLM, M., HIESINGER, P.R., SCHULZE, K.L., RUBIN, G.M., *et al.* (2004) The BDGP gene disruption project: single transposon insertions associated with 40% of *Drosophila* genes. *Genetics* **167**: 761-781.
- BETRÁN, E., SANTOS, M. and RUIZ, A. (1998) Antagonistic pleiotropic effect of second-chromosome inversions on body size and early life-history traits in *Drosophila buzzatii*. *Evolution* **52**: 144-154.
- BIRNEY, E., STAMATOYANNOPOULOS, J.A., DUTTA, A., GUIGO, R., GINGERAS, T.R., MARGULIES, E.H., WENG, Z., SNYDER, M., DERMITZAKIS, E.T., THURMAN, R.E., *et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**: 799-816.
- BLANCO, E., PIGNATELLI, M., BELTRAN, S., PUNSET, A., PEREZ-LLUCH, S., SERRAS, F., GUIGO, R. and COROMINAS, M. (2008) Conserved chromosomal clustering of genes governed by chromatin regulators in *Drosophila*. *Genome Biol* **9**: R134.
- BORNBERG-BAUER, E., RIVALS, E. and VINGRON, M. (1998) Computational approaches to identify leucine zippers. *Nucleic Acids Res* **26**: 2740-2746.
- BORSANI, O., ZHU, J., VERSLUES, P.E., SUNKAR, R. and ZHU, J.K. (2005) Endogenous siRNAs derived from a pair of natural *cis*-antisense transcripts regulate salt tolerance in *Arabidopsis*. *Cell* **123**: 1279-1291.
- BOWEN, N.J. and JORDAN, I.K. (2007) Exaptation of protein coding sequences from transposable elements. *Genome Dyn* **3**: 147-162.
- BRAYER, K.J. and SEGAL, D.J. (2008) Keep your fingers off my DNA: protein-protein interactions mediated by C2H2 zinc finger domains. *Cell Biochem Biophys* **50**: 111-131.
- BRITTEN, R.J. (1996a) Cases of ancient mobile element DNA insertions that now affect gene regulation. *Mol Phylogenet Evol* **5**: 13-17.
- BRITTEN, R.J. (1996b) DNA sequence insertion and evolutionary variation in gene regulation. *Proc Natl Acad Sci U S A* **93**: 9374-9377.
- BRITTEN, R.J. (2004) Coding sequences of functioning human genes derived entirely from mobile element sequences. *Proc Natl Acad Sci U S A* **101**: 16825-16830.
- BRONNER, G., TAUBERT, H. and JACKLE, H. (1995) Mesoderm-specific *B104* expression in the *Drosophila* embryo is mediated by internal *cis*-acting elements of the transposon. *Chromosoma* **103**: 669-675.
- BRUDNO, M., DO, C.B., COOPER, G.M., KIM, M.F., DAVYDOV, E., GREEN, E.D., SIDOW, A. and BATZOGLOU, S. (2003) LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res* **13**: 721-731.
- BURKHARD, P., STETEFELD, J. and STRELKOV, S.V. (2001) Coiled coils: a highly versatile protein folding motif. *Trends Cell Biol* **11**: 82-88.
- CÁCERES, M., RANZ, J.M., BARBADILLA, A., LONG, M. and RUIZ, A. (1999) Generation of a widespread *Drosophila* inversion by a transposable element. *Science* **285**: 415-418.
- CÁCERES, M., PUIG, M. and RUIZ, A. (2001) Molecular characterization of two natural hotspots in the *Drosophila buzzatii* genome induced by transposon insertions. *Genome Res* **11**: 1353-1364.
- CÁCERES, M., LACHUER, J., ZAPALA, M.A., REDMOND, J.C., KUDO, L., GESCHWIND, D.H., LOCKHART, D.J., PREUSS, T.M. and BARLOW, C. (2003)

- Elevated gene expression levels distinguish human from non-human primate brains. *Proc Natl Acad Sci U S A* **100**: 13030-13035.
- CALVETE, O. (2010) Dinámica evolutiva de las reordenaciones cromosómicas y coincidencia de los puntos de rotura: análisis molecular de las inversiones fijadas en el cromosoma 2 de *Drosophila buzzatii*. Doctoral thesis. Departament de Genètica i Microbiologia, Universitat Autònoma de Barcelona, Bellaterra, Spain.
- CAMPANELLA, J.J., BITINCKA, L. and SMALLEY, J. (2003) MatGAT: an application that generates similarity/identity matrices using protein or DNA sequences. *BMC Bioinformatics* **4**: 29.
- CANALES, R.D., LUO, Y., WILLEY, J.C., AUSTERMILLER, B., BARBACIORU, C.C., BOYSEN, C., HUNKAPILLER, K., JENSEN, R.V., KNIGHT, C.R., LEE, K.Y., *et al.* (2006) Evaluation of DNA microarray results with quantitative gene expression platforms. *Nat Biotechnol* **24**: 1115-1122.
- CARNINCI, P., SANDELIN, A., LENHARD, B., KATAYAMA, S., SHIMOKAWA, K., PONJAVIC, J., SEMPLE, C.A., TAYLOR, M.S., ENGSTROM, P.G., FRITH, M.C., *et al.* (2006) Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet* **38**: 626-635.
- CARROLL, S.B. (2005) Evolution at two levels: on genes and form. *PLoS Biol* **3**: e245.
- CARTHEW, R.W. (2003) RNAi applications in *Drosophila melanogaster*. In *RNAi: a guide to gene silencing* (ed. Hannon, G.J.), pp. 361-400. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
- CARTHEW, R.W. and SONTHEIMER, E.J. (2009) Origins and Mechanisms of miRNAs and siRNAs. *Cell* **136**: 642-655.
- CASALS, F., CÁCERES, M. and RUIZ, A. (2003) The foldback-like transposon *Galileo* is involved in the generation of two different natural chromosomal inversions of *Drosophila buzzatii*. *Mol Biol Evol* **20**: 674-685.
- CASALS, F., CÁCERES, M., MANFRIN, M.H., GONZÁLEZ, J. and RUIZ, A. (2005) Molecular characterization and chromosomal distribution of *Galileo*, *Kepler* and *Newton*, three foldback transposable elements of the *Drosophila buzzatii* species complex. *Genetics* **169**: 2047-2059.
- CASTERMANS, D., VERMEESCH, J.R., FRYNS, J.P., STEYAERT, J.G., VAN DE VEN, W.J., CREEMERS, J.W. and DEVRIENDT, K. (2007) Identification and characterization of the *TRIP8* and *REEP3* genes on chromosome 10q21.3 as novel candidate genes for autism. *Eur J Hum Genet* **15**: 422-431.
- CAWLEY, S., BEKIRANOV, S., NG, H.H., KAPRANOV, P., SEKINGER, E.A., KAMPA, D., PICCOLBONI, A., SEMENTCHENKO, V., CHENG, J., WILLIAMS, A.J., *et al.* (2004) Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* **116**: 499-509.
- CELNIKER, S.E., DILLON, L.A., GERSTEIN, M.B., GUNSALUS, K.C., HENIKOFF, S., KARPEN, G.H., KELLIS, M., LAI, E.C., LIEB, J.D., MACALPINE, D.M., *et al.* (2009) Unlocking the secrets of the genome. *Nature* **459**: 927-930.
- CERDEÑO-TÁRRAGA, A.M., PATRICK, S., CROSSMAN, L.C., BLAKELY, G., ABRATT, V., LENNARD, N., POXTON, I., DUERDEN, B., HARRIS, B., QUAIL, M.A., *et al.* (2005) Extensive DNA inversions in the *B. fragilis* genome control variable gene expression. *Science* **307**: 1463-1465.
- CLARK, A.G., EISEN, M.B., SMITH, D.R., BERGMAN, C.M., OLIVER, B., MARKOW, T.A., KAUFMAN, T.C., KELLIS, M., GELBART, W., IYER, V.N., *et al.* (2007) Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* **450**: 203-218.
- CLARK, R.M., WAGLER, T.N., QUIJADA, P. and DOEBLEY, J. (2006) A distant upstream enhancer at the maize domestication gene *tb1* has pleiotropic effects on plant and inflorescent architecture. *Nat Genet* **38**: 594-597.
- CONNE, B., STUTZ, A. and VASSALLI, J.D. (2000) The 3' untranslated region of messenger RNA: A molecular 'hotspot' for pathology? *Nat Med* **6**: 637-641.
- CORTESE, M.D., NORRY, F.M., PICCINALI, R. and HASSON, E. (2002) Direct and correlated responses to artificial selection on developmental time and wing length in *Drosophila buzzatii*. *Evolution* **56**: 2541-2547.
- COULOMBE-HUNTINGTON, J. and MAJEWSKI, J. (2007) Intron loss and gain in *Drosophila*. *Mol Biol Evol* **24**: 2842-2850.

