Comparative genomics: chromosome and gene evolution in two cactophilic Drosophila species, *D. buzzatii* and *D. mojavensis*

DOCTORAL THESIS

Yolanda Guillén Montalbán

Universitat Autònoma de Barcelona

Comparative genomics: chromosome and gene evolution in two cactophilic Drosophila species, *D. buzzatii* and *D. mojavensis*

Genómica comparativa: evolución cromosómica y génica de dos especies cactófilas del género Drosophila, *D. buzzatii* y *D. mojavensis*.

Doctoral thesis

Yolanda Guillén Montalbán

Universitat Autònoma de Barcelona

Departament de Genètica i Microbiologia

The cover was kindly designed by Miguel Miranda.

Memòria presentada per la Llicenciada en Biotecnologia Yolanda Guillén Montalbán per a optar al grau de Doctora en Genètica.

Yolanda Guillén Montalbán

Bellaterra,   de Maig de 2014

El Doctor Alfredo Ruiz Panadero, Catedràtic del Departament de Genètica i Microbiologia de la Facultat de Biociències de la Universitat Autònoma de Barcelona,

CERTIFICA que la Yolanda Guillén Montalbán 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 "Comparative Genomics: chromosome and gene evolution in two cactophlic Drosophila species, *D. buzzatii* and *D. mojavensis*".

I perquè consti als efectes oportuns, signa el present certificat a Bellaterra, a   de Maig de 2014.

Dr. Alfredo Ruiz Panadero

Table of contents

1. ABSTRACT 1

RESUMEN 2

2. INTRODUCTION 5

2.1 Comparative Genomics 5

2.2 Drosophila and the beginning of the Genomic Era 11

2.3 Cactophilic Drosophila species 14

2.4 Genetic diversity 18

2.4.1 Genetic variation 18

2.4.2 Tracking natural selection in comparative genomics 21

2.4.3 Codon substitution models 23

2.5 The plasticity of the genome 25

2.5.1 Structural variations 25

2.5.2 Transposable elements and their impact on the genome 31

2.6 Emergence of new genetic functions 33

2.7 Regulatory changes in adaptive evolution 37

3. OBJECTIVES 41

4. RESULTS 43

4.1 Gene alterations at Drosophila inversion breakpoints provide prima facie evidence for natural selection as an explanation for rapid chromosomal evolution 43

4.2 Genomics of ecological adaptation in cactophilic Drosophila: hundreds of genes under positive selection in the *D. buzzatii* and *D. mojavensis* lineages 73

5. DISCUSSION 203

5.1 Facing a *de novo* genome assembly 203

5.2 Comparative genomics and evolution 205

5.3 Chromosomal inversions and their role in adaptation 207

5.4 TE role in genome evolution 210

5.5 Divergence patterns and genomic determinants of gene evolution 212

5.6 Inferring positive selection 217
5.7 From Genomics to Transcriptomics 221
5.8 GBrowse and web resource 223
6. CONCLUSIONS 227
APPENDIX 231
BIBLIOGRAPHY 235
Index of tables 259
Index of figures 259
ACKNOWLEDGEMENTS 261

A mis padres, mi hermana y

mi yaya

1. ABSTRACT

The genetic basis of ecological adaptation has been long investigated by exploring particular regions of the genomes, like chromosomal rearrangements, morphological polymorphisms or allozymes. The increasingly appreciated power of comparative genomics and the explosive number of sequenced genomes have offered the opportunity to better understand how molecular evolution relates to adaptation and phenotypic variation at the organismic level. Adaptive changes have been attributed to different genomic features including (i) changes in the coding sequences of the genes; (ii) gain or loss of functional genes; (iii) alterations of gene expression regulation; (iv) TE activity; and (v) chromosomal rearrangements. In this work we have focused on the adaptive value of two genomic features: chromosomal inversions and genes evolving under positive selection.

We first investigated seven inversions fixed in chromosome 2 of *D. mojavensis*, a cactophilic species that lives under extreme ecological conditions. Different mechanisms were found responsible for their generation, including TE-mediated ectopic recombination and breakage and repair by NHEJ. In addition important gene alterations were identified at some of the breakpoint regions, suggesting that natural selection was the main force driving the fixation of these inversions. Secondly we compared the genomes of two cactophilic flies, *D. buzzatii* and *D. mojavensis*, in order to characterize the patterns of protein-coding gene divergence between two species with a well-defined ecology. To accomplish this objective the genome of *D. buzzatii* was sequenced and annotated. Furthermore, we provided an overview of the transcriptional profile along the *D. buzzatii* development using RNAseq-based experiments. By using codon substitution models we have detected more than 1000 protein-coding genes evolving under positive selection, likely indicative of adaptive evolution.

RESUMEN

Las bases genéticas de la adaptación ecológica han sido investigadas durante muchos años mediante la exploración de regiones particulares del genoma tales como las reordenaciones cromosómicas, los polimorfismos morfológicos o las aloenzimas. El poder cada vez más apreciado de la genómica comparativa y el creciente número de genomas secuenciados ofrecen la oportunidad de comprender como se relacionan la evolución molecular, la adaptación y la variación fenotípica. Los cambios adaptativos han sido atribuidos a diferentes factores genómicos incluyendo (i) cambios en las regiones codificadoras de los genes; (ii) ganancia o pérdida de genes funcionales; (iii) alteraciones en la regulación de la expresión génica; (iv) actividad asociada a los elementos transponibles; y (v) reordenaciones cromosómics. En este trabajo nos hemos centrado en el valor adaptativo de dos factores genómicos: las inversiones cromosómicas y los genes sometidos a selección positiva.

En primer lugar se investigaron siete inversiones fijadas en el cromosoma 2 de *D. mojavensis*, una especie cactófila que vive bajo condiciones ecológicas extremas. Diferentes mecanismos son responsables de la generación de estas inversiones, incluyendo la recombinación ectópica entre elementos transponibles y la rotura y reparación por unión de extremos no homólogos (NHEJ). Asimismo se identificaron importantes alteraciones génicas en algunas regiones asociadas a los puntos de rotura. En segundo lugar se compararon los genomas de dos especies cactófilas, *D. buzzatii* y *D. mojavensis*, con tal de caracterizar los patrones de divergencia de los genes codificantes entre dos especies con una ecología bien definida. Para cumplir con estos objetivos, el genoma de *D. buzzatii* fue secuenciado y anotado. Además se analizó el perfil de expresión génica a lo largo del desarrollo de *D. buzzatii* usando experimentos basados en la tecnología del RNAseq. Finalmente, mediante el uso de modelos de sustitución de

codones se detectaron más de 1000 genes codificantes bajo selección positiva, probablemente indicativos de evolución adaptativa.

2. INTRODUCTION

2.1 Comparative Genomics

The comparison of genomes from different organisms has become a practical and powerful approach to understand the patterns of genome evolution. By comparing the sequence, structure and content of genomes we are able to detect the sources of molecular differences within and among species. Comparative genomics definitely provides an efficient tool for tracking evolutionary changes among organisms, allowing for the detection of highly conserved regions preserved from a common ancestor, as well as lineage-specific changes. Lately, the development of deep-sequencing-based technologies (Mardis 2008) has empowered the generation not only of DNA sequences but also of transcriptomes, i.e. the collection of all the RNA molecules produced in one or more cells, and their comparison between different species, individuals and even cell types (Wang et al. 2009). The increasing number of studies focusing on comparative transcriptomics at different levels has revealed that gene expression plasticity represents an important source for adaptive responses to environmental changes (Knight et al. 2006; Larsen et al. 2007; Smith et al. 2013).

Prior to the development of sequence-based approaches, other procedures were carried out to compare genomes based mainly on chromosomes observation. Karyotyping became one of the first techniques to compare genomes by examining the number, relative sizes and shapes of the chromosomes (Gregory 2011). With the availability of techniques that allow reading the nucleotide sequence of DNA molecules, computer-based comparison of multiple genomes have been done at a nucleotide level. Consequently, fascinating differences in the number of genes and DNA content among organisms have been reported (Table 1).

TABLE 1. Summary of genome properties of different organisms sequenced between 1996 and 2005.

Organism	Genome size (Mb)	Chromosome number	Estimated number of gene models	Reference
Escherichia coli	4.6	1	3200	(Blattner et al. 1997)
Saccharomyces cerevisiae **(unicellular yeast)**	12.4	32	6000	(Goffeau et al. 1996)
Caenorhabditis elegans **(nematode)**	100	12	19000	(C. elegans Sequencing Consortium 1998)
Arabidopsis thaliana **(mustard)**	157	10	25000	(Arabidopsis Genome Initiative 2000)
Oryza sativa **(rice)**	470	14	51000	(Goff et al. 2002)
Drosophila melanogaster **(fruitfly)**	165	8	13600	(Adams et al. 2000)
Gallus gallus **(chicken)**	1000	78	20000	(Hillier et al. 2004)
Canis familiaris **(domestic dog)**	2400	78	19000	(Lindblad-Toh et al. 2005)
Mus musculus **(mouse)**	2900	40	25000	(Waterston et al. 2002)
Homo sapiens **(human)**	3000	46	25000	(Lander et al. 2001)

Nowadays, genome size estimates for more than 4500 animals are available (Gregory 2014), 65% of them vertebrates; and a total of 18887 genome projects have been completed, including 330 archaeal, 17649 bacterial and 906 eukaryal genomes (Pagani et al. 2012). The smallest genome found so far is that of the microsporidian *Encephalitozoon intestinalis*, a useful model for exceptional genome compaction comprising only 2.3 Mb (Corradi et al. 2010). On the other side, the plant *Paris japonica* has the largest recorded genome, with 150000 Mb (Pellicer et al. 2010). Even so, the dramatic differences in terms of size and gene content reveal little about biological complexity, especially among eukaryotes (Gregory 2005a; Straalen 2012).

According to the C-value paradox, where C-value is the total amount of DNA in a haploid genome (Swift 1950), the complexity of an organism is not directly correlated with the number of genes nor with genome size (Thomas 1971; Hartl 2000; Gregory 2005b) (Figure 1). Different explanations have been proposed to disentangle this puzzling fact along the history (Lynch 2007). Today it is generally accepted that transposable elements (TEs) account for the major contribution to eukaryotic genome size variation, providing a partial explanation for the C-value paradox (Kidwell 2002). Indeed, TEs have been shown to comprise ~15% of the *D. melanogaster* genome (Kaminker et al. 2002; Bergman et al. 2006; Krassovsky and Henikoff 2014), and approximately half of the sequence content of a typical mammalian genome (de Koning et al. 2011). On the other hand, it has been suggested that the lack of correlation between complexity and DNA content seems to derive from a spotlighting on extreme outliers rather than a measure of central tendency (Lynch 2007), as evidenced by the clear ranking from viruses to prokaryotes to unicellular eukaryotes to multicellular eukaryotes in terms of genome size, gene and mobile element content and intron number and size.