- CHEN, J., SUN, M., KENT, W.J., HUANG, X., XIE, H., WANG, W., ZHOU, G., SHI, R.Z. and ROWLEY, J.D. (2004) Over 20% of human transcripts might form sense-antisense pairs. *Nucleic Acids Res* **32**: 4812-4820.
- CHINTAPALLI, V.R., WANG, J. and DOW, J.A. (2007) Using FlyAtlas to identify better *Drosophila melanogaster* models of human disease. *Nat Genet* **39**: 715-720.
- CHUNG, H., BOGWITZ, M.R., MCCART, C., ANDRIANOPOULOS, A., FFRENCH-CONSTANT, R.H., BATTERHAM, P. and DABORN, P.J. (2007) *Cis*-regulatory elements in the *Accord* retrotransposon result in tissue-specific expression of the *Drosophila melanogaster* insecticide resistance gene *Cyp6g1*. *Genetics* **175**: 1071-1077.
- CHUNG, W.J., OKAMURA, K., MARTIN, R. and LAI, E.C. (2008) Endogenous RNA interference provides a somatic defense against *Drosophila* transposons. *Curr Biol* **18**: 795-802.
- DATHAN, N., ZACCARO, L., ESPOSITO, S., ISERNIA, C., OMICHINSKI, J.G., RICCIO, A., PEDONE, C., DI BLASIO, B., FATTORUSSO, R. and PEDONE, P.V. (2002) The *Arabidopsis* SUPERMAN protein is able to specifically bind DNA through its single Cys2-His2 zinc finger motif. *Nucleic Acids Res* **30**: 4945-4951.
- DE, S.K., MCMASTER, M.T. and ANDREWS, G.K. (1990) Endotoxin induction of murine metallothionein gene expression. *J Biol Chem* **265**: 15267-15274.
- DEININGER, P.L. and BATZER, M.A. (1999) *Alu* repeats and human disease. *Mol Genet Metab* **67**: 183-193.
- DELPRAT, A., NEGRE, B., PUIG, M. and RUIZ, A. (2009) The transposon *Galileo* generates natural chromosomal inversions in *Drosophila* by ectopic recombination. *PLoS One* **4**: e7883.
- DENNIS, G., JR., SHERMAN, B.T., HOSACK, D.A., YANG, J., GAO, W., LANE, H.C. and LEMPICKI, R.A. (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* **4**: P3.
- DIMOVA, D.K., STEVAUX, O., FROLOV, M.V. and DYSON, N.J. (2003) Cell cycle-dependent and cell cycle-independent control of transcription by the *Drosophila* E2F/RB pathway. *Genes Dev* **17**: 2308-2320.
- DOBZHANSKY, T. (1970) *Genetics of the evolutionary process*. Columbia University Press, New York.
- DOMBRÁDI, V., AXTON, J.M., BREWIS, N.D., DA CRUZ E SILVA, E.F., ALPHEY, L. and COHEN, P.T. (1990) *Drosophila* contains three genes that encode distinct isoforms of protein phosphatase 1. *Eur J Biochem* **194**: 739-745.
- DONG, J., FELDMANN, G., HUANG, J., WU, S., ZHANG, N., COMERFORD, S.A., GAYYED, M.F., ANDERS, R.A., MAITRA, A. and PAN, D. (2007) Elucidation of a universal size-control mechanism in *Drosophila* and mammals. *Cell* **130**: 1120-1133.
- DOPMAN, E.B. and HARTL, D.L. (2007) A portrait of copy-number polymorphism in *Drosophila melanogaster*. *Proc Natl Acad Sci U S A* **104**: 19920-19925.
- DOYLE, J.J. and DICKSON, E.E. (1987) Preservation of plant samples for DNA restriction endonuclease analysis. *Taxon* **36**: 715-722.
- DUHRING, U., AXMANN, I.M., HESS, W.R. and WILDE, A. (2006) An internal antisense RNA regulates expression of the photosynthesis gene *isi4*. *Proc Natl Acad Sci U S A* **103**: 7054-7058.
- DUMAN-SCHEEL, M., WENG, L., XIN, S. and DU, W. (2002) Hedgehog regulates cell growth and proliferation by inducing Cyclin D and Cyclin E. *Nature* **417**: 299-304.
- DUNN, C.A., VAN DE LAGEMAAT, L.N., BAILLIE, G.J. and MAGER, D.L. (2005) Endogenous retrovirus long terminal repeats as ready-to-use mobile promoters: the case of primate *beta3GAL-T5*. *Gene* **364**: 2-12.
- DYE, M.J. and PROUDFOOT, N.J. (2001) Multiple transcript cleavage precedes polymerase release in termination by RNA polymerase II. *Cell* **105**: 669-681.
- EDGAR, B.A. and NIJHOUT, H.F. (2004) Growth and cell cycle control in *Drosophila*. In *Cell Growth: control of cell size* (eds. Hall, M.N., Raff, M. and Thomas, G.), pp. 23-83. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
- EDGAR, B.A. (2006) How flies get their size: genetics meets physiology. *Nat Rev Genet* **7**: 907-916.
- EDGAR, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**: 1792-1797.

- EISEN, M.B., SPELLMAN, P.T., BROWN, P.O. and BOTSTEIN, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* **95**: 14863-14868.
- EMERY, P., STRUBIN, M., HOFMANN, K., BUCHER, P., MACH, B. and REITH, W. (1996) A consensus motif in the RFX DNA binding domain and binding domain mutants with altered specificity. *Mol Cell Biol* **16**: 4486-4494.
- ETIENNE, W., MEYER, M.H., PEPPERS, J. and MEYER, R.A., JR. (2004) Comparison of mRNA gene expression by RT-PCR and DNA microarray. *Biotechniques* **36**: 618-620, 622, 624-616.
- FAHRENKROG, B. and AEBI, U. (2003) The nuclear pore complex: nucleocytoplasmic transport and beyond. *Nat Rev Mol Cell Biol* **4**: 757-766.
- FANTAUZZO, K.A., TADIN-STRAPPS, M., YOU, Y., MENTZER, S.E., BAUMEISTER, F.A., CIANFARANI, S., VAN MALDERGEM, L., WARBURTON, D., SUNDBERG, J.P. and CHRISTIANO, A.M. (2008) A position effect on *TRPS1* is associated with Ambras syndrome in humans and the *Koala* phenotype in mice. *Hum Mol Genet* **17**: 3539-3551.
- FAULKNER, G.J., KIMURA, Y., DAUB, C.O., WANI, S., PLESSY, C., IRVINE, K.M., SCHRODER, K., CLOONAN, N., STEPTOE, A.L., LASSMANN, T., *et al.* (2009) The regulated retrotransposon transcriptome of mammalian cells. *Nat Genet* **41**: 563-571.
- FERNÁNDEZ IRIARTE, P. and HASSON, E. (2000) The role of the use of different host plants in the maintenance of the inversion polymorphism in the cactophilic *Drosophila buzzatii*. *Evolution* **54**: 1295-1302.
- FERRIGNO, O., VIROLLE, T., DJABARI, Z., ORTONNE, J.P., WHITE, R.J. and ABERDAM, D. (2001) Transposable B2 SINE elements can provide mobile RNA polymerase II promoters. *Nat Genet* **28**: 77-81.
- FESCHOTTE, C. (2008) Transposable elements and the evolution of regulatory networks. *Nat Rev Genet* **9**: 397-405.
- FEUK, L., MACDONALD, J.R., TANG, T., CARSON, A.R., LI, M., RAO, G., KHAJA, R. and SCHERER, S.W. (2005) Discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee DNA sequence assemblies. *PLoS Genet* **1**: e56.
- FISHER, R.A. (1930) *The genetical theory of natural selection*. Clarendon Press, Oxford.
- FONTDEVILA, A., RUIZ, A., ALONSO, G. and OCAÑA, J. (1981) Evolutionary history of *Drosophila buzzatii*. I. Natural chromosomal polymorphism in colonized populations of the old world. *Evolution* **35**: 148-157.
- FONTDEVILA, A., RUIZ, A., OCAÑA, J. and ALONSO, G. (1982) Evolutionary history of *Drosophila buzzatii*. II. How much has chromosomal polymorphism changed in colonization? *Evolution* **36**: 843-851.
- FRAZER, K.A., PACTER, L., POLIAKOV, A., RUBIN, E.M. and DUBCHAK, I. (2004) VISTA: computational tools for comparative genomics. *Nucleic Acids Res* **32**: W273-279.
- FRISCHER, L.E., HAGEN, F.S. and GARBER, R.L. (1986) An inversion that disrupts the *Antennapedia* gene causes abnormal structure and localization of RNAs. *Cell* **47**: 1017-1023.
- FUKAMI, M., KATO, F., TAJIMA, T., YOKOYA, S. and OGATA, T. (2006) Transactivation function of an approximately 800-bp evolutionarily conserved sequence at the *SHOX* 3' region: implication for the downstream enhancer. *Am J Hum Genet* **78**: 167-170.
- GAHAN, L.J., GOULD, F. and HECKEL, D.G. (2001) Identification of a gene associated with Bt resistance in *Heliothis virescens*. *Science* **293**: 857-860.
- GAMA-CARVALHO, M. and CARMO-FONSECA, M. (2001) The rules and roles of nucleocytoplasmic shuttling proteins. *FEBS Lett* **498**: 157-163.
- GDULA, D.A., GERASIMOVA, T.I. and CORCES, V.G. (1996) Genetic and molecular analysis of the *gpy* chromatin insulator of *Drosophila*. *Proc Natl Acad Sci U S A* **93**: 9378-9383.
- GENTLEMAN, R.C., CAREY, V.J., BATES, D.M., BOLSTAD, B., DETTLING, M., DUDOIT, S., ELLIS, B., GAUTIER, L., GE, Y., GENTRY, J., *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5**: R80.
- GERSHENZON, N.I., TRIFONOV, E.N. and IOSHIKHES, I.P. (2006) The features of *Drosophila* core promoters revealed by statistical analysis. *BMC Genomics* **7**: 161.

- GILBERT, D.G. (2007) DroSpeGe: rapid access database for new *Drosophila* species genomes. *Nucleic Acids Res* **35**: D480-485.
- GIROUX, M.J., CLANCY, M., BAIER, J., INGHAM, L., MCCARTY, D. and HANNAH, L.C. (1994) *De novo* synthesis of an intron by the maize transposable element *Dissociation*. *Proc Natl Acad Sci U S A* **91**: 12150-12154.
- GIRTON, J.R. and JOHANSEN, K.M. (2008) Chromatin structure and the regulation of gene expression: the lessons of PEV in *Drosophila*. *Adv Genet* **61**: 1-43.
- GOMPEL, N., PRUD'HOMME, B., WITTKOPP, P.J., KASSNER, V.A. and CARROLL, S.B. (2005) Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in *Drosophila*. *Nature* **433**: 481-487.
- GONZÁLEZ, J., NEFEDOV, M., BOSDET, I., CASALS, F., CALVETE, O., DELPRAT, A., SHIN, H., CHIU, R., MATHEWSON, C., WYE, N., *et al.* (2005) A BAC-based physical map of the *Drosophila buzzatii* genome. *Genome Res* **15**: 885-892.
- HALL, T.A. (1999) BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp Ser* **41**: 95-98.
- HALL, T.M. (2005) Multiple modes of RNA recognition by zinc finger proteins. *Curr Opin Struct Biol* **15**: 367-373.
- HARENDZA, C.J. and JOHNSON, L.F. (1990) Polyadenylation signal of the mouse thymidylate synthase gene was created by insertion of an *L1* repetitive element downstream of the open reading frame. *Proc Natl Acad Sci U S A* **87**: 2531-2535.
- HASSON, E., NAVEIRA, H. and FONTDEVILA, A. (1992) The breeding sites of Argentinian cactophilic species of the *Drosophila mulleri* complex (subgenus *Drosophila* - *repleta* group). *Rev. chil. hist. nat.* **65**: 319-326.
- HASSON, E., RODRÍGUEZ, C., FANARA, J.J., NAVEIRA, H., REIG, O.A. and FONTDEVILA, A. (1995) The evolutionary history of *Drosophila buzzatii*. XXVI. Macrogeographic patterns of inversion polymorphism in New World populations *J Evol Biol* **8**: 369-384.
- HASTINGS, M.L., MILCAREK, C., MARTINCIC, K., PETERSON, M.L. and MUNROE, S.H. (1997) Expression of the thyroid hormone receptor gene, *erbAa*, in B lymphocytes: alternative mRNA processing is independent of differentiation but correlates with antisense RNA levels. *Nucleic Acids Res* **25**: 4296-4300.
- HAYAKAWA, T., SATTI, Y., GAGNEUX, P., VARKI, A. and TAKAHATA, N. (2001) *Alu*-mediated inactivation of the human CMP-N-acetylneuraminic acid hydroxylase gene. *Proc Natl Acad Sci U S A* **98**: 11399-11404.
- HAYNES, K.A., CAUDY, A.A., COLLINS, L. and ELGIN, S.C. (2006) Element *1360* and RNAi components contribute to HP1-dependent silencing of a pericentric reporter. *Curr Biol* **16**: 2222-2227.
- HIROSE, F., YAMAGUCHI, M., HANDA, H., INOMATA, Y. and MATSUKAGE, A. (1993) Novel 8-base pair sequence (*Drosophila* DNA replication-related element) and specific binding factor involved in the expression of *Drosophila* genes for DNA polymerase alpha and proliferating cell nuclear antigen. *J Biol Chem* **268**: 2092-2099.
- HOCHHEIMER, A., ZHOU, S., ZHENG, S., HOLMES, M.C. and TJIAN, R. (2002) TRF2 associates with DREF and directs promoter-selective gene expression in *Drosophila*. *Nature* **420**: 439-445.
- HOFFMANN, A.A., SGRÒ, C.M. and WEEKS, A.R. (2004) Chromosomal inversion polymorphisms and adaptation. *Trends Ecol Evol* **19**: 482-488.
- HOFFMANN, A.A. and RIESEBERG, L.H. (2008) Revisiting the Impact of Inversions in Evolution: From Population Genetic Markers to Drivers of Adaptive Shifts and Speciation? *Annu Rev Ecol Evol Syst* **39**: 21-42.
- HONGAY, C.F., GRISAFI, P.L., GALITSKI, T. and FINK, G.R. (2006) Antisense transcription controls cell fate in *Saccharomyces cerevisiae*. *Cell* **127**: 735-745.
- HOUGH, R.B., LENGELING, A., BEDIAN, V., LO, C. and BUCAN, M. (1998) *Rump white* inversion in the mouse disrupts dipeptidyl aminopeptidase-like protein 6 and causes dysregulation of *Kit* expression. *Proc Natl Acad Sci U S A* **95**: 13800-13805.
- HUANG, D.W., SHERMAN, B.T. and LEMPICKI, R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**: 44-57.
- IMAMURA, T., YAMAMOTO, S., OHGANE, J., HATTORI, N., TANAKA, S. and SHIOTA, K. (2004) Non-

- coding RNA directed DNA demethylation of *Sphk1* CpG island. *Biochem Biophys Res Commun* **322**: 593-600.
- INGHAM, P.W. and MCMAHON, A.P. (2001) Hedgehog signaling in animal development: paradigms and principles. *Genes Dev* **15**: 3059-3087.
- IRIZARRY, R.A., HOBBS, B., COLLIN, F., BEAZER-BARCLAY, Y.D., ANTONELLIS, K.J., SCHERF, U. and SPEED, T.P. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**: 249-264.
- JANS, D.A., XIAO, C.Y. and LAM, M.H. (2000) Nuclear targeting signal recognition: a key control point in nuclear transport? *Bioessays* **22**: 532-544.
- JARMAN, A.P., GRAU, Y., JAN, L.Y. and JAN, Y.N. (1993) *atonal* is a proneural gene that directs chordotonal organ formation in the *Drosophila* peripheral nervous system. *Cell* **73**: 1307-1321.
- JORDAN, I.K., ROGOZIN, I.B., GLAZKO, G.V. and KOONIN, E.V. (2003) Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends Genet* **19**: 68-72.
- KALDERON, D., ROBERTS, B.L., RICHARDSON, W.D. and SMITH, A.E. (1984) A short amino acid sequence able to specify nuclear location. *Cell* **39**: 499-509.
- KAPITONOV, V.V. and JURKA, J. (1999) The long terminal repeat of an endogenous retrovirus induces alternative splicing and encodes an additional carboxy-terminal sequence in the human leptin receptor. *J Mol Evol* **48**: 248-251.
- KASHKUSH, K., FELDMAN, M. and LEVY, A.A. (2003) Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat. *Nat Genet* **33**: 102-106.
- KATAYAMA, S., TOMARU, Y., KASUKAWA, T., WAKI, K., NAKANISHI, M., NAKAMURA, M., NISHIDA, H., YAP, C.C., SUZUKI, M., KAWAI, J., *et al.* (2005) Antisense transcription in the mammalian transcriptome. *Science* **309**: 1564-1566.
- KATZENBERGER, R.J., MARENGO, M.S. and WASSARMAN, D.A. (2009) Control of alternative splicing by signal-dependent degradation of splicing-regulatory proteins. *J Biol Chem* **284**: 10737-10746.
- KAZAZIAN, H.H., JR. (1998) Mobile elements and disease. *Curr Opin Genet Dev* **8**: 343-350.
- KEHRER-SAWATZKI, H., SANDIG, C., CHUZHANOVA, N., GOIDTS, V., SZAMALEK, J.M., TANZER, S., MULLER, S., PLATZER, M., COOPER, D.N. and HAMEISTER, H. (2005) Breakpoint analysis of the pericentric inversion distinguishing human chromosome 4 from the homologous chromosome in the chimpanzee (*Pan troglodytes*). *Hum Mutat* **25**: 45-55.
- KEHRER-SAWATZKI, H. and COOPER, D.N. (2008) Molecular mechanisms of chromosomal rearrangement during primate evolution. *Chromosome Res* **16**: 41-56.
- KENNINGTON, W.J., HOFFMANN, A.A. and PARTRIDGE, L. (2007) Mapping regions within cosmopolitan inversion *In(3R)Payne* associated with natural variation in body size in *Drosophila melanogaster*. *Genetics* **177**: 549-556.
- KERBER, B., FELLERT, S., TAUBERT, H. and HOCH, M. (1996) Germ line and embryonic expression of *Fex*, a member of the *Drosophila* F-element retrotransposon family, is mediated by an internal *cis*-regulatory control region. *Mol Cell Biol* **16**: 2998-3007.
- KIDD, J.M., COOPER, G.M., DONAHUE, W.F., HAYDEN, H.S., SAMPAS, N., GRAVES, T., HANSEN, N., TEAGUE, B., ALKAN, C., ANTONACCI, F., *et al.* (2008) Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**: 56-64.
- KIDWELL, M.G. and LISCH, D. (1997) Transposable elements as sources of variation in animals and plants. *Proc Natl Acad Sci U S A* **94**: 7704-7711.
- KIDWELL, M.G. and LISCH, D.R. (2001) Perspective: transposable elements, parasitic DNA, and genome evolution. *Evolution* **55**: 1-24.
- KIGER, A.A., GIGLIOTTI, S. and FULLER, M.T. (1999) Developmental genetics of the essential *Drosophila* nucleoporin *nup154*: allelic differences due to an outward-directed promoter in the *P*-element 3' end. *Genetics* **153**: 799-812.
- KING, M.C. and WILSON, A.C. (1975) Evolution at two levels in humans and chimpanzees. *Science* **188**: 107-116.
- KIRKPATRICK, M. and BARTON, N. (2006) Chromosome inversions, local adaptation and speciation. *Genetics* **173**: 419-434.
- KIRKPATRICK, M. (2010) How and why chromosome inversions evolve. *PLoS Biol* **8**.

- KLEINJAN, D.J. and VAN HEYNINGEN, V. (1998) Position effect in human genetic disease. *Hum Mol Genet* **7**: 1611-1618.
- KLUG, A. and SCHWABE, J.W. (1995) Protein motifs 5. Zinc fingers. *FASEB J* **9**: 597-604.
- KNOBLICH, J.A., SAUER, K., JONES, L., RICHARDSON, H., SAINT, R. and LEHNER, C.F. (1994) Cyclin E controls S phase progression and its down-regulation during *Drosophila* embryogenesis is required for the arrest of cell proliferation. *Cell* **77**: 107-120.
- KOMATSU, A., OTSUKA, A. and ONO, M. (2002) Novel regulatory regions found downstream of the rat B29/Ig- β gene. *Eur J Biochem* **269**: 1227-1236.
- KORBEL, J.O., URBAN, A.E., AFFOURTIT, J.P., GODWIN, B., GRUBERT, F., SIMONS, J.F., KIM, P.M., PALEJEV, D., CARRIERO, N.J., DU, L., *et al.* (2007) Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**: 420-426.
- KORNEEV, S.A., PARK, J.H. and O'SHEA, M. (1999) Neuronal expression of neural nitric oxide synthase (nNOS) protein is suppressed by an antisense RNA transcribed from an NOS pseudogene. *J Neurosci* **19**: 7711-7720.
- KRAMER, C., LOROS, J.J., DUNLAP, J.C. and CROSTHWAITE, S.K. (2003) Role for antisense RNA in regulating circadian clock function in *Neurospora crassa*. *Nature* **421**: 948-952.
- KRIMBAS, C.B. and POWELL, J.R. (1992) *Drosophila inversion polymorphism*. CRC Press, Boca Raton, Florida.
- LA COUR, T., KIEMER, L., MOLGAARD, A., GUPTA, R., SKRIVER, K. and BRUNAK, S. (2004) Analysis and prediction of leucine-rich nuclear export signals. *Protein Eng Des Sel* **17**: 527-536.
- LAAYOUNI, H., HASSON, E., SANTOS, M. and FONTDEVILA, A. (2003) The evolutionary history of *Drosophila buzzatii*. XXXV. Inversion polymorphism and nucleotide variability in different regions of the second chromosome. *Mol Biol Evol* **20**: 931-944.
- LAKICH, D., KAZAZIAN, H.H., JR., ANTONARAKIS, S.E. and GITSCHIER, J. (1993) Inversions disrupting the factor VIII gene are a common cause of severe haemophilia A. *Nat Genet* **5**: 236-241.
- LANDRY, J.R., MEDSTRAND, P. and MAGER, D.L. (2001) Repetitive elements in the 5' untranslated region of a human zinc-finger gene modulate transcription and translation efficiency. *Genomics* **76**: 110-116.
- LAPIDOT, M. and PILPEL, Y. (2006) Genome-wide natural antisense transcription: coupling its regulation to its different regulatory mechanisms. *EMBO Rep* **7**: 1216-1222.
- LARKIN, M.A., BLACKSHIELDS, G., BROWN, N.P., CHENNA, R., MCGETTIGAN, P.A., MCWILLIAM, H., VALENTIN, F., WALLACE, I.M., WILM, A., LOPEZ, R., *et al.* (2007) ClustalW and ClustalX version 2.0. *Bioinformatics* **23**: 2947-2948.
- LEEB, M., STEFFEN, P.A. and WUTZ, A. (2009) X chromosome inactivation sparked by non-coding RNAs. *RNA Biol* **6**: 94-99.
- LEHMANN, R. and TAUTZ, D. (1994) *In situ* hybridization to RNA. In *Drosophila melanogaster: Practical Uses in Cellular and Molecular Biology*, vol. 44 (eds. Goldstein, L.S.B. and Fyberg, E.A.), pp. 576-597. Academic Press, New York.
- LERMAN, D.N. and FEDER, M.E. (2005) Naturally occurring transposable elements disrupt *hsp70* promoter function in *Drosophila melanogaster*. *Mol Biol Evol* **22**: 776-783.
- LEV-MAOR, G., SOREK, R., SHOMRON, N. and AST, G. (2003) The birth of an alternatively spliced exon: 3' splice-site selection in *Alu* exons. *Science* **300**: 1288-1291.
- LEVITAN, M. (2001) Studies of linkage in populations. XIV. Historical changes in frequencies of gene arrangements and arrangement combinations in natural populations of *Drosophila robusta*. *Evolution* **55**: 2359-2362.
- LEVY, S., SUTTON, G., NG, P.C., FEUK, L., HALPERN, A.L., WALENZ, B.P., AXELROD, N., HUANG, J., KIRKNESS, E.F., DENISOV, G., *et al.* (2007) The diploid genome sequence of an individual human. *PLoS Biol* **5**: e254.
- LEWIS, E.B. (1950) The phenomenon of position effect. *Adv Genet* **3**: 73-115.
- LEWONTIN, R.C. (1965) Selection for colonizing ability. In *The genetics of colonizing species* (eds. Baker, H.G. and Stebbins, G.L.), pp. 79-94. Academic Press, New York.
- LI, C. and WONG, W.H. (2001) Model-based analysis of oligonucleotide arrays: expression index