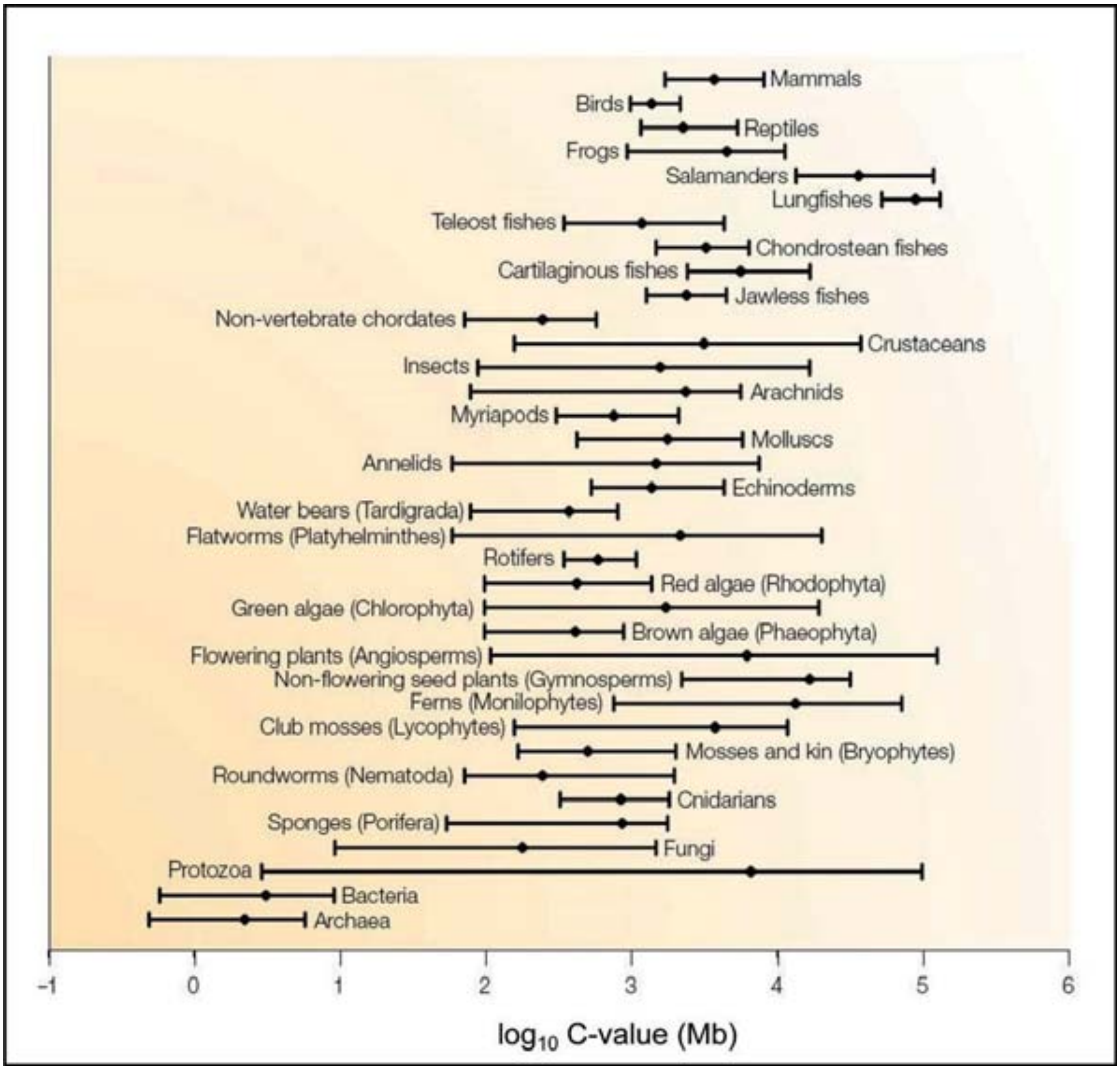


FIGURE 1. **Genome size variation within and among the main groups of life**. The mean and overall ranges of genome size for the main groups of living organisms are depicted. In prokaryotes it seems to exist a correlation between genome size and protein-coding gene content. However the vast majority of nuclear DNA in eukaryotes is non-coding. It has been apparent that genome sizes greatly vary within eukaryotes and thus this is not correlated to common ideas of both morphological and functional complexity. Figure extracted from Gregory (2005a).

The rising number of studies describing the transcription dynamics has disclosed that organisms complexity is correlated with transcriptome intricacy rather than DNA content (Adams 2008). Functional complexity is said to derive from the increasing number of mechanisms producing multiple mRNA variants from a single gene, like alternative splicing, RNA edition, transcript fusion or alternative initiation and termination sites. For example, the *Dscam* (Down Syndrome Cell Adhesion Molecule) gene found in Drosophila has 24 exons and presents more than 38000 isoforms differentially expressed in a wide variety of cell types and individual cells (Neves et al. 2004; Sawaya et al. 2008), and the regulation of the expressed variants is controlled by both spatial and temporal factors (Figure 2). In addition, several non-protein-coding sequences that are transcribed have been widely described (Eddy 2001), including microRNAs, snRNAs, piwiRNAs and lincRNAs (Griffiths-Jones et al. 2005; Mattick and Makunin 2006); and the content of non coding RNA (ncRNA) genes within a genome seems to scale with functional complexity (Mattick 2004). Finally, recent controversial analyses based on human genome content (ENCODE Project Consortium et al. 2012) have shown that the human genome is pervasively transcribed, calling for the need for a more RNA-centric viewpoint to understand the evolution of organism complexity.

The rapidly emerging field of comparative genomics and the accumulation of new genome sequences have already yielded impressive results that have fascinated the researcher's community, affecting multiple areas of Biology. Due to the easy and affordable accessibility to next generation sequencing (NGS) technologies, genomic information is rapidly accumulating in the public databases and so large-scale analyses are becoming the norm. For instance, obtaining the sequence of a human genome today (~3000 Mb) is a relative inexpensive task that a single researcher could do in a few weeks (Fox and Kling 2010). As a consequence, the exponential increase of public available genome sequences is becoming a challenge to massive store development.

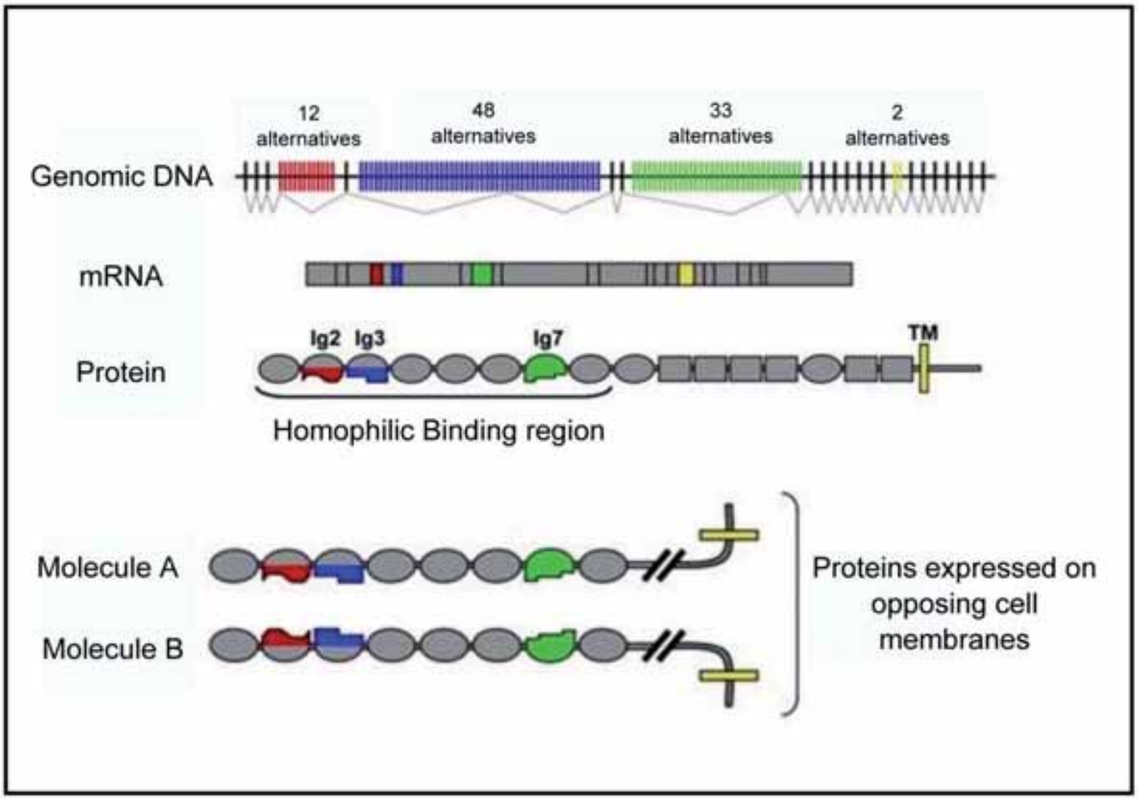


FIGURE 2. **The structure of the *Dscam* gene in *D. melanogaster***. *Dscam* is an essential gene for fruitfly development, involved in neuronal wiring and adaptive immunity system. *Dscam* locus is 61 kb long and comprises four exon clusters spliced in a mutually exclusive manner generating a repertoire of up to 38016 transcripts. Variable exon clusters are shown in colour: exon 4 cluster in red, exon 6 cluster in blue, exon 9 cluster in green and exon 17 cluster in yellow. Constant exons are shown as black boxes. *Dscam* encodes for a set of complex cell surface proteins comprising immunoglobulin (Ig) domains (ovals), fibronectin type III domains (rectangles), a transmembrane domain (yellow line), and a cytoplasmic tail. In essence, *Dscam is* an example of a genetic mechanism that leads to huge morphological and physiological diversification. Figure extracted from Sawaya et al. (2008).

In summary, genomic tools have made it possible to design genome-wide studies to deeply explore genetic changes accumulated in different genomes and to identify genetic traits responsible for adaptive evolution (Stapley et al. 2010). Furthermore, the integration of biogeography, field experimentation and long-term life history research with cutting edge genomics tools will make it possible to test and develop new theories and advance our understanding about adaptation. As a consequence, new objectives will arise in the study of comparative genomics like the effects of climate change on genetic variation, conservation of genetic resources and even crop and animal production improvement.

2.2 Drosophila and the beginning of the Genomic Era

Drosophila melanogaster is one of the most popular research tools in Biology that provided major theoretical and technical progresses in this field during the last century. Modern Drosophila Genetics first originated with Thomas Hunt Morgan's discovery of the white eye mutation and its X-linkage inheritance in 1910 (Morgan 1910). Indeed, he was the first geneticist to clearly link a trait inheritance to a specific chromosome.

Several reasons contributed to the election of *Drosophila melanogaster* as the central focus in the study of transmission genetics in the origins of the Modern Genetics (Hartwell 2011). First, its life cycle is relatively short, making it easy to obtain thousands of progeny in a short period of time (Figure 3). This little fruitfly also has huge salivary gland chromosomes exhibiting finer bands simply visible by microscope examination (Bridges 1935). Thus, they provided geneticists with a ready-made detailed physical map of the genome making it possible to identify chromosomal rearrangements with a high precision (Muller and Painter 1932; Horton 1938; Dobzhansky and Sturtevant 1938). Furthermore crossing-over events are restricted to Drosophila females, a phenomenon that was first discovered by T. H. Morgan in 1914 (Morgan 1914), though several exception exist (Philip 1944; Kale 1969; Hiraizumi 1971). This fact has greatly simplified several experimental manipulations allowing for a variety of selective genetic screens through generations.