- computation and outlier detection. *Proc Natl Acad Sci U S A* **98**: 31-36.
- LISTER, C., JACKSON, D. and MARTIN, C. (1993) Transposon-induced inversion in *Antirrhinum* modifies *nivea* gene expression to give a novel flower color pattern under the control of *cycloidea^{radialis}*. *Plant Cell* **5**: 1541-1553.
- LITTLETON, J.T. and GANETZKY, B. (2000) Ion channels and synaptic organization: analysis of the *Drosophila* genome. *Neuron* **26**: 35-43.
- LIU, M., LEE, B.H. and MATHEWS, M.B. (1999) Involvement of RFX1 protein in the regulation of the human proliferating cell nuclear antigen promoter. *J Biol Chem* **274**: 15433-15439.
- LOOTS, G.G., OVCHARENKO, I., PACTER, L., DUBCHAK, I. and RUBIN, E.M. (2002) rVista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Res* **12**: 832-839.
- LOWE, C.B., BEJERANO, G. and HAUSSLER, D. (2007) Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. *Proc Natl Acad Sci U S A* **104**: 8005-8010.
- LOWRY, D.B. and WILLIS, J.H. (2010) A widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation. *PLoS Biol* **8**.
- LUKOV, G.L. and GOODELL, M.A. (2010) LYL1 degradation by the proteasome is directed by a N-terminal PEST rich site in a phosphorylation-independent manner. *PLoS One* **5**.
- LUPAS, A., VAN DYKE, M. and STOCK, J. (1991) Predicting coiled coils from protein sequences. *Science* **252**: 1162-1164.
- MALLET, F., BOUTON, O., PRUDHOMME, S., CHEYNET, V., ORIOL, G., BONNAUD, B., LUCOTTE, G., DURET, L. and MANDRAND, B. (2004) The endogenous retroviral locus ERVWE1 is a bona fide gene involved in hominoid placental physiology. *Proc Natl Acad Sci U S A* **101**: 1731-1736.
- MARGULIES, M., EGHOLM, M., ALTMAN, W.E., ATTIYA, S., BADER, J.S., BEMBEN, L.A., BERKA, J., BRAVERMAN, M.S., CHEN, Y.J., CHEN, Z., et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**: 376-380.
- MARZO, M., PUIG, M. and RUIZ, A. (2008) The *Foldback*-like element *Galileo* belongs to the P superfamily of DNA transposons and is widespread within the *Drosophila* genus. *Proc Natl Acad Sci U S A* **105**: 2957-2962.
- MASON, D.A. and GOLDFARB, D.S. (2009) The nuclear transport machinery as a regulator of *Drosophila* development. *Semin Cell Dev Biol* **20**: 582-589.
- MATHIOPOULOS, K.D., DELLA TORRE, A., PREDAZZI, V., PETRARCA, V. and COLUZZI, M. (1998) Cloning of inversion breakpoints in the *Anopheles gambiae* complex traces a transposable element at the inversion junction. *Proc Natl Acad Sci U S A* **95**: 12444-12449.
- MAZO, A., HODGSON, J.W., PETRUK, S., SEDKOV, Y. and BROCK, H.W. (2007) Transcriptional interference: an unexpected layer of complexity in gene regulation. *J Cell Sci* **120**: 2755-2761.
- MCDONALD, J.F. (1995) Transposable elements: possible catalysts of organismic evolution. *Trends Ecol Evol* **10**: 123-126.
- MCGINNIS, S. and MADDEN, T.L. (2004) BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res* **32**: W20-25.
- MEDSTRAND, P., LANDRY, J.R. and MAGER, D.L. (2001) Long terminal repeats are used as alternative promoters for the endothelin B receptor and apolipoprotein C-I genes in humans. *J Biol Chem* **276**: 1896-1903.
- MEDSTRAND, P., VAN DE LAGEMAAT, L.N., DUNN, C.A., LANDRY, J.R., SVENBACK, D. and MAGER, D.L. (2005) Impact of transposable elements on the evolution of mammalian gene regulation. *Cytogenet Genome Res* **110**: 342-352.
- MICALI, C.O. and SMITH, M.L. (2006) A nonself recognition gene complex in *Neurospora crassa*. *Genetics* **173**: 1991-2004.
- MIGNONE, F., GISSI, C., LIUNI, S. and PESOLE, G. (2002) Untranslated regions of mRNAs. *Genome Biol* **3**: REVIEWS0004.
- MILLER, W.J., MCDONALD, J.F., NOUAUD, D. and ANXOLABEHERE, D. (1999) Molecular domestication--more than a sporadic episode in evolution. *Genetica* **107**: 197-207.

- MISRA, S., CROSBY, M.A., MUNGALL, C.J., MATTHEWS, B.B., CAMPBELL, K.S., HRADECKY, P., HUANG, Y., KAMINKER, J.S., MILLBURN, G.H., PROCHNIK, S.E., *et al.* (2002) Annotation of the *Drosophila melanogaster* euchromatic genome: a systematic review. *Genome Biol* **3**: RESEARCH0083.
- MOORE, F.L., JARUZELSKA, J., DORFMAN, D.M. and REIJO-PERA, R.A. (2004) Identification of a novel gene, *DZIP* (DAZ-interacting protein), that encodes a protein that interacts with DAZ (deleted in azoospermia) and is expressed in embryonic stem cells and germ cells. *Genomics* **83**: 834-843.
- MOREY, J.S., RYAN, J.C. and VAN DOLAH, F.M. (2006) Microarray validation: factors influencing correlation between oligonucleotide microarrays and real-time PCR. *Biol Proced Online* **8**: 175-193.
- MORGAN, H.D., SUTHERLAND, H.G., MARTIN, D.I. and WHITELAW, E. (1999) Epigenetic inheritance at the *agouti* locus in the mouse. *Nat Genet* **23**: 314-318.
- MOUNT, S.M. and SALZ, H.K. (2000) Pre-messenger RNA processing factors in the *Drosophila* genome. *J Cell Biol* **150**: F37-44.
- MUNROE, S.H. and ZHU, J. (2006) Overlapping transcripts, double-stranded RNA and antisense regulation: a genomic perspective. *Cell Mol Life Sci* **63**: 2102-2118.
- MYERS, A.J., GIBBS, J.R., WEBSTER, J.A., ROHRER, K., ZHAO, A., MARLOWE, L., KALEEM, M., LEUNG, D., BRYDEN, L., NATH, P., *et al.* (2007) A survey of genetic human cortical gene expression. *Nat Genet* **39**: 1494-1499.
- NAKAI, K. and KANEHISA, M. (1992) A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics* **14**: 897-911.
- NARACHI, M.A. and BOYD, J.B. (1985) The *giant* (*gl*) mutants of *Drosophila melanogaster* alter DNA metabolism. *Mol Gen Genet* **199**: 500-506.
- NEKRUTENKO, A. and LI, W.H. (2001) Transposable elements are found in a large number of human protein-coding genes. *Trends Genet* **17**: 619-621.
- NEUFELD, T.P., DE LA CRUZ, A.F., JOHNSTON, L.A. and EDGAR, B.A. (1998) Coordination of growth and cell division in the *Drosophila* wing. *Cell* **93**: 1183-1193.
- NIEDERMAIER, M., SCHWABE, G.C., FEES, S., HELMRICH, A., BRIESKE, N., SEEMANN, P., HECHT, J., SEITZ, V., STRICKER, S., LESCHIK, G., *et al.* (2005) An inversion involving the mouse *Sbh* locus results in brachydactyly through dysregulation of *Sbh* expression. *J Clin Invest* **115**: 900-909.
- NIGUMANN, P., REDIK, K., MATLIK, K. and SPEEK, M. (2002) Many human genes are transcribed from the antisense promoter of *L1* retrotransposon. *Genomics* **79**: 628-634.
- NORRY, F.M., VILARDI, J.C., FANARA, J.J., HASSON, E. and RODRIGUEZ, C. (1995) An adaptive chromosomal polymorphism affecting size-related traits, and longevity selection in a natural population of *Drosophila buzzatii*. *Genetica* **96**: 285-291.
- NOTREDAME, C., HIGGINS, D.G. and HERINGA, J. (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* **302**: 205-217.
- O'FARRELL, P.H. (2004) How metazoans reach their full size: the natural history of bigness. In *Cell growth: control of cell size* (eds. Hall, M.N., Raff, M. and Thomas, G.), pp. 1-22. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
- OBBARD, D.J., GORDON, K.H., BUCK, A.H. and JIGGINS, F.M. (2009) The evolution of RNAi as a defence against viruses and transposable elements. *Philos Trans R Soc Lond B Biol Sci* **364**: 99-115.
- OHLER, U., LIAO, G.C., NIEMANN, H. and RUBIN, G.M. (2002) Computational analysis of core promoters in the *Drosophila* genome. *Genome Biol* **3**: RESEARCH0087.
- OHLER, U. (2006) Identification of core promoter modules in *Drosophila* and their application in accurate transcription start site prediction. *Nucleic Acids Res* **34**: 5943-5950.
- OLIVER, F., CHRISTIANS, J.K., LIU, X., RHIND, S., VERMA, V., DAVISON, C., BROWN, S.D., DENNY, P. and KEIGHTLEY, P.D. (2005) Regulatory variation at glypican-3 underlies a major growth QTL in mice. *PLoS Biol* **3**: e135.
- OTSUKI, K., HAYASHI, Y., KATO, M., YOSHIDA, H. and YAMAGUCHI, M. (2004) Characterization of dRFX2, a novel RFX family protein in *Drosophila*. *Nucleic Acids Res* **32**: 5636-5648.

- PAINTER, T.S. (1933) A new method for the study of chromosome rearrangements and the plotting of chromosome maps. *Science* **78**: 585-586.
- PAL-BHADRA, M., LEIBOVITCH, B.A., GANDHI, S.G., RAO, M., BHADRA, U., BIRCHLER, J.A. and ELGIN, S.C. (2004) Heterochromatic silencing and HP1 localization in *Drosophila* are dependent on the RNAi machinery. *Science* **303**: 669-672.
- PARDUE, M.L., RASHKOVA, S., CASACUBERTA, E., DEBARYSHE, P.G., GEORGE, J.A. and TRAVERSE, K.L. (2005) Two retrotransposons maintain telomeres in *Drosophila*. *Chromosome Res* **13**: 443-453.
- PAULER, F.M., STRICKER, S.H., WARCZOK, K.E. and BARLOW, D.P. (2005) Long-range DNase I hypersensitivity mapping reveals the imprinted *Igf2r* and *Air* promoters share *cis*-regulatory elements. *Genome Res* **15**: 1379-1387.
- PEDONE, P.V., GHIRLANDO, R., CLORE, G.M., GRONENBORN, A.M., FELSENFELD, G. and OMICHINSKI, J.G. (1996) The single Cys2-His2 zinc finger domain of the GAGA protein flanked by basic residues is sufficient for high-affinity specific DNA binding. *Proc Natl Acad Sci U S A* **93**: 2822-2826.
- PEI, J. and GRISHIN, N.V. (2001) AL2CO: calculation of positional conservation in a protein sequence alignment. *Bioinformatics* **17**: 700-712.
- PENNISI, E. (2007) Breakthrough of the year. Human genetic variation. *Science* **318**: 1842-1843.
- PETRUK, S., SEDKOV, Y., RILEY, K.M., HODGSON, J., SCHWEISGUTH, F., HIROSE, S., JAYNES, J.B., BROCK, H.W. and MAZO, A. (2006) Transcription of *bxd* noncoding RNAs promoted by trithorax represses *Ubx* in *cis* by transcriptional interference. *Cell* **127**: 1209-1221.
- PILOT, F., PHILIPPE, J.M., LEMMERS, C., CHAUVIN, J.P. and LECUIT, T. (2006) Developmental control of nuclear morphogenesis and anchoring by *charleston*, identified in a functional genomic screen of *Drosophila* cellularisation. *Development* **133**: 711-723.
- PIMPINELLI, S., BERLOCO, M., FANTI, L., DIMITRI, P., BONACCORSI, S., MARCHETTI, E., CAZZI, R., CAGGESE, C. and GATTI, M. (1995) Transposable elements are stable structural components of *Drosophila melanogaster* heterochromatin. *Proc Natl Acad Sci U S A* **92**: 3804-3808.
- PIÑOL, J., FRANCINO, O., FONTDEVILA, A. and CABRÉ, O. (1988) Rapid isolation of *Drosophila* high molecular weight DNA to obtain genomic libraries. *Nucleic Acids Res* **16**: 2736.
- PRADA, C.F. (2010) Evolución cromosómica del cluster *Drosophila martensis*: origen de las inversiones y reutilización de los puntos de rotura. Doctoral thesis. Departament de Genètica i Microbiologia, Universitat Autònoma de Barcelona, Bellaterra, Spain.
- PRAZERES DA COSTA, O., GONZÁLEZ, J. and RUIZ, A. (2009) Cloning and sequencing of the breakpoint regions of inversion 5g fixed in *Drosophila buzzatii*. *Chromosoma* **118**: 349-360.
- PRESCOTT, E.M. and PROUDFOOT, N.J. (2002) Transcriptional collision between convergent genes in budding yeast. *Proc Natl Acad Sci U S A* **99**: 8796-8801.
- PREVOSTI, A., RIBÓ, G., SERRA, L., AGUADÉ, M., BALANYÀ, J., MONCLUS, M. and MESTRES, F. (1988) Colonization of America by *Drosophila subobscura*: Experiment in natural populations that supports the adaptive role of chromosomal-inversion polymorphism. *Proc Natl Acad Sci U S A* **85**: 5597-5600.
- PROUDFOOT, N. (2004) New perspectives on connecting messenger RNA 3' end formation to transcription. *Curr Opin Cell Biol* **16**: 272-278.
- PROUDFOOT, N.J., FURGER, A. and DYE, M.J. (2002) Integrating mRNA processing with transcription. *Cell* **108**: 501-512.
- PUIG, M., CÁCERES, M. and RUIZ, A. (2004) Silencing of a gene adjacent to the breakpoint of a widespread *Drosophila* inversion by a transposon-induced antisense RNA. *Proc Natl Acad Sci U S A* **101**: 9013-9018.
- QUINLAN, A.R., CLARK, R.A., SOKOLOVA, S., LEIBOWITZ, M.L., ZHANG, Y., HURLES, M.E., MELL, J.C. and HALL, I.M. (2010) Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. *Genome Res* **20**: 623-635.
- RAKO, L., ANDERSON, A.R., SGRO, C.M., STOCKER, A.J. and HOFFMANN, A.A. (2006) The association between inversion *In(3R)Payne* and clinally varying traits in *Drosophila melanogaster*. *Genetica* **128**: 373-384.

- RANZ, J.M., CASALS, F. and RUIZ, A. (2001) How malleable is the eukaryotic genome? Extreme rate of chromosomal rearrangement in the genus *Drosophila*. *Genome Res* **11**: 230-239.
- RANZ, J.M., MAURIN, D., CHAN, Y.S., VON GROTHUSS, M., HILLIER, L.W., ROOTE, J., ASHBURNER, M. and BERGMAN, C.M. (2007) Principles of genome evolution in the *Drosophila melanogaster* species group. *PLoS Biol* **5**: e152.
- RECHSTEINER, M. and ROGERS, S.W. (1996) PEST sequences and regulation by proteolysis. *Trends Biochem Sci* **21**: 267-271.
- REESE, M.G. (2001) Application of a time-delay neural network to promoter annotation in the *Drosophila melanogaster* genome. *Comput Chem* **26**: 51-56.
- RIO, D.C. (2002) P transposable elements in *Drosophila melanogaster*. In *Mobile DNA II* (eds. Craig, N.L., Craigie, R., Gellert, M. and Lambowitz, A.M.), pp. 484-518. ASM Press, Washington D.C.
- ROCKMAN, M.V., HAHN, M.W., SORANZO, N., ZIMPRICH, F., GOLDSTEIN, D.B. and WRAY, G.A. (2005) Ancient and recent positive selection transformed opioid *cis*-regulation in humans. *PLoS Biol* **3**: e387.
- RODRÍGUEZ-TRELLES, F., ÁLVAREZ, G. and ZAPATA, C. (1996) Time-series analysis of seasonal changes of the O inversion polymorphism of *Drosophila subobscura*. *Genetics* **142**: 179-187.
- RODRÍGUEZ-TRELLES, F., TARRÍO, R. and AYALA, F.J. (2006) Origins and evolution of spliceosomal introns. *Annu Rev Genet* **40**: 47-76.
- ROFF, D.A. (2000) Trade-offs between growth and reproduction: an analysis of the quantitative genetic evidence. *J Evol Biol* **13**: 434-445.
- ROGERS, S., WELLS, R. and RECHSTEINER, M. (1986) Amino acid sequences common to rapidly degraded proteins: the PEST hypothesis. *Science* **234**: 364-368.
- ROMANISH, M.T., LOCK, W.M., VAN DE LAGEMAAT, L.N., DUNN, C.A. and MAGER, D.L. (2007) Repeated recruitment of LTR retrotransposons as promoters by the anti-apoptotic locus NAIP during mammalian evolution. *PLoS Genet* **3**: e10.
- RUIZ, A., SANTOS, M., BARBADILLA, A., QUEZADA-DIAZ, J.E., HASSON, E. and FONTDEVILA, A. (1991) Genetic variance for body size in a natural population of *Drosophila buzzatii*. *Genetics* **128**: 739-750.
- SAMBROOK, J., FRITSCH, E.F. and MANIATIS, T. (1989) *Molecular cloning: a laboratory manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
- SANDOVAL, A., OVIEDO, N., TADMOURI, A., AVILA, T., DE WAARD, M. and FELIX, R. (2006) Two PEST-like motifs regulate Ca²⁺/calpain-mediated cleavage of the Cavβ₃ subunit and provide important determinants for neuronal Ca²⁺ channel activity. *Eur J Neurosci* **23**: 2311-2320.
- SANTOS, M., RUIZ, A., QUEZADA-DIAZ, J.E., BARBADILLA, A. and FONTDEVILA, A. (1992) The evolutionary history of *Drosophila buzzatii*. XX. Positive phenotypic covariance between field adult fitness components and body size. *J Evol Biol* **5**: 403-422.
- SCHAEFFER, S.W., GOETTING-MINESKY, M.P., KOVACEVIC, M., PEOPLES, J.R., GRAYBILL, J.L., MILLER, J.M., KIM, K., NELSON, J.G. and ANDERSON, W.W. (2003) Evolutionary genomics of inversions in *Drosophila pseudoobscura*: evidence for epistasis. *Proc Natl Acad Sci U S A* **100**: 8319-8324.
- SCHLENKE, T.A. and BEGUN, D.J. (2004) Strong selective sweep associated with a transposon insertion in *Drosophila simulans*. *Proc Natl Acad Sci U S A* **101**: 1626-1631.
- SCHMIDT, J.M., GOOD, R.T., APPLETON, B., SHERRARD, J., RAYMANT, G.C., BOGWITZ, M.R., MARTIN, J., DABORN, P.J., GODDARD, M.E., BATTERHAM, P., *et al.* (2010) Copy number variation and transposable elements feature in recent, ongoing adaptation at the *Cyp6g1* locus. *PLoS Genet* **6**: e1000998.
- SCHWARTZ, M.B., IMBERSKI, R.B. and KELLY, T.J. (1984) Analysis of metamorphosis in *Drosophila melanogaster*: characterization of *giant*, an ecdysteroid-deficient mutant. *Dev Biol* **103**: 85-95.
- SEKIMIZU, K., NISHIOKA, N., SASAKI, H., TAKEDA, H., KARLSTROM, R.O. and KAWAKAMI, A. (2004) The zebrafish *iguana* locus encodes Dzip1, a novel zinc-finger protein required for proper regulation of Hedgehog signaling. *Development* **131**: 2521-2532.