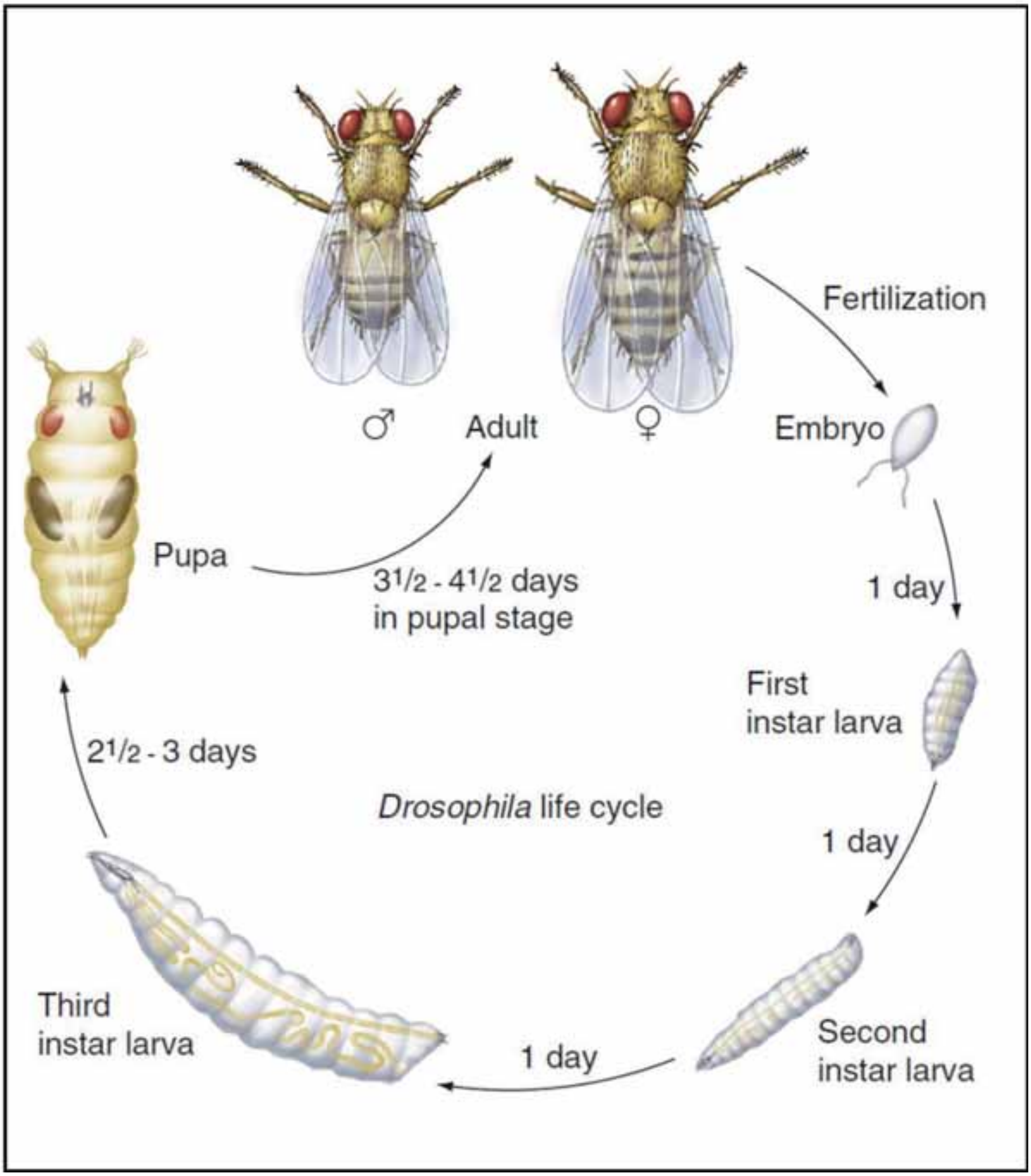


FIGURE 3. **The *Drosophila melanogaster* life cycle.** The transition from an embryo to a first instal larva is called hatching. The transitions between larval instars are molts. The process that converts a third instar larva to a pupa is pupariation. Emergence of the adult from the pupal case is called eclosion. The Drosophila life cycle is completed in approximately 12 days. Figure extracted from Hartwell (2011).

By and large *D. melanogaster* has been an important model organism not only for classical genetics but also for animal development (Lewis 1978; Kaufman et al. 1980) and behavior studies (Konopka and Benzer 1971) in the last decades. Indeed it has been described as “a little person with wings” since it was discovered that both human and fruitfly share a core set of genes, including ~60% of genes associated to human diseases (Schneider 2000). Thus, this tiny insect can even serve as a competent model for testing therapies targeting hereditary diseases. In summary *Drosophila* system has become an essential model in multiple research fields for a wide range of eukaryotic organisms.

The genome of *D. melanogaster* was the second metazoan genome to be sequenced (Table 1) (Adams et al. 2000; Rubin and Lewis 2000). Since the first publication of the *D. melanogaster* sequence in 2000, there have been subsequent genome releases that have incorporated quality and gene annotation improvements (The FlyBase Consortium 2002; Ashburner and Bergman 2005). Nowadays, the genome of *D. melanogaster* is considered one of the best characterized eukaryotic genomes at both, gene content and transcriptome levels (modENCODE Consortium et al. 2010; Graveley et al. 2011; Brown et al. 2014). Nowadays, more than 20 Drosophila genomes have been already sequenced and annotated (www.flybase.org/), providing a valuable resource to Comparative Genomics. The ecological diversity of the complete sequenced Drosophila genomes is staggering, including species inhabiting different geographical locations separated by a wide range of evolutionary distances (Drosophila 12 Genomes Consortium et al. 2007; Markow and O’Grady 2007; Singh et al. 2009; Russo et al. 2013) (Figure 4). This genomic data has made it possible to better understand the patterns of genome evolution in a fine-scale approach.

2.3 Cactophilic Drosophila species

The chemical ecology of insects has been the center of many studies focused on ecological genetics. Different species from Drosophila genus have been used as model organisms in several works about evolutionary genetics in the last century. The Drosophila genus is large and diverse with about 2,000 known species. Phylogenetic analyses indicate that two main lineages exist, which diverged 40-60 myr ago (Tamura et al. 2004). One lineage led to the Sophophora subgenus comprising more than 300 species, whereas the other one led to the subgenus Drosophila, with about 1700 species. Out of the 24 Drosophila genomes already sequenced and available in FlyBase (The FlyBase Consortium 2002), only five belong to the Drosophila subgenus: *D. virilis*, *D. mojavensis*, *D. grimshawii*, *D. americana* and *D. albomicans;* whereas the remaining nineteen species belong to the Sophophora subgenus.

The Drosophila subgenus includes the repleta group (Figure 5), which comprises many cactophilic species living in the necrotic stems of different cactus (Wasserman 1992; Oliveira et al. 2012). The fruitfly community inhabiting rotting tissues of these distinctive plants in arid zones provides a valuable model for gene-environment interaction and ecological adaptation comprehension (Barker and Starmer 1982; Etges et al. 1999; Fogleman and Danielson 2001).

Some Drosophila species are able to colonize cactus widely distributed along different geographical areas. However, specialists are restricted to certain environments and have limited growing conditions (Patterson and Stone 1953; Wasserman 1982; Vilela 1983). Niche specificity depends on a variety of ecological factors like the availability of nutrition resources or tolerance to toxic compounds present in the host plant (Heed 1978; Kircher 1982; Ruiz and Heed 1988). For instance, senita cactus (*Lophocereus schottii*) is the unique host plant of *Drosophila pachea*, one of the four endemic Drosophila species inhabiting the Sonora Desert (Heed 1978). This plant has a characteristic chemical composition making it impossible for other Drosophila species to

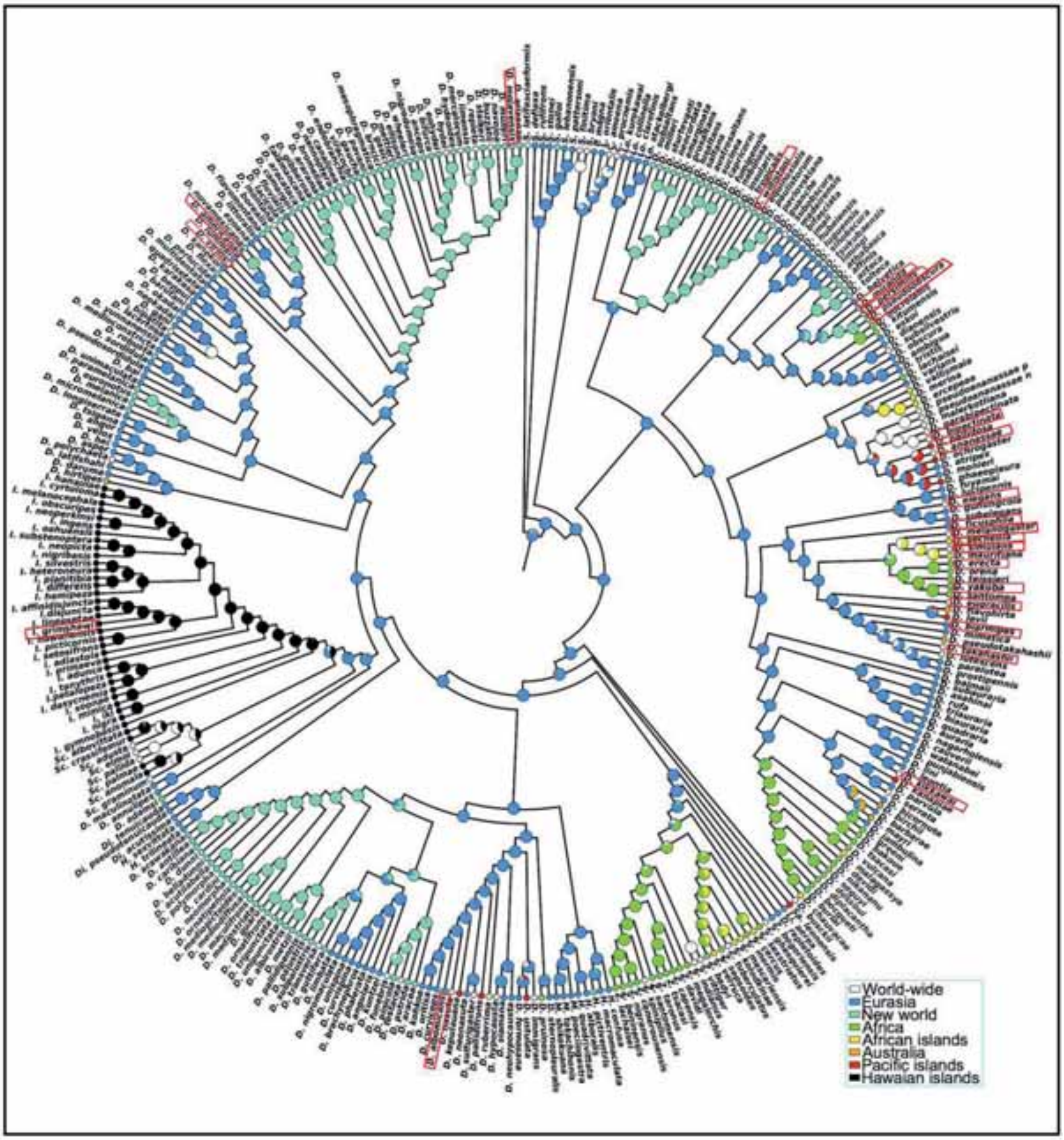


FIGURE 4. **Phylogenetic tree reconstructed from for a large drosophilid data set.** Both geographical distribution and phylogenetic relationships among Drosophila species representing up to 14 genera, help to infer the evolutionary history of this genus. Twenty-two out of the 24 drosophila species whose genome have been already sequenced are contained in red rectangles (*D. suzuki* and *D. rhopaloa* are not included in the tree). Figure modified from Russo et al. (2013).

inhabit it (Kircher et al. 1967). Lang et al. (2012) showed that few changes in nucleotide sequence of *Neverland* gene restricted the host plant of this fruitfly. These results evidenced that the ecological niche can be determined by little but crucial mutations.