- SHAPIRO, M.D., MARKS, M.E., PEICHEL, C.L., BLACKMAN, B.K., NERENG, K.S., JONSSON, B., SCHLUTER, D. and KINGSLEY, D.M. (2004) Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. *Nature* **428**: 717-723.
- SIJEN, T. and PLASTERK, R.H. (2003) Transposon silencing in the *Caenorhabditis elegans* germ line by natural RNAi. *Nature* **426**: 310-314.
- SILVERMAN, M. and SIMON, M. (1980) Phase variation: genetic analysis of switching mutants. *Cell* **19**: 845-854.
- SIMPSON, P., BERREUR, P. and BERREUR-BONNENFANT, J. (1980) The initiation of pupariation in *Drosophila*: dependence on growth of the imaginal discs. *J Embryol Exp Morphol* **57**: 155-165.
- SIMPSON, P. and MORATA, G. (1980) The control of growth in the imaginal discs of *Drosophila*. In *Development and Neurobiology of Drosophila* (eds. Siddiqi, O., Babu, P., Hall, L.M. and Hall, J.C.), pp. 129-140. Plenum Press, New York.
- SIRONEN, A., THOMSEN, B., ANDERSSON, M., AHOLA, V. and VILKKI, J. (2006) An intronic insertion in *KPL2* results in aberrant splicing and causes the immotile short-tail sperm defect in the pig. *Proc Natl Acad Sci U S A* **103**: 5006-5011.
- SISSON, B.E., ZIEGENHORN, S.L. and HOLMGREN, R.A. (2006) Regulation of Ci and Su(fu) nuclear import in *Drosophila*. *Dev Biol* **294**: 258-270.
- SLEUTELS, F., ZWART, R. and BARLOW, D.P. (2002) The non-coding *Air* RNA is required for silencing autosomal imprinted genes. *Nature* **415**: 810-813.
- SLOTKIN, R.K. and MARTIENSSON, R. (2007) Transposable elements and the epigenetic regulation of the genome. *Nat Rev Genet* **8**: 272-285.
- SOKAL, R.R. and ROHLF, F.J. (2000) *Biometry*. W.H. Freeman and Company, New York.
- SPELEMAN, F., CAUWELIER, B., DASTUGUE, N., COOLS, J., VERHASSELT, B., POPPE, B., VAN ROY, N., VANDESOMPELE, J., GRAUX, C., UYTTEBROECK, A., *et al.* (2005) A new recurrent inversion, inv(7)(p15q34), leads to transcriptional activation of *HOXA10* and *HOXA11* in a subset of T-cell acute lymphoblastic leukemias. *Leukemia* **19**: 358-366.
- SPERLICH, D. (1966) Equilibria for inversions induced by X-rays in isogenic strains of *Drosophila pseudoobscura*. *Genetics* **53**: 835-842.
- SPERLICH, D. and PFRIEM, P. (1986) Chromosomal polymorphism in natural and experimental populations. In *The genetics and biology of Drosophila*, vol. 3e (eds. Ashburner, M., Carson, H.L. and Thompson, J.N.), pp. 257-309. Academic Press, London.
- STAVENHAGEN, J.B. and ROBINS, D.M. (1988) An ancient provirus has imposed androgen regulation on the adjacent mouse sex-limited protein gene. *Cell* **55**: 247-254.
- STEFANSSON, H., HELGASON, A., THORLEIFSSON, G., STEINTHORSDOTTIR, V., MASSON, G., BARNARD, J., BAKER, A., JONASDOTTIR, A., INGASON, A., GUDNADOTTIR, V.G., *et al.* (2005) A common inversion under selection in Europeans. *Nat Genet* **37**: 129-137.
- STERN, D.L. (1998) A role of Ultrabithorax in morphological differences between *Drosophila* species. *Nature* **396**: 463-466.
- STONE, W.S., GUEST, W.C. and WILSON, F.D. (1960) The evolutionary implications of the cytological polymorphism and phylogeny of the virilis group of *Drosophila*. *Proc Natl Acad Sci U S A* **46**: 350-361.
- STRANGER, B.E., FORREST, M.S., DUNNING, M., INGLE, C.E., BEAZLEY, C., THORNE, N., REDON, R., BIRD, C.P., DE GRASSI, A., LEE, C., *et al.* (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315**: 848-853.
- STURTEVANT, A.H. (1917) Genetic factors affecting the strength of linkage in *Drosophila*. *Proc Natl Acad Sci U S A* **3**: 555-558.
- SU, T.T. and O'FARRELL, P.H. (1998) Size control: cell proliferation does not equal growth. *Curr Biol* **8**: R687-689.
- SUN, F.L., HAYNES, K., SIMPSON, C.L., LEE, S.D., COLLINS, L., WULLER, J., EISSENBERG, J.C. and ELGIN, S.C. (2004) *cis*-acting determinants of heterochromatin formation on *Drosophila melanogaster* chromosome four. *Mol Cell Biol* **24**: 8210-8220.
- SUN, M., HURST, L.D., CARMICHAEL, G.G. and CHEN, J. (2006) Evidence for variation in abundance of antisense transcripts between multicellular animals but no relationship between antisense

- transcription and organismic complexity. *Genome Res* **16**: 922-933.
- TANAKA-MATAKATSU, M. and DU, W. (2008) Direct control of the proneural gene *atonal* by retinal determination factors during *Drosophila* eye development. *Dev Biol* **313**: 787-801.
- THOMAS, J.W., CÁCERES, M., LOWMAN, J.J., MOREHOUSE, C.B., SHORT, M.E., BALDWIN, E.L., MANEY, D.L. and MARTIN, C.L. (2008) The chromosomal polymorphism linked to variation in social behavior in the white-throated sparrow (*Zonotrichia albicollis*) is a complex rearrangement and suppressor of recombination. *Genetics* **179**: 1455-1468.
- THOMAS, P.D., CAMPBELL, M.J., KEJARIWAL, A., MI, H., KARLAK, B., DAVERMAN, R., DIEMER, K., MURUGANUJAN, A. and NARECHANIA, A. (2003) PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res* **13**: 2129-2141.
- TING, C.N., ROSENBERG, M.P., SNOW, C.M., SAMUELSON, L.C. and MEISLER, M.H. (1992) Endogenous retroviral sequences are required for tissue-specific expression of a human salivary amylase gene. *Genes Dev* **6**: 1457-1465.
- TRIMARCHI, J.M. and LEES, J.A. (2002) Sibling rivalry in the E2F family. *Nat Rev Mol Cell Biol* **3**: 11-20.
- TUFARELLI, C., STANLEY, J.A., GARRICK, D., SHARPE, J.A., AYYUB, H., WOOD, W.G. and HIGGS, D.R. (2003) Transcription of antisense RNA leading to gene silencing and methylation as a novel cause of human genetic disease. *Nat Genet* **34**: 157-165.
- TUSHER, V.G., TIBSHIRANI, R. and CHU, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* **98**: 5116-5121.
- TWEEDIE, S., ASHBURNER, M., FALLS, K., LEYLAND, P., MCQUILTON, P., MARYGOLD, S., MILLBURN, G., OSUMI-SUTHERLAND, D., SCHROEDER, A., SEAL, R., *et al.* (2009) FlyBase: enhancing *Drosophila* Gene Ontology annotations. *Nucleic Acids Res* **37**: D555-559.
- UMINA, P.A., WEEKS, A.R., KEARNEY, M.R., MCKECHNIE, S.W. and HOFFMANN, A.A. (2005) A rapid shift in a classic clinal pattern in *Drosophila* reflecting climate change. *Science* **308**: 691-693.
- VAN DE LAGEMAAT, L.N., LANDRY, J.R., MAGER, D.L. and MEDSTRAND, P. (2003) Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. *Trends Genet* **19**: 530-536.
- VOKES, S.A. and MCMAHON, A.P. (2004) Hedgehog signaling: iguana debuts as a nuclear gatekeeper. *Curr Biol* **14**: R668-670.
- WANG, T., ZENG, J., LOWE, C.B., SELLERS, R.G., SALAMA, S.R., YANG, M., BURGESS, S.M., BRACHMANN, R.K. and HAUSSLER, D. (2007) Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. *Proc Natl Acad Sci U S A* **104**: 18613-18618.
- WASSERMAN, M. (1954) Cytological studies of the *repleta* group. *Univ Texas Publ* **5422**: 130-152.
- WASSERMAN, M. (1992) Cytological evolution of the *Drosophila repleta* species group. In *Drosophila inversion polymorphism* (eds. Krimbas, C.B. and Powell, J.R.), pp. 455-552. CRC Press, Boca Raton, Florida.
- WEEDON, M.N. and FRAYLING, T.M. (2008) Reaching new heights: insights into the genetics of human stature. *Trends Genet* **24**: 595-603.
- WEIGMANN, K., COHEN, S.M. and LEHNER, C.F. (1997) Cell cycle progression, growth and patterning in imaginal discs despite inhibition of cell division after inactivation of *Drosophila* Cdc2 kinase. *Development* **124**: 3555-3563.
- WEILER, K.S. and WAKIMOTO, B.T. (1995) Heterochromatin and gene expression in *Drosophila*. *Annu Rev Genet* **29**: 577-605.
- WESLEY, C.S. and EANES, W.F. (1994) Isolation and analysis of the breakpoint sequences of chromosome inversion *In(3L)Payne* in *Drosophila melanogaster*. *Proc Natl Acad Sci U S A* **91**: 3132-3136.
- WICKER, T., SABOT, F., HUA-VAN, A., BENNETZEN, J.L., CAPY, P., CHALHOUB, B., FLAVELL, A., LEROY, P., MORGANTE, M., PANAUD, O., *et al.* (2007) A unified classification system for eukaryotic transposable elements. *Nat Rev Genet* **8**: 973-982.
- WITTKOPP, P.J., HAERUM, B.K. and CLARK, A.G. (2008) Independent effects of cis- and trans-regulatory variation on gene expression in *Drosophila melanogaster*. *Genetics* **178**: 1831-1835.

- WITTKOPP, P.J. (2010) Variable transcription factor binding: a mechanism of evolutionary change. *PLoS Biol* **8**: e1000342.
- WOLFF, C., ROY, S., LEWIS, K.E., SCHAUERTE, H., JOERG-RAUCH, G., KIRN, A., WEILER, C., GEISLER, R., HAFFTER, P. and INGHAM, P.W. (2004) *iguana* encodes a novel zinc-finger protein with coiled-coil domains essential for Hedgehog signal transduction in the zebrafish embryo. *Genes Dev* **18**: 1565-1576.
- WRAY, G.A., HAHN, M.W., ABOUHEIF, E., BALHOFF, J.P., PIZER, M., ROCKMAN, M.V. and ROMANO, L.A. (2003) The evolution of transcriptional regulation in eukaryotes. *Mol Biol Evol* **20**: 1377-1419.
- WRAY, G.A. (2007) The evolutionary significance of *cis*-regulatory mutations. *Nat Rev Genet* **8**: 206-216.
- WURMBACH, E., YUEN, T. and SEALFON, S.C. (2003) Focused microarray analysis. *Methods* **31**: 306-316.
- XING, J., WANG, H., BELANCIO, V.P., CORDAUX, R., DEININGER, P.L. and BATZER, M.A. (2006) Emergence of primate genes by retrotransposon-mediated sequence transduction. *Proc Natl Acad Sci U S A* **103**: 17608-17613.
- YOCHUM, G.S., CLELAND, R. and GOODMAN, R.H. (2008) A genome-wide screen for β -catenin binding sites identifies a downstream enhancer element that controls *c-Myc* gene expression. *Mol Cell Biol* **28**: 7368-7379.
- ZAPALA, M.A., LOCKHART, D.J., PANKRATZ, D.G., GARCIA, A.J., BARLOW, C. and LOCKHART, D.J. (2002) Software and methods for oligonucleotide and cDNA array data analysis. *Genome Biol* **3**: SOFTWARE0001.
- ZATSEPINA, O.G., VELIKODVORSKAIA, V.V., MOLODTSOV, V.B., GARBUZ, D., LERMAN, D.N., BETTENCOURT, B.R., FEDER, M.E. and EVGENEV, M.B. (2001) A *Drosophila melanogaster* strain from sub-equatorial Africa has exceptional thermotolerance but decreased *Hsp70* expression. *J Exp Biol* **204**: 1869-1881.
- ZDOBNOV, E.M. and APWEILER, R. (2001) InterProScan – an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**: 847-848.
- ZHOU, Y.H., ZHENG, J.B., GU, X., SAUNDERS, G.F. and YUNG, W.K. (2002) Novel PAX6 binding sites in the human genome and the role of repetitive elements in the evolution of gene regulation. *Genome Res* **12**: 1716-1722.

DATABASES

AAA: 12 *Drosophila* Genomes – Assembly/Alignment/Annotation of 12 related *Drosophila* species <http://rana.lbl.gov/drosophila/>

Affymetrix – NetAffx™ Analysis Center
<http://www.affymetrix.com/analysis/index.affx>

DroSpeGe – *Drosophila* species genome browser
<http://insects.eugenes.org/species/>

GenBank – Sequence archive from the NCBI (National Center for Biotechnology Information)
<http://www.ncbi.nlm.nih.gov/>

Gene Ontology – The Gene Ontology Database
<http://www.geneontology.org/>

FlyAtlas – Gene expression in multiple tissues of *Drosophila melanogaster*
<http://www.flyatlas.org/>

FlyBase – A database for *Drosophila* genetics and molecular biology
<http://flybase.org/>

FlyExpress – A Knowledge-base of Spatiotemporal Expression Patterns at a Genomic-scale in the Fruit-fly Embryogenesis
<http://www.flyexpress.net/>

FlyTF – Integrated database of genomic and protein data for *Drosophila* site-specific transcription factors
<http://www.flytf.org>

KEGG PATHWAY Database – Kyoto Encyclopedia of Genes and Genomes

<http://www.genome.jp/kegg>

PANTHER – Protein ANalysis THrough Evolutionary Relationships

<http://www.pantherdb.org/>

VectorBase – Bioinformatics Resource Center for Invertebrate Vectors of Human Pathogens

<http://www.vectorbase.org/index.php>

SOFTWARE

AL2CO – Sequence conservation analysis server

<http://prodata.swmed.edu/al2co/al2co.php>

BLAST – Sequence similarity searches (NCBI)

<http://blast.ncbi.nlm.nih.gov/Blast.cgi>

CLUSTALW – Multiple sequence alignment of DNA and proteins (EBI)

<http://www.ebi.ac.uk/Tools/clustalw2/index.html>

COILS – Prediction of coiled coil regions in proteins

http://www.ch.embnet.org/software/COILS_form.html

EPESTFIND – Finds PEST motifs as potential proteolytic cleavage sites

<http://emboss.bioinformatics.nl/cgi-bin/emboss/epestfind>

Functional Annotation Clustering tool – Identification of enriched biological themes within a gene list (DAVID Bioinformatics Resources)

<http://david.abcc.ncifcrf.gov>

INTERPROSCAN – Protein search against InterPro database of protein signatures (EBI)

<http://www.ebi.ac.uk/Tools/InterProScan/>

MATCH™ public version 1.0 – Matrix search for transcription factor binding sites

<http://www.gene-regulation.com/cgi-bin/pub/programs/match/bin/match.cgi>

McPROMOTER – Statistical tool for the prediction of transcription start sites in eukaryotic DNA

<http://tools.genome.duke.edu/generegulation/McPromoter/>

MUSCLE – Multiple sequence alignment (EBI)

<http://www.ebi.ac.uk/Tools/muscle/>

NETNES 1.1 SERVER – Prediction of leucine-rich nuclear export signals (NES) in eukaryotic proteins

<http://www.cbs.dtu.dk/services/NetNES/>

NEURAL NETWORK PROMOTER PREDICTION – Neural network based program to find possible transcription promoters

<http://www.fruitfly.org/seq-tools/promoter.html>

PSORTII – Prediction of protein localization sites within cells

<http://psort.ims.u-tokyo.ac.jp/>

T-COFFEE – Multiple sequence alignment

<http://www.tcoffee.org/>

VISTA – Programs and databases for comparative analysis of genomic sequences

<http://genome.lbl.gov/vista/index.shtml>

Abbreviations

aa	amino acid
Ago	Argonaute
BAC	bacterial artificial chromosome
bp	base pairs
BP	breakpoint
BSA	bovine serum albumin
cDNA	complementary DNA
CF	cleavage factor
CNS	conserved non-coding sequence
CPSF	cleavage and polyadenylation specificity factor
CstF	cleavage stimulatory factor
cRNA	complementary RNA
DAZ	deleted in azoospermia
df	degrees of freedom
DNA	deoxyribonucleic acid
DRE	DNA-replication related element
DSE	downstream sequence element
dsDNA	double-stranded DNA
dsRNA	double-stranded RNA
DZIP1	DAZ interacting protein 1
DZIP1L	DAZ interacting protein 1-like
EDTA	ethylenediaminetetraacetic acid
ERV	endogenous retrovirus
EST	expressed sequence tag
FC	fold change
HERV	human endogenous retrovirus
Hh	Hedgehog
kb	kilobase
LINE	long interspersed element

LTR	long terminal repeat
miRNA	microRNA
mRNA	messenger RNA
MS	mean of squares
NAHR	non-allelic homologous recombination
NES	nuclear export signal
NHEJ	non-homologous end-joining
NLS	nuclear localization signal
nt	nucleotide
ORF	open reading frame
<i>P</i>	probability
PAP	polyA polymerase
PCR	polymerase chain reaction
PEV	position effect variegation
Pol II	RNA polymerase II
polyA	polyadenylation
PVP	polyvinylpyrrolidone
RACE	rapid amplification of cDNA ends
RdRP	RNA-dependent RNA polymerase
RISC	RNA-induced silencing complex
RNA	ribonucleic acid
RNAi	RNA interference
rNTPs	ribonucleotides
rRNA	ribosomal RNA
RT	reverse transcriptase
RT-PCR	reverse transcription polymerase chain reaction
SD	standard deviation
SDS	sodium dodecyl sulfate
SINE	short interspersed element
siRNAs	small interfering RNAs
SNP	single nucleotide polymorphism
SS	sum of squares

SSC	saline-sodium citrate buffer
TE	transposable element
TFBS	transcription factor binding site
TIR	terminal inverted repeat
tRNA	transfer RNA
TSD	target site duplication
TSS	transcription start site
UTR	untranslated region
[³² P]UTP	uridine triphosphate radioactively labeled with phosphorus-32

Units

Da	dalton
g	gram
h	hour
M	molar
min	minute
ml	milliliter
mM	millimolar
ng	nanogram
sec	second
μg	microgram
μl	microliter
°C	degree Celsius

Drosophila genes

<i>Art1</i>	arginine methyltransferase 1
<i>Ci</i>	cubitus interruptus
<i>CycE</i>	cyclin E
<i>Cos2</i>	costal 2
<i>DebB</i>	developmental embryonic B
<i>Disp</i>	dispatched
<i>dnk</i>	deoxyribonucleoside kinase
<i>dUTPase</i>	deoxyuridine triphosphatase
<i>Est-6</i>	esterase 6
<i>Fu</i>	fused
<i>Gapdh</i>	glyceraldehyde 3 phosphate dehydrogenase
<i>gt</i>	giant
<i>Hsp70</i>	heat shock protein 70
<i>Hsp83</i>	heat shock protein 83
<i>Lcp4</i>	larval cuticle protein 4
<i>Mcm2</i>	minichromosome maintenance 2
<i>Mcm5</i>	minichromosome maintenance 5
<i>Mcm7</i>	minichromosome maintenance 7
<i>mus209</i>	mutagen-sensitive 209
<i>nAChRβ-96A</i>	nicotinic acetylcholine receptor β at 96A
<i>Obp56a</i>	odorant-binding protein 56a
<i>Pka</i>	cAMP-dependent protein kinase 1
<i>pont</i>	pontin
<i>Pp1α-96A</i>	protein phosphatase 1 α at 96A
<i>Ptc</i>	patched
<i>RanGap</i>	Ran GTPase activating protein
<i>RnrL</i>	ribonucleoside diphosphate reductase large subunit
<i>RnrS</i>	ribonucleoside diphosphate reductase small subunit
<i>Smo</i>	smoothened
<i>Su(fu)</i>	suppressor of fused
<i>Ts</i>	thymidylate synthase

Drosophila species

Dana	<i>D. ananassae</i>
Dbuz	<i>D. buzzatii</i>
Dere	<i>D. erecta</i>
Dgri	<i>D. grimshawi</i>
Dmar	<i>D. martensis</i>
Dmel	<i>D. melanogaster</i>
Dmoj	<i>D. mojavensis</i>
Dper	<i>D. persimilis</i>
Dpse	<i>D. pseudoobscura</i>
Dsec	<i>D. sechellia</i>
Dsim	<i>D. simulans</i>
Dvir	<i>D. virilis</i>
Dwil	<i>D. willistoni</i>
Dyak	<i>D. yakuba</i>

Index of figures

1 | INTRODUCTION

FIGURE 1 900-kb inversion polymorphism detected in human chromosome 17.	24
FIGURE 2 Examples of possible mechanisms for position effects altering the expression of genes flanking inversion breakpoints.	32
FIGURE 3 TEs can influence gene expression through different mechanisms.	37
FIGURE 4 Classification and structure of TEs.	40
FIGURE 5 Specific cases of genes whose expression is affected by TEs.	46
FIGURE 6 Epigenetic regulation of TEs is able to influence gene expression.	49
FIGURE 7 Inversion $2j$ in <i>D. buzzatii</i> chromosome 2.	52
FIGURE 8 TEs inserted at the distal and proximal breakpoint junctions in different $2j$ chromosomes.	54
FIGURE 9 Genes flanking inversion $2j$ breakpoints in $2st$ and $2j$ chromosomes.	57
FIGURE 10 Proximity of <i>Rox8</i> and <i>Pp1α-96A</i> transcripts to inversion $2j$ breakpoints.	59

2 | MATERIALS AND METHODS

FIGURE 11 dsRNA synthesis in <i>D. melanogaster</i> .	80
FIGURE 12 RNA interference mechanism.	83
FIGURE 13 Labeling of RNA samples to be hybridized in Affymetrix microarrays for gene expression profiling.	87
FIGURE 14 Affymetrix GeneChip [®] microarray.	88
FIGURE 15 Structure and spatial organization of Affymetrix microarrays and hybridization of the labeled mRNAs to their corresponding complementary probes.	88
FIGURE 16 Each transcript is analyzed by a specific probe set formed by 14 probe pairs.	89