Drosophila mojavensis, a specialist living in the deserts of SW United States and NW Mexico (Heed and Mangan 1986; Ruiz and Heed 1988; Etges et al. 1999), is composed of four ecologically distinct subspecies, and each of them feeds from nectrotic tissue of cactus with different chemical composition (Kircher 1982; Fogleman and Kircher 1986). The populations living in the Sonoran Desert feeds from agria (*Stenocereus gummosus*) and organ pipe (*Stenocereus thurberi*) cacti. In the Mojave and Anza-Borrego Deserts they use as a substrate necrotic tissues from barrel cactus (*Ferocactus cylindraceus)* (Fellows and Heed 1972; Heed 1978; Fogleman and Armstrong 1989). In Santa Catalina Island they feed from the fruits of *Opuntia "demissa"* cactus.

D. buzzatii, unlike its sibling *D. mojavensis*, is a widespread species found in many continents. It chiefly feeds and breeds in rotting tissues of cactus from Opuntia genus. The geographical diffusion of this plant by humans is considered the main cause of *D. buzzatii* world-wide colonization (Fontdevila et al. 1981; Barker and Starmer 1982; Hasson et al. 1992; Ruiz et al. 2000).

The karyotypes of both *D. mojavensis* and *D. buzzatii* consist of five pairs of rod chromosomes (2, 3, 4, 5, and X or Y) and a pair of dot chromosomes (6). The phylogenetic relationship between these two species was first inferred by combining both biogeographical and cytogenetical data (Ruiz et al. 1990; Ruiz and Wasserman 1993). Cytological-based studies showed that *D. mojavensis* had a relatively high rate of fixation of chromosomal rearrangements compared to other species of the repleta

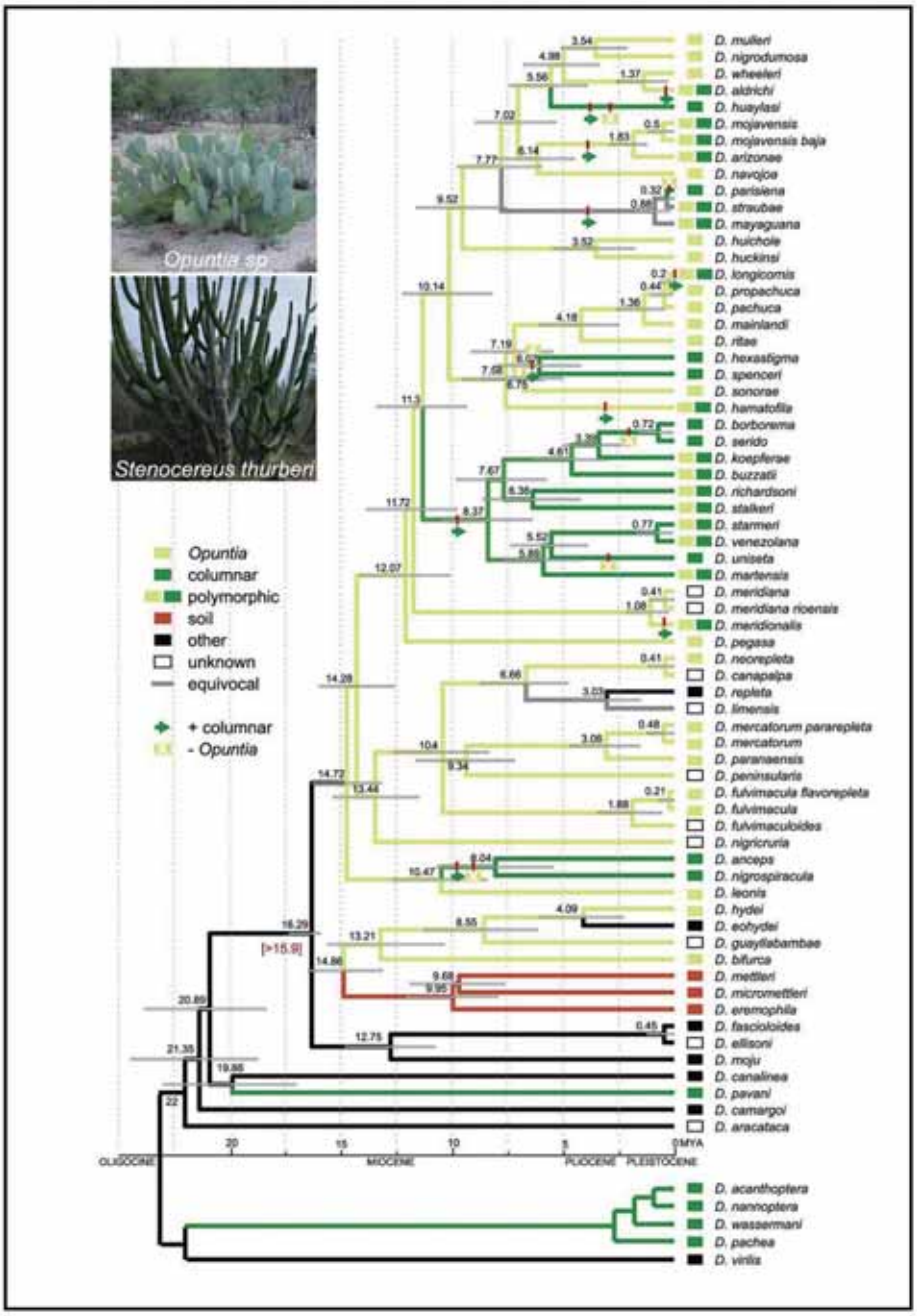


FIGURE 5. **Phylogenetic tree including species from repleta group.** Time estimates are depicted next to tree nodes and the bars represent their 95% confidence interval. Host substrates are color coded. "Soil" refers to cactus exudate-soaked soils, and "other" refers to other substrates, but not cactus. Typical Opuntia and columnar cactus growth forms are represented in the top left pictures. Figure extracted from Oliveira et al. (2012).

group (Ruiz et al. 1990; González et al. 2007). Nowadays *D. mojavensis* is the only cactophilic species whose genome has been sequenced and annotated (Drosophila 12 Genomes Consortium et al. 2007). The genome sequence of this fruitfly has been included in several genome-wide studies that explored the gene and chromosome evolution within Drosophila genus (Drosophila 12 Genomes Consortium et al. 2007; Heger and Ponting 2007; Bhutkar et al. 2008; Singh et al. 2009). In addition, *D. mojavensis* has been used as an excellent model to examine the role of transcriptional differentiation in ecological adaptation (Matzkin 2012; Matzkin and Markow 2013).

2.4 Genetic diversity

2.4.1 Genetic variation

Genetic variation is considered the raw material for biological evolution. It is ultimately originated by mutations, i.e. changes that randomly occur in DNA molecules by multiple causes (errors in DNA replication, TE activity, exposure to ionizing radiation, mutagenic chemicals or infection by viruses) that can be transmitted through successive generations. Mutations occur at different scales, including single changes in the nucleotide sequence of a gene as well as chromosomal rearrangements, which encompass many classes of events such inversions, insertions, deletions or translocations (Hartl and Clark 1997) (Figure 6).

The fate of mutations is driven by multiple forces, chiefly natural selection and genetic drift. Recombination joins mutations of different genomic regions together into the same chromosome, generating new combinations of alleles. Mutations are also spread among different populations by migration, resulting in the addition of new alleles to the gene pool of a particular population.

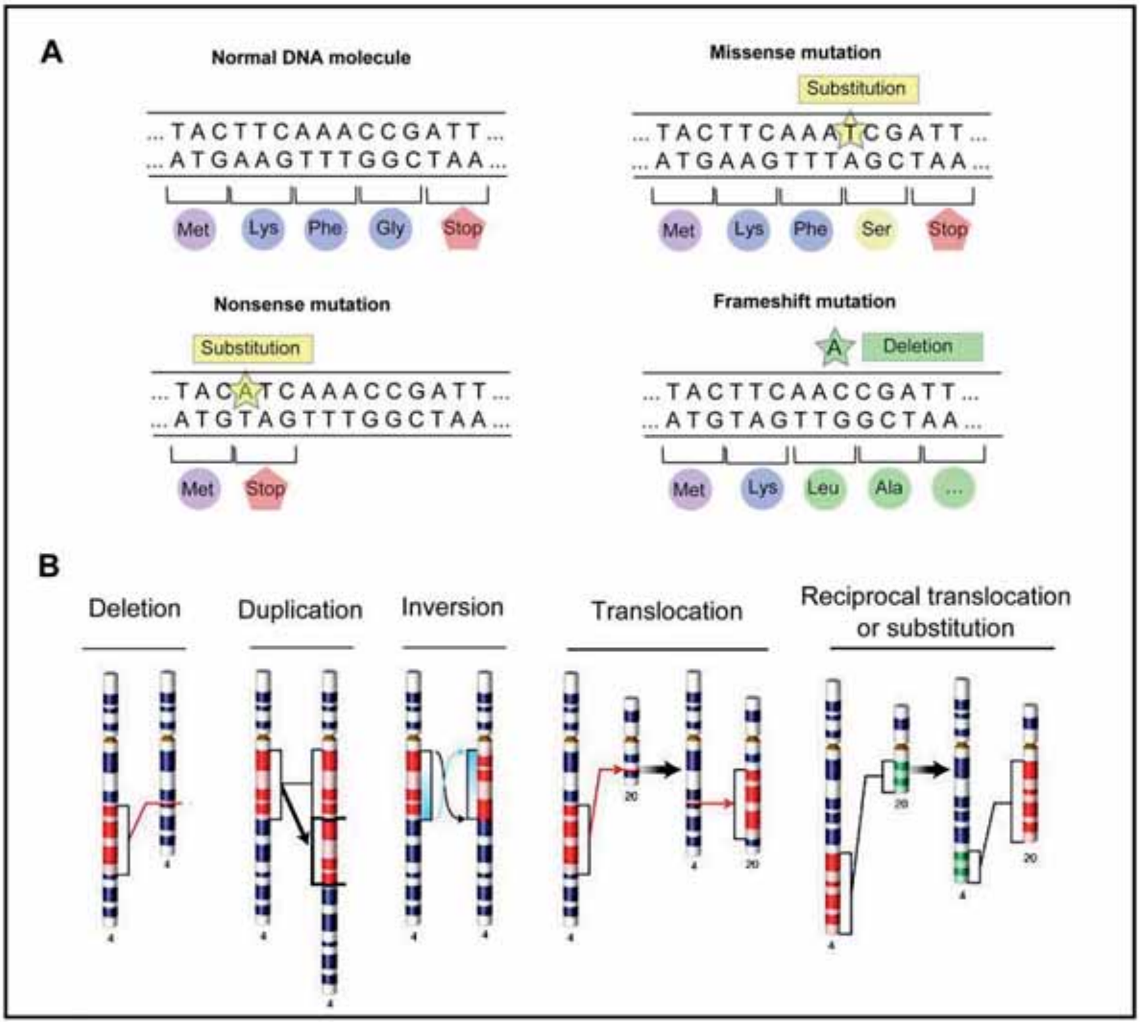


FIGURE 6. **General classification of DNA mutations.** Mutations can occur at a nucleotide level (A) or can involve larger portions of the genome resulting in chromosomal rearrangements (B). Point mutations (deletions, insertions or substitutions) can affect the coding region of a gene altering the protein function. Missense mutations refer to the substitution of a different amino acid in the protein, which can alter or not its functionality. Mutations that cause the appearance of a premature stop codon within a coding gene are called nonsense mutations. They lead to the production of a shortened and likely nonfunctional protein. Finally frameshift mutations are caused by a nucleotide deletion or insertion that shifts the way the coding sequence is read. Figure (B) modified from National Human Genome Research Institute website (www.genome.gov).