3 | RESULTS

Article – FIGURE 1 Schematic representation of inversion <i>2j</i> proximal breakpoint sequence in lines st-1 and j-1.	95
Article – FIGURE 2 Nucleotide sequence of the 3' end of <i>CG13617</i> in lines st-1 and j-1.	96
Article – FIGURE 3 Expression analysis of <i>CG13617</i> in four different developmental stages of two lines with (<i>2j</i>) and without (<i>2st</i>) the inversion.	97
Article – FIGURE 4 Relative quantification of <i>CG13617</i> mRNA and antisense RNA and <i>Pp1α-96A</i> mRNA by real-time RT-PCR.	97
Article – FIGURE 5 Strand-specific RT-PCR of <i>CG13617</i> .	97
Article – FIGURE 6 Detection of <i>CG13617</i> double-stranded RNA (dsRNA) in st-1 and j-19 embryos.	98
Article – SUPPORTING FIGURE 7 Expression pattern of <i>CG13617</i> in <i>Drosophila buzzatii</i> embryos analyzed by <i>in situ</i> hybridization.	100
Article – SUPPORTING FIGURE 8 Diagram of the <i>Drosophila</i> <i>CG13617</i> protein and similar proteins of other species.	101
FIGURE 17 Experimental design of the functional analysis of the consequences of <i>CG13617</i> silencing.	106
FIGURE 18 Silencing of <i>CG13617</i> gene expression in <i>D. melanogaster</i> by RNAi.	109
FIGURE 19 Venn diagram showing the results of the microarray analysis using three independent methods.	110
FIGURE 20 Expression data of genes differentially expressed in samples with (CONTROL) and without (DSRNA) <i>CG13617</i> expression according to microarray analysis.	113
FIGURE 21 Schematic representation of the location of microarray probes, dsRNA molecule, and amplification products in the <i>D. melanogaster</i> <i>CG13617</i> gene sequence.	114
FIGURE 22 Gene ontology analysis of the genes that are differentially expressed when <i>CG13617</i> is silenced.	115
FIGURE 23 Diagram showing the DNA replication complex in <i>D. melanogaster</i> .	117

FIGURE 24 Differentially expressed genes after <i>CG13617</i> silencing involved in nucleotide metabolism in <i>D. melanogaster</i> .	118
FIGURE 25 Real-time RT-PCR results for the eight genes analyzed in <i>D. melanogaster</i> .	120
FIGURE 26 Real-time RT-PCR results for the ten genes analyzed in <i>D. buzzatii</i> .	124
FIGURE 27 <i>CG13617</i> gene structure in different <i>Drosophila</i> species.	128
FIGURE 28 Alignment and conservation plot of <i>CG13617</i> proteins of 14 different <i>Drosophila</i> species.	133
FIGURE 29 Sliding window graph of <i>CG13617</i> protein sequence conservation in the 14 <i>Drosophila</i> species.	137
FIGURE 30 Analysis of the sequences upstream gene <i>CG13617</i> using VISTA.	142
FIGURE 31 Sequence alignments of the CNSs located in the intergenic region separating genes <i>nAcRβ-96A</i> and <i>CG13617</i> .	144

4 | DISCUSSION

FIGURE 32 Fold reduction of <i>CG13617</i> expression level in different developmental stages of <i>D. buzzatii</i> 2j lines.	150
FIGURE 33 Factors and sequence elements required for mRNA 3' end processing and transcription termination.	153
FIGURE 34 Nucleotide sequences of the 3' UTRs of <i>CG13617</i> mRNAs in lines st-1 and j-1.	155
FIGURE 35 Examples of natural antisense RNAs with regulatory functions.	163
FIGURE 36 Mechanisms of antisense regulation.	165
FIGURE 37 Animal groups with proteins homologous to <i>CG13617</i> .	188
FIGURE 38 Alignment of <i>D. buzzatii</i> <i>CG13617</i> with human DZIP1 and zebrafish Iguana proteins.	190
FIGURE 39 Hedgehog signaling pathway in <i>Drosophila</i> .	191
FIGURE 40 Cyclin E and RanGap cellular functions.	193
FIGURE 41 <i>D. melanogaster</i> RNA-Seq developmental profile.	196
FIGURE 42 Biological processes in which human genes associated to variation in height are involved compared with all genes in the genome.	201

Index of tables

1 | INTRODUCTION

Table 1 Traits associated with chromosomal polymorphism in several species.	30
Table 2 Effects on gene expression of TE insertions within genes.	42
Table 3 Genes adjacent to inversion <i>2j</i> breakpoints.	60

2 | MATERIALS AND METHODS

TABLE 4 <i>D. buzzatii</i> lines used in this study.	68
TABLE 5 Primers used in this study.	70

3 | RESULTS

Article – TABLE 1 <i>D. buzzatii</i> isogenic lines used in this study.	95
Article – TABLE 2 Structure of the CG13617 coding region in <i>D. buzzatii</i> and <i>D. melanogaster</i> .	96
Article – SUPPORTING TABLE 3 Primers used in this study.	102
Article – SUPPORTING TABLE 4 Structural variation in the <i>CG13617</i> gene region analyzed by PCR amplification and restriction enzyme digestion of PCR products.	103
Article – SUPPORTING TABLE 5 Real-time RT-PCR relative measurements of <i>CG13617</i> mRNA, antisense RNA, and <i>Pp1α-96A</i> mRNA expression levels in embryos of homozygous lines with and without inversion <i>2j</i> .	104
Article – SUPPORTING TABLE 6 Real-time RT-PCR relative measurements of <i>CG13617</i> mRNA and antisense RNA in embryos homozygous and heterozygous for inversion <i>2j</i> .	105

TABLE 6 List of differentially expressed genes determined by the combination of the results of the three different microarray analysis methods.	112
TABLE 7 Significantly enriched functional categories in the list of genes that are differentially expressed when <i>CG13617</i> is silenced.	116
TABLE 8 Genes partially sequenced in <i>D. buzzatii</i> prior to the expression analyses by real-time RT-PCR.	122
TABLE 9 <i>CG13617</i> orthologous genes in <i>Drosophila</i> species.	128
TABLE 10 Structure of gene <i>CG13617</i> in the different <i>Drosophila</i> species.	129
TABLE 11 <i>CG13617</i> protein identity and similarity matrix in <i>Drosophila</i> species.	132
TABLE 12 <i>CG13617</i> protein orthologues in species other than <i>Drosophila</i> identified using BLASTP.	139
TABLE 13 Additional <i>CG13617</i> protein orthologues in species other than <i>Drosophila</i> obtained using tBLASTN.	140
TABLE 14 Conserved non-coding sequences upstream of gene <i>CG13617</i> identified by VISTA in the alignment of <i>D. buzzatii</i> and <i>D. mojavensis</i> sequences.	143

Index of boxes

1 | INTRODUCTION

Box 1 | Classes of transposable elements 39

2 | MATERIALS AND METHODS

Box 2 | RNA interference 82

Box 3 | Microarrays 88

ACKNOWLEDGEMENTS | AGRAÏMENTS | AGRADECIMIENTOS

No voldria acabar sense donar les gràcies a totes les persones que durant aquests anys han estat implicades d'una manera o altra en l'elaboració d'aquesta tesi:

A Alfredo por darme la oportunidad de trabajar en su laboratorio, por confiar en mi capacidad para sacar adelante este proyecto, por tener paciencia conmigo, por dejarme hacer las cosas a mi manera y a la vez corregirme y ayudarme siempre que ha hecho falta. Y también por creerse desde el principio que había encontrado un RNA *antisense*.

To John McDonald for welcoming me into his lab where I learnt many new things, for the enthusiasm he showed for my project since day one, and for all the good ideas he contributed.

A tots meus companys al llarg d'aquests anys. A Alejandra, por haber traído tan buen rollo al laboratorio, por todos los cafés, las charlas y las lecturas compartidas. A la Sònia per estar sempre disposada a ajudar, per tots els cafès i els esmorzars i, mes recentment, les xerrades sobre nens i els consells sobre el format de la tesi. Vosaltres dues sou les que més m'heu hagut d'aguantar tots aquests anys. A la Cristina Santa, per estar sempre allà els primers anys d'aquest treball quan tot semblaven obstacles molt complicats. A la Mar, perquè va ser tota una experiència treballar juntes buscant *Galileos* amb la nostra bioinformàtica "artesanal". A Fernando Ayllón y Deodoro por la música y las tertulias mientras preparábamos PCRs en el lab. I també a l'Oriol i el Fernando Prada, que encara que van començar les seves tesis abans que jo, les han acabat abans. También a Antonio Barbadilla, por estar siempre dispuesto a resolver cualquier tipo de dudas y confiar en mi a la hora de dar clases.

To everybody at Georgia Tech. To DeEtte, for your smile, for being always so nice and keeping in touch, and for thinking of me every spring when the robin is in her nest again. To Lilya, for teaching me how to microinject the embryos and trusting me to do the microarray experiments. To Nathan, for his valuable help with the microarray analysis and for the conversations about our research. To Nalini and Nina, for all those lunches, conversations and laughs while we were working together in the office.

A la Montse Sales, per la teva simpatia i per ajudar-me sempre amb les mosques (i també per renyar-me sempre que m'oblidava de canviar-les, que han estat moltes vegades). A la Julia, la

Maite, la Conchi i la resta de personal de secretaria, per resoldre sempre qualsevol problema amb eficiència.

A totes aquelles persones que m'han ajudat en algun moment amb la feina experimental d'aquesta tesi. Potser aquestes col·laboracions han estat breus, però no per això menys importants. A l'Anna Barceló, per tots els seus consells sobre com millorar les meves seqüències, a l'Armand Sánchez i l'Olga Francino per deixar-me fer servir el seu aparell de PCR en temps real, al Juan Hidalgo i el Javier Carrasco per la seva ajuda amb els Northern blots, i al Jordi Casanovas i la Marta Llimargas per ensenyar-me a fer hibridacions *in situ* en embrions de *Drosophila*.

Al Departament d'Universitats, Recerca i Societat de la Informació de la Generalitat de Catalunya pels quatre anys de beca predoctoral FI que em van permetre començar aquesta tesi.

Als meus pares, la iaia i l'Anna, per ajudar-me d'una manera o una altra sempre que ha fet falta.

A la Sara i el Joan. La vostra arribada ha contribuït considerablement a endarrerir l'acabament d'aquesta tesi perquè vosaltres us heu convertit en el meu projecte més important. Acabar aquesta tesi ha estat tota una aventura gràcies a vosaltres. Sense vosaltres hauria pogut dedicar-hi més temps, però segur que no m'ho hauria passat tan bé!

Al Mario, per tantes coses que no sé per on començar. Per ensenyar-me a treballar al laboratori, per resoldre dubtes sempre que ha fet falta, per ensenyar-me a ser crítica fins i tot amb els meus propis resultats, i per animar-me quan el projecte no anava bé o els experiments no sortien. Però sobretot per estar sempre amb mi, sempre a prop, encara que sovint ens separessin milers de quilòmetres de distància. No sé si ho hauria fet sense tu i en qualsevol cas, no hauria après tantes coses ni hauria estat tan interessant.

A tots vosaltres, per totes aquestes i mil altres coses més, gràcies!

To all of you, for these and a thousand other things, thank you!