Mutations can be classified according to their impact on individuals' fitness into deleterious, neutral and advantageous. Deleterious mutations are those that negatively impact on the individuals' ability to reproduce and they are rapidly removed by natural selection (purifying selection) in large populations. By contrast, beneficial mutations improve individuals' fitness and they are rapidly fixated by natural selection (positive selection) in large populations (see below). According to the neutral theory of molecular evolution (Kimura 1968, 1983), which attempts to describe the dynamics of molecular polymorphism within a population, most observed polymorphisms are neutral. Neutral mutations (or selectively neutral) do not influence the individuals' fitness, and their frequency within populations only depends on genetic drift, a stochastic process by which genetic variants are fixed or removed from the population by random. Thus, Kimura's theory postulates that neutral divergence among species only depends on divergence time and mutation rate (μ), i.e. the rate at which changes are incorporated in a nucleotide sequence during replication.

The nearly neutral theory of molecular evolution (Ohta 1973), a modification of the original neutral theory proposed by Kimura (1968), assumes that (i) each mutation is associated to a particular selection coefficient (s), which is a measure of the relative fitness of the mutation (from s=0 denoting neutrality to s=1 complete lethality), and (ii) the rate of molecular evolution depends on the effective population size (N_e) (Lynch 2007). Accordingly the probability of fixation of a certain mutation depends on two factors: its selective coefficient and the population size. In large populations, the probability of fixation for beneficial mutations is higher than in small populations, whereas a considerable accumulation of fixed mildly deleterious mutations in populations with lower N_e is expected (Lynch 2007). Thus, at low N_e, selection is less efficient in removing disadvantageous mutations, with genetic drift leading to the fixation of mildly deleterious variants, and selection against deleterious mutations is strong only if they reduce fitness by $s >> 1/4N_e$.

2.4.2 Tracking natural selection in comparative genomics

The rapid accumulation of molecular sequence data allows for the detection of natural selection footprint at a genomic scale. The development of large-scale methods for comparative analysis of DNA and protein sequences enables to minimize the stochastic effects inherent to small sequence samples (Ellegren 2008). Thus, the genome-wide estimation of selection pressures helps to better understand how natural selection operates in different lineages and in relation to different life histories.

In order to identify the selective forces acting on protein-coding genes it is essential to establish a correct orthology relationship between genes from species to be compared. Orthology is defined as the relationship between homologous genes that arose by speciation at their most recent point of origin (Fitch 1970). The inference of orthologous genes tends to be a difficult task since there are different homologous relationships between genes beyond orthology, such as paralogy or co-orthology, terms that can be easily confused (Kristensen et al. 2011) (Figure 7). When two genes diverged after a duplication event within the same species they are said to be paralogous. However, gene duplications following the speciation create two or more genes in one lineage that are, collectively orthologous to one or more genes in another lineage, and they are denoted as co-orthologs (Koonin 2005). The prevalence of complex evolutionary events makes it difficult to assess orthologous, paralogous and co-orthologous genes in genomes containing large gene families.

Genes or regions of the genome that are affected by negative or purifying selection are highly conserved, whereas an accelerated evolution is indicative of positive or Darwinian selection. The most common test to detect signatures of adaptive evolution is based on the count of nucleotide substitutions observed when aligning protein-coding gene sequences from different species. This statistical method based on divergence data is known as ka/ks or dn/ds (ω ratio) test (Yang and Bielawski 2000), and it has been widely used to scan for positive selected genes on many lineages from both prokaryotic and eukaryotic organisms (Waterston et al. 2002; Richards et al. 2005; Nielsen et al. 2005; Petersen et al. 2007).

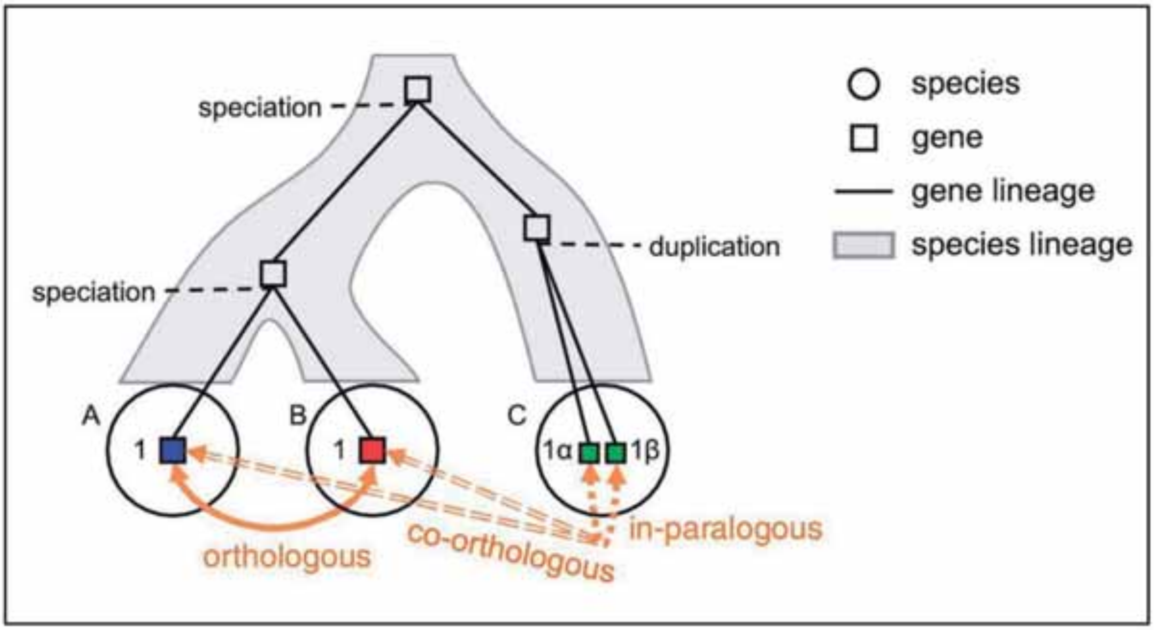


FIGURE 7. **Different evolutionary relationships among genes**. A, B and C represent three hypothetical species that have diverged from a single common ancestor. Genes that arise from a duplication event within a species (1α and 1β) are said to be in-paralogs. Homologous genes from related species that have diverged from a common ancestor are orthologs (1 from A and 1 from B). Orthologous genes are co-orthologs of homologous genes duplicated in related species. Figure modified from Kristensen et al. (2011).

When aligning sequences of the same protein-coding gene from two species (orthologs) we can observe two types of nucleotide substitutions. The differences that lead to changes in the amino acids of the encoded proteins are said to be nonsynonymous and they occur at nonsynonymous positions. Ka (or dn) is then defined as the number of nonsynonymous substitutions per nonsynonymous site. However, some differences leave the protein unchanged because of the degeneracy of the genetic code. They are called synonymous or silent changes and they occur at synonymous positions. Then, the number of synonymous substitutions per synonymous site is denoted by Ks (or ds). Synonymous and nonsynonymous mutations are under very different selective pressures and are fixated at different rates (Kimura 1977; Miyata and Yasunaga 1980). Thus the Ka/Ks statistics or ω ratio can reveal the direction and strength of natural selection acting on the gene.

Assuming that synonymous substitutions are neutral (because they do not affect the protein sequence and we do not expect them to affect the protein functionality), we can consider that a gene has undergone adaptive or positive selection if ω is higher than 1. This implies that nonsynonymous changes have been fixated at a higher rate than synonymous mutations as they provided a fitness advantage to the protein. However, most positions in functional genes are conserved, and the average value of ω tends to be much lower than 1, even in genes that have experienced positive selection in many sites (Figure 8), and thus we strictly infer that they evolve under purifying selection. On the other hand, genes are said to evolve neutrally when ω = 1, i.e. the likelihood that a nonsynonymous mutation is fixated is the same as that for a synonymous mutation. However, if one part of the gene experienced positive selection whereas others evolved under purifying selection, we might get also an average ω = 1. To account for this fact, more powerful methods have been developed to scan for positive selection at the codon level (Nielsen and Yang 1998; Yang et al. 2000; Lindblad-Toh et al. 2011, Villanueva-Cañas et al. 2013), revealing much more positive selection than previously suspected.

2.4.3 Codon substitution models

Although the ω ratio is a useful method to identify genes evolving under positive selection, it is considered a conservative test as it only accounts for an overall selective pressure. Codon substitution models were originally developed to consider heterogeneous ω ratios among amino acid sites using phylogenetics analyses of protein-coding DNA sequences (Goldman and Yang 1994; Muse and Gaut 1994). These statistical models, implemented in the package PAML (Yang 2007), consider the evolution of codons on a phylogeny of species using a maximum likelihood framework, allowing for heterogeneous ω ratios not only among sites (site models) but also among branches (branch site models).

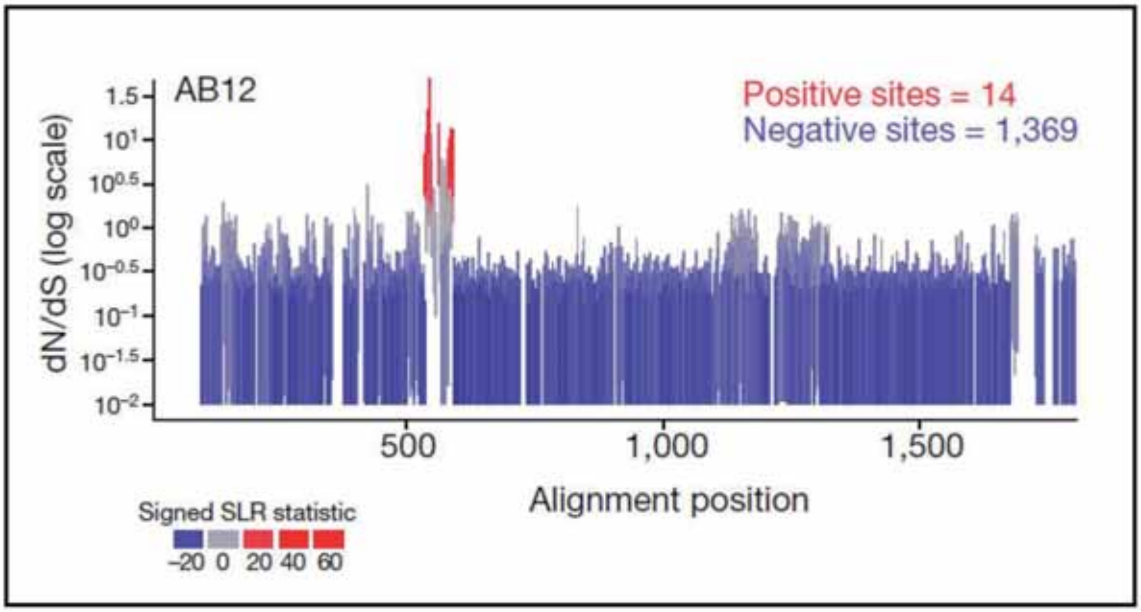


FIGURE 8. **Divergence ratio distribution along *AB12* gene sequence**. The alignment of AB12 gene sequences contained in the genomes of 29 mammals reveals that localized regions of genes may evolve under positive selection even detecting an overall negative selection. Bars are colored according to a signed version of the simple linear regression (SLR) statistic for non-neutral evolution: sites under positive selection (red), sites under purifying selection (blue) and neutral sites (grey). Figure modified from Lindblad-Toh et al. (2011).

By comparing the likelihood of the data under multiple models that make different assumptions about how ω varies among sites or among lineages, we can test different evolutionary hypotheses (Yang 2002). However, these statistical models assume that i) silent substitutions are always neutral and ii) the mutational process is at equilibrium, which are premises rarely true in real data (Sharp et al. 1995; Hartl and Clark 1997; Plotkin and Kudla 2011). However it has been reported that these assumptions do not bias the detection of positive selection (Larracuente et al. 2008). Codon substitution models have been successfully applied to screen for positive selection in a wide variety of organisms, including viruses (Zanotto et al. 1999; Fares et al. 2001), prokaryotes (Farfán et al. 2009) and eukaryotes (Swanson et al. 2001; Drosophila 12 Genomes Consortium et al. 2007; Amemiya et al. 2013; Ometto et al. 2013).

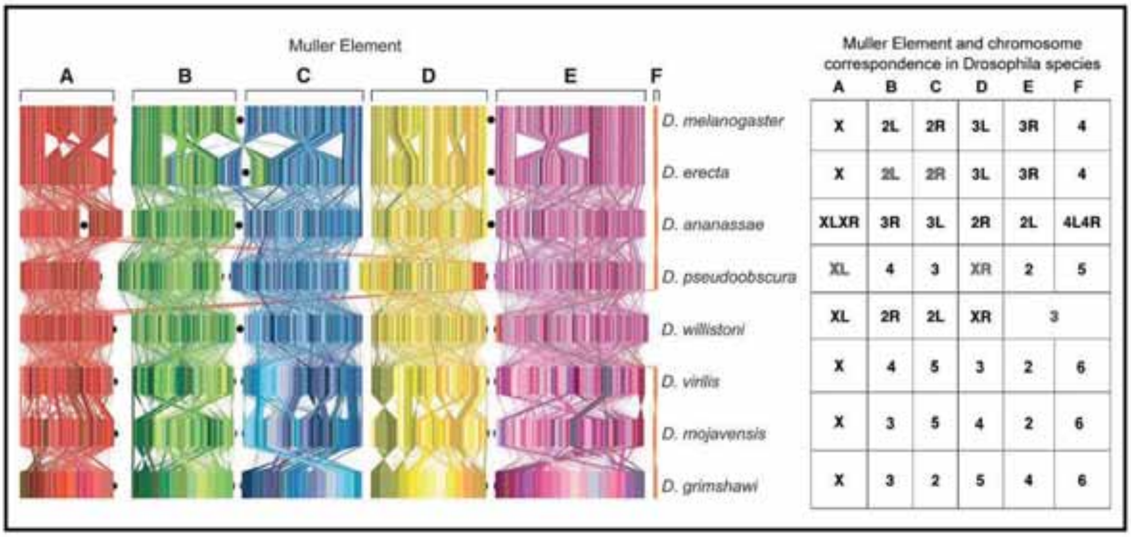
2.5 The plasticity of the genome

2.5.1 Structural variations

Structural variation (SV) is the variation in structure of an organism's chromosome. Structural variants can be classified into different types: insertions, deletions, copy number variations (CNVs), inversions or translocations (Figure 6). It has been reported that SV is pervasive and important in genome evolution, making significant contributions to genetic diversity and even disease susceptibility (Feuk et al. 2006). The rate at which chromosomal rearrangements are fixated within populations radically varies among species. It has been observed that fruitfly genomes evolve up to five order of magnitude faster than the most dynamic plant genomes included in the *Arabidopsis-Brassica* clade (Ranz et al. 2001). In turn, *Caenorhabditis* chromosomes have a faster rearrangement rate than those of *Drosophila* (Coghlan and Wolfe 2002). Different factors have been suggested to influence the fixation rate of structural variants in Drosophila, like generation time, population size, mutation rate (caused for example by the activity of transposable elements), and the meiotic cost of infertility in heterozygotes (Krimbas and Powell 1992; Coghlan et al. 2005; Hoffmann and Willi 2008). The large-scale analysis of chromosomal rearrangements of the complete sequence of 12 Drosophila genomes revealed that rearrangements fixation rate clearly differ among Drosophila lineages (Drosophila 12 Genomes Consortium et al. 2007; Bhutkar et al. 2008) (Figure 9). Finally variation in the number of fixed rearrangements is also observed between chromosomal elements, i.e. some chromosomes are able to accumulate multiple rearrangements whereas no rearrangements are observed in others (Bhutkar et al. 2008). The causes of these phenomena remain still unclear since no convincing hypotheses have been suggested to explain them.

Chromosomal inversions

Chromosomal inversions occur when a chromosomal segment that may include one or more genes breaks in two places defined as breakpoints. This segment -which can span a few kb or cover a substantial part of a chromosome arm-, is then re-inserted in the chromosome joining the two end fragments, acquiring a new orientation (Figure 6). Paracentric inversions are those that do not include the centromere because the breakpoints occur on the same arm, whereas pericentric inversions do span the centromere. Inversions are highly abundant in species from Drosophila genus, and the breakpoints of different polymorphic (Table 2) and fixed inversions (Cirera et al. 1995, Ranz et al. 2007; Runcie and Noor 2009; Prazeres da Costa et al. 2009; Calvete et al. 2012) have been already characterized at a molecular level.



Muller Element and chromosome correspondence in Drosophila species

	A	B	C	D	E	F
D. melanogaster	X	2L	2R	3L	3R	4
D. erecta	X	2L	2R	3L	3R	4
D. ananassae	XLXR	3R	3L	2R	2L	4L4R
D. pseudoobscura	XL	4	3	XR	2	5
D. willistoni	XL	2R	2L	XR	3	
D. virilis	X	4	5	3	2	6
D. mojavensis	X	3	5	4	2	6
D. grimshawi	X	3	2	5	4	6

FIGURE 9. **Overview of rearrangement events ocurred during the divergence of eight Drosophila species.** Vertical lines correspond to single genes, which are connected among different species according to the movement they have undergone as a consequence of the rearrangements. Muller Element and chromosome correspondence is represented next to each species' name. The vast majority of rearrangements occurred within a chromosomal arm, though several exceptions are observed. Figure modified from Bhutkar et al. (2008).

TABLE 2. Summary of polymorphic inversions with characterized breakpoints in Drosophila and Anopheles.

Species	Inversion	Breakpoint	Mechanism	Reference
D. melanogaster	*In(3L)Payne*	Lacking of repetitive sequences (including TEs)	Chromosomal breakage and NHEJ	(Wesley and Eanes 1994)
	In(2L)t	Lacking of repetitive sequences (including TEs)	Chromosomal breakage and NHEJ	(Andolfatto and Kreitman 2000)
	In(3R)Payne	Inverted duplications	Chromosomal breakage and NHEJ	(Matzkin et al. 2005)
D. buzzatii	*2j*	TE insertions	Ectopic recombination	(Cáceres et al. 1999, 2001)
	$2q^7$	TE insertions	Ectopic recombination	(Casals et al. 2003)
	$2z^3$	TE insertions	Ectopic recombination	(Delprat et al. 2009)
D. pseudoobscura	*Arrowhead*	128 and 315-bp repetitive sequences	Ectopic recombination	(Richards et al. 2005)
D. subobscura	*3O*	Lacking of repetitive sequences (including TEs)	Chromosomal breakage and NHEJ	(Papaceit et al. 2013)
A. gambiae	*2Rd'*	TE insertion	Unknown	(Mathiopoulos et al. 1998)
	2La	Inverted duplications and TE insertion	Unknown	(Sharakhov et al. 2006)
	2Rj	Segmental duplications	Ectopic recombination	(Coulibaly et al. 2007)

Inversions are mainly generated by two mechanisms: ectopic recombination (or non-allelic homologous recombination, NAHR) (Cáceres et al. 1999; Coulibaly et al. 2007) and chromosomal breakage and erroneous repair by non-homologous end-joining (NHEJ) (Sonoda et al. 2006; Casals and Navarro 2007) (Figure 10). Polymorphic inversions can be cytologically identified in Drosophila and other Diptera by examining the banding pattern of salivary gland chromosomes (Ruiz et al. 1990; Ruiz and Wasserman 1993). Inverted and noninverted (standard) forms of chromosomes usually coexist within the same population (Krimbas and Powell 1992). The chromosomal pairing between inverted and standard rearrangements generates the formation of characteristic loops clearly detectable by microscope observation. On the other hand, lineage-specific inversions, i.e. rearrangements that have been fixated in a species, can be cytologically detectable by comparing the order and orientation of chromosomal bands from different species.

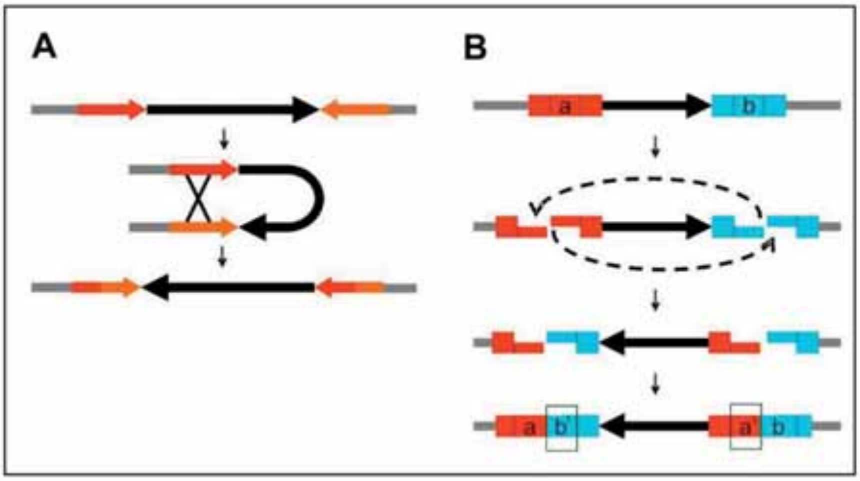


FIGURE 10. **Chief mechanisms that generate chromosomal inversions**. Ectopic recombination (A) and chromosomal breakage and erroneous repair by NHEJ (B) are two of the proposed mechanisms that originate inversions. Black arrows represent the chromosomal fragment involved in the inversion. In (A) red and orange arrows represent repetitive sequences (segmental duplications or TEs). In (B) the non-homologous regions are represented as blue and red rectangles. Single staggered breakages occurred at both breakpoints, resulting in the duplication of the unique sequences a' and b' distanced from the respective parental copies (a and b) by the inversion. Figure modified from Casals and Navarro (2007).

To test for the presence of chromosomal inversions at a fine-scale, different experimental approaches have been developed (Bailey et al. 1996; Iafrate et al. 2004; Tuzun et al. 2005; Redon et al. 2006; Korbel et al. 2007b). Although methods based on *polymerase chain reaction* (PCR) (Saiki et al. 1988) have been widely used in the last years to scan for chromosomal inversions along genome sequences, they are laborious and do not allow for the detection of small and/or *a priori* unknown inversions since a previous design of proves to target the rearrangement location is needed.

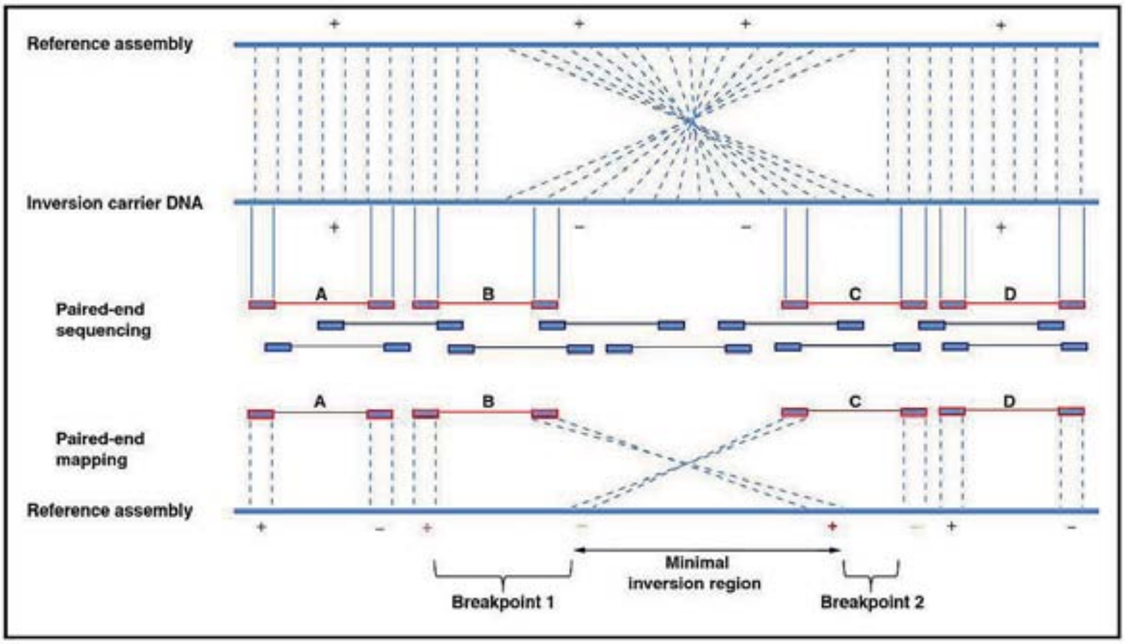


FIGURE 11. **Detection of a chromosomal inversion by paired-end mapping (PEM)**. An inversion can be characterized by aligning paired-end sequences from a genome containing the inversion (inversion carrier DNA) against a genome with the standard arrangement (Reference assembly) (or vice-versa). Figure modified from Feuk (2010).

With the recent advance of high-throughput DNA sequencing technologies and computational algorithms, new large-scale and powerful methods have been applied to identify chromosomal inversions reporting successful results (Medvedev et al. 2009). One of the most popular techniques is called paired-end mapping (PEM), a recent approach associated to NGS technologies that enables the identification of hundreds of structural rearrangements rapidly together with sophisticated algorithms that interpret the PEM data (Korbel et al. 2007a; Feuk 2010) (Figure 11).

Inversions and adaptive evolution

Chromosomal inversions are thought to play an important role in adaptive evolution and speciation (Rieseberg 2001; Coghlan et al. 2005), not only in animals, including insects (Feder et al. 2003; Joron et al. 2011; Ayala et al. 2011), fish (Jones et al. 2012) and mammals (Coghlan et al. 2005; Stefansson et al. 2005), but also in plants (Lowry and Willis 2010). Several studies have provided compelling evidence of the adaptive significance of polymorphic chromosomal inversions in Drosophila. These evidences include latitudinal clines, alterations of inversion frequency associated to seasonal and long-term environmental changes and even correlation between inversion and quantitative traits like body size and developmental time (Krimbas and Powell 1992; Powell 1997; Hoffmann et al. 2004). Thus, it is conceivable that inversion fixation within populations can be also driven by natural selection and not only depends on genetic drift.

Several hypotheses have been put forth to explain the adaptive significance of chromosomal inversions (Hoffmann and Rieseberg 2008). Some of them are based on the reduction of recombination within the inverted segment that occurs in heterokaryotypes. The coadaptation hypothesis (Dobzhansky 1970) postulates that the recombination reduction associated to inversions helps to maintain positive epistatic interactions within local populations. This implies that the allele combination trapped by the inversion likely have higher fitness than that predicted from the sum of their independent effects. A different but not excluding hypothesis is the local adaptation hypothesis (Kirkpatrick and Barton 2006). According to this hypothesis, inversions are favored even without epistasis because reduced recombination in inversions heterokaryotypes joins together locally adapted alleles and stabilizes them against gene exchange with immigrant chromosomes.

The position effect hypothesis proposes that the adaptive value of an inversion depends on fitness effects caused by breakpoints or position effects (Sperlich and Pfreim 1986; Puig 2011). Inversions can alter the functionality of genes adjacent to breakpoints by disrupting their nucleotide sequence, modifying their associated regulatory elements or even generating new genetic material (Ranz et al. 2007). But only a few genetic disorders associated to inversion position effects have been yet discovered in humans and Drosophila. For example, in *Drosophila melanogaster*, the Antp73b inversion mutation results in *Antp* transcription in an abnormal location (Frischer et al. 1986). Puig et al. (2004) and Puig (2011) also demonstrated the existence of a position effect caused by the *2j* inversion in *Drosophila buzzatii*, presumably resulting in phenotypic differences in body size and developmental time. Finally in humans, the principal cause of the severe haemophilia A disease has been attributed to an inversion that alters the coding region of factor VIII gene (Lakich et al. 1993). Moreover inversions can down-regulate or silence a gene by moving it to a heterochromatic region, an effect known as position effect with variegation (Henikoff 1990).

The three hypotheses mentioned above (co-adaptation, local selection and position effects) are not mutually exclusive, and all of them can jointly influence the fate of an inversion within a population.

2.5.2 Transposable elements and their impact on the genome

One of the main contributors to the eukaryotic genome plasticity is transposable elements (TEs) activity (Cordaux et al. 2006). TEs are DNA fragments that move from one location in the genome to another. They are found in many eukaryotic species, and their abundance and variety is considerable (Wicker et al. 2007). TEs are classified into two groups: retrotransposons and DNA transposons. Retrotransposons are able to copy themselves using an RNA intermediate, whereas DNA transposons can excise themselves out of the genome and be re-inserted somewhere else without the help of and RNA intermediate.

TEs are an important cause of mutations, basically insertions and deletions, and they are considered potential sources of adaptive selection (Casacuberta and González 2013). Although TEs usually do not encode cellular proteins, genomes can acquire new genes by recruiting them, a process called TE protein domestication, which has been observed in Drosophila (Casola et al. 2007) and in mammals (Casola et al. 2008). Moreover, TEs can positively or negatively impact on gene functionality depending on the genome site at which they are inserted. An insertion of a TE within a coding sequence will likely affect the gene fitness by truncating its product due to alterations in the associated reading frame. However remarkable exceptions exist, like the adaptive insertion of a Doc element within a Drosophila gene sequence, leading to a new coding gene associated to pesticide resistance (Aminetzach et al. 2005). On the other hand, the insertion of TEs in intronic sequences is expected to have less impact on gene functionality. Nevertheless, abnormal splicing events can occur as a result of these insertions.

Active transposable elements not only produce mutations at a structural level, including inversions mediated by ectopic recombination (see above), but they can also lead to nucleotide changes affecting gene expression. The insertion of TEs within regulatory elements in the genome may cause alterations in gene regulation by, for example, up- or down- regulating gene expression or modifying the tissue-expression pattern (Lerman and Feder 2005; Romanish et al. 2007). Another role attributed to TEs is the so-called process ‘exaptation’, by which traces from inactive TEs acquire new regulatory functions highly conserved among genomes (Muotri et al. 2007).

All these evidences suggest that TEs are important factors shaping the genome through evolution rather than selfish and parasite sequences. The important impact of TEs in the genome is rapidly being demonstrated thanks to the large-scale analysis and the availability of huge amount of genome sequences.

2.6 Emergence of new genetic functions

The origin of new genes is a source of evolutionary innovation in all organisms (Toll-Riera et al. 2009; Long et al. 2013). New genes usually take on novel biological functions that allow individuals coping with new niches and changing environmental conditions. By and large they are considered to mediate, jointly with protein-coding gene mutations and changes in regulatory regions, habitat-specific adaptations (Figure 12) (Long and Langley 1993; Begun 1997; Nurminsky et al. 1998; Khalturin et al. 2009; Long et al. 2013).

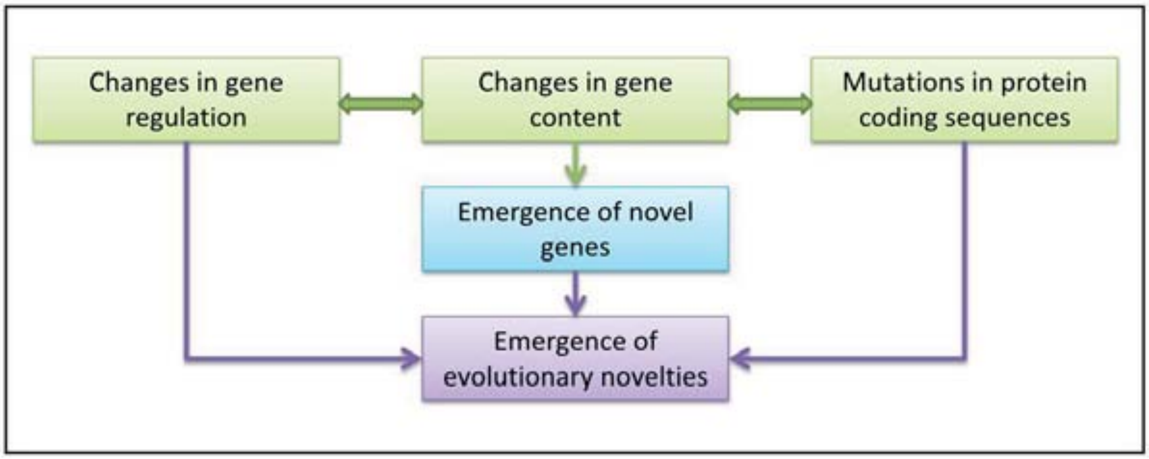

Changes in gene regulation
Changes in gene content
Mutations in protein coding sequences
Emergence of novel genes
Emergence of evolutionary novelties

FIGURE 12. **Overview of genomic changes that lead to evolutionary novelties.** Different genetic alterations, including changes in gene structure and regulation, and new genes lead to new functions.

It has been reported that ~10-20% of genes contained in eukaryotic genomes are novel genes because they do not present any significant sequence similarity to genes of other known species (Khalturin et al. 2009). Thus, new genes are commonly named orphans or taxonomically-restricted genes (TRGs)(Wilson et al. 2005). There exist multiple mechanisms responsible for the arising of new genes, not only protein-coding genes but also non-coding RNAs (ncRNA) (Long et al. 2003). Some of them are summarized below.

Gene duplications

New genetic material usually arises as a product of chromosomal abnormalities. Gene duplication is one of the most recurrent mechanisms that originated novel genes (Ohno 1970). Duplications occur when a DNA fragment is duplicated. Duplicated regions can involve one or many genes or even the whole genome of an individual (polyploidy), a phenomenon more common in plants than in other organisms (Adams and Wendel 2005; Cui et al. 2006). The main mechanisms causing DNA duplications are ectopic recombination, duplication-dependent strand annealing (DDSA) (Fiston-Lavier et al. 2007), DNA duplicative transposition (Bailey and Eichler 2006) and retrotransposition (Cordaux and Batzer 2009). According to the original theory of Ohno (1970), a new duplicated gene can acquire new and beneficial functions distinct from those of the original copies. However the classic model also predicted that a duplicate gene can lose its function (pseudogenization) because of the accumulation of deleterious mutations in one of the copies balanced by the initial functional redundancy (Lynch and Walsh 1998). Duplicated genes can be preserved in genomes by natural selection, and it can be explained by the functional divergence process. The adaptive radiation model predicts that the preservation of a duplicated gene is favored by the increased dosage compensation of a gene product which can lately take on new functions different from that retained by the original copies by accumulating adaptive mutations (neofunctionalization) (Long et al. 2013). On the other hand, original genes and new duplicated copies can retain a subset of the original ancestral function, i.e. the original functional capabilities are divided among the gene copies (subfunctionalitazion) (Conrad and Antonarakis 2007). Functional divergence occurs not only at a coding-sequence level but it is also induced by changes in regulatory elements of duplicated copies (Force et al. 1999) and even by alterations in gene splicing patterns (Su et al. 2006). In Drosophila, tandem duplication seems to be the most common mechanism generating multigenic families (Zhou et al. 2008). The rate at which fruitfly genes are gained and lost within a multigenic family is remarkably high (on average 17 genes arise from duplication events and 17 are lost per myr). This fact results in the rapid gain of species-specific genes, which may be implied in environmental adaptation. Finally, it has been postulated that gene duplication events followed by geographic isolation lead to hybrid incompatibility, and thus, duplications can contribute to speciation (Presgraves 2010).

Inversions

Inversions can also make a genome to gain new genes depending on the mechanism that generates the rearrangement. In Drosophila it has been shown that inversions caused by staggered single-strand break and repair by NHEJ (Figure 10) produce inverted duplications of DNA at the two breakpoints (Ranz et al. 2007). Only in *Helycobacter pilori* it has been demonstrated that new functional genes can be generated by this mechanism, also called duplication association to inversion (DDAI) (Furuta et al. 2011).

De novo gene origination

The recent availability of genome-wide data have revealed that *de novo* gene origination could be a common mechanism responsible for the great variation of genes in different lineages (Begun et al. 2007). By this process, originally noncoding DNA sequences become functional due to certain mutational events (Figure 13). In *D. melanogaster* 142 *cis*-regulated coding genes have been identified to come from ancestral nongenic sequences (Zhao et al. 2014). A total of 60 putative coding genes originated *de novo* seem to be present in the human genome since its divergence from the chimpanzee (Guerzoni and McLysaght 2011). These genes are suggested to be potential sources for the great phenotypic differences shown between humans and chimpanzees.

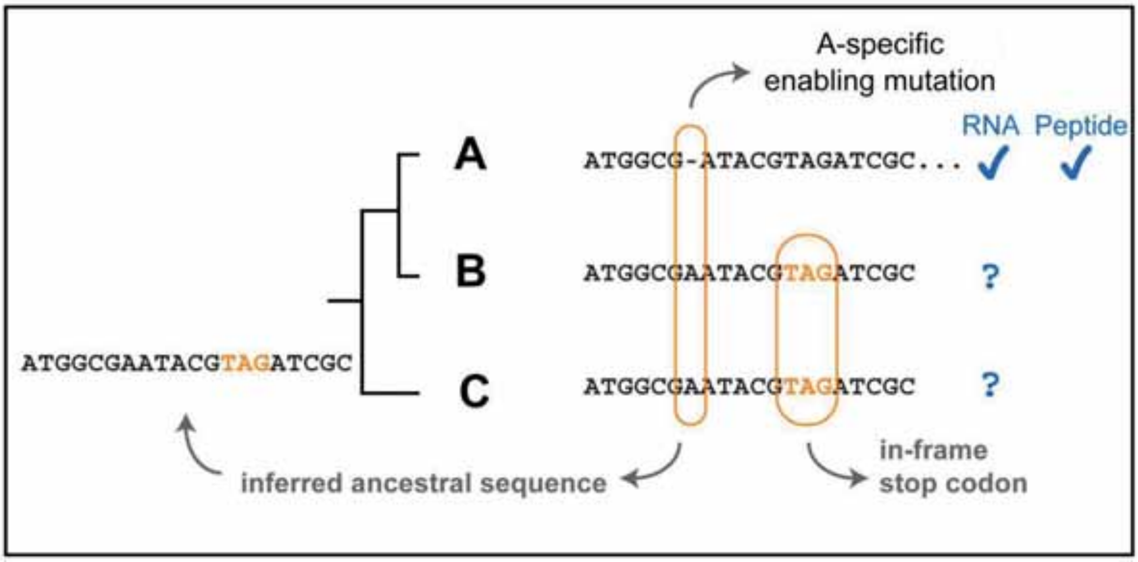


FIGURE 13. **Hypothetical example of a lineage-specific gene arised by de novo gene formation**. A single nucleotide deletion shifts a stop codon out of the new reading frame in species A. The comparison of the homologous sequences among sibling species (B and C) provides information about the ancestral sequence. The putative novel gene discovery can be confirmed with experimental evidences. Figure extracted from Guerzoni and McLysaght (2011).

Gene fusion and fission

The fusion of existing genes can also lead to new transcripts with a different function than that performed by the parental proteins, resulting in chimeric genes (Long 2000). However, many of the discovered gene fusion events in humans seem to be related to different diseases, mainly cancer (Mitelman et al. 2007). In Drosophila 14 chimeric functional genes have been recently identified (Rogers and Hartl 2012). The analysis of their sequence evolution as well as their expression pattern revealed that somehow they play an important role in adaptive evolution. On the other hand, by the gene fission process a single transcript can break into multiple transcripts carrying independent functions. For instance, the monkey-king gene (*mkg*) family, conserved in four related Drosophila species, is an example of a young gene family originated by gene fission (Wang et al. 2004).

Horizontal gene transfer

Organisms can transfer genes from each other (reciprocally or not) by horizontal (or lateral) gene transfer (HGT), i.e. genes are not sexually inherited from parents to progeny but they come from distantly related genomes (Roger 1999). Horizontal gene transfer is a common process between bacterial microorganisms, but only a few evidences have been reported for gene transfer movements between eukaryotic and prokaryotic genomes (Dunning Hotopp et al. 2007; Acuña et al. 2012). In addition eukaryote-eukaryote gene transfer has been also reported between fungi (Keeling and Palmer 2008) and it is though that the number of gene transfers between eukaryotes is underestimated as a consequence of the limitations associated to the methods used to detect HGT. Although nonsexual transmission of genetic material cannot be strictly considered a mechanism of gain of new genetic material, since the gene previously exist in other species, it has an important evolutionary impact (Keeling and Palmer 2008).

2.7 Regulatory changes in adaptive evolution

It has been clearly demonstrated that structural changes in genes, as well as the generation of new genetic material, have an important role in adaptive shifts in response to environmental changes (Hoffmann and Willi 2008). However, the enormous morphological and physiological diversity existing within organisms cannot be explained only by the contribution of these changes (Wilkins 1998). The structural and functional constrain of transcription factors (TFs), which are implicated in essential pathways controlling processes related to organisms' development, indicate that differences in gene expression likely impact on morphological diversification.

Hox genes are an essential set of transcription factors considered major regulators of animal development and it has been shown that both their sequence structure and genome colinearity are highly conserved among a wide range of species (McGinnis et al. 1990; McGinnis 1994; Kmita and Duboule 2003). This fact suggests that the accumulation of changes in hox gene expression pattern, rather than structural alterations in the coding sequence, greatly contributed to animal development diversification. Consequently, modifications in promoter regions or other regulatory elements controlling gene transcription, mainly cis-regulatory elements (CRE), considerably impact on adaptive evolution (Prud'homme et al. 2007). Hox gene complex' content and structure have been thoroughly studied in Drosophila (Negre et al. 2005; Negre and Ruiz 2007).

The study of the evolution of heat shock genes has also revealed the importance of mutations affecting regulatory patterns in key genes. Heat shock protein (*Hsp*) genes are involved in thermal responses. They encode intra-cellular chaperone proteins that help to protect other macromolecules from degradation, among other functions (Hoffmann et al. 2003). *Hsp* genes have been linked with adaptation to thermal environments across a wide range of organisms (Riehle et al. 2005; Fangue et al. 2006; Huang and Kang 2007). In Drosophila, differences in the expression of Hsp genes can be caused by the insertion of TEs in promoter regions of the genomes (Lerman and Feder 2005; Chen et al. 2007).

As a concluding remark, unlike other kinds of genetic alterations, regulatory changes are said to be more favored in the process of morphological evolution at a wide range of taxonomical levels since they are able to generate novelty by exploiting available genetic components.

3. OBJECTIVES

The recent availability of new sequencing technologies has made it possible to explore genome sequences and to assess the DNA changes directly involved in responding to environmental shifts. In this work we seek to identify genetic changes responsible for the peculiar ecology of two cactophilic species: *D. buzzatii* and *D. mojavensis*. To accomplish this objective we have focused on the adaptive value of two genomic features: chromosomal inversions and genes evolving under positive selection. Accordingly, this thesis is divided in two main objectives and eight specific objectives. In the first part we characterize all the inversions fixed in the chromosome 2 of *D. mojavensis*, the most dynamic of the five major chromosomes, and analyze their genomic distribution as well as their molecular causes and functional consequences. In the second part, the genomes of *D. mojavensis* and *D. buzzatii* are compared, allowing us for the analysis of the evolutionary patterns across genome sequences as well as the detection of genes under positive selection and other genomic features likely affecting niche specificity. A brief description of the proposed objectives is presented below.

Objective 1. To characterize the chromosomal inversions fixed in Drosophila mojavensis

1.1 To compare the organization of chromosomes between *D. buzzatii* and *D. mojavensis* to identify the number and extent of chromosomal inversions fixed during the divergence of the two species.

1.2 To map and characterize the breakpoints of the chromosomal inversions fixed in *D. mojavensis*.

1.3 To provide information on the molecular mechanisms that generated the inversions fixed in *D. mojavensis*.

1.4 To provide an explanation for the accelerated chromosomal evolution of the *D. mojavensis* lineage.

Objective 2. To compare the genome sequence of D. buzzatii and D. mojavensis in order to investigate the evolution of these cactophilic flies at the chromosome and gene levels.

2.1 To sequence, assemble and annotate the genome of *D. buzzatii*.

2.2. To study the developmental transcriptome of *D. buzzatii*

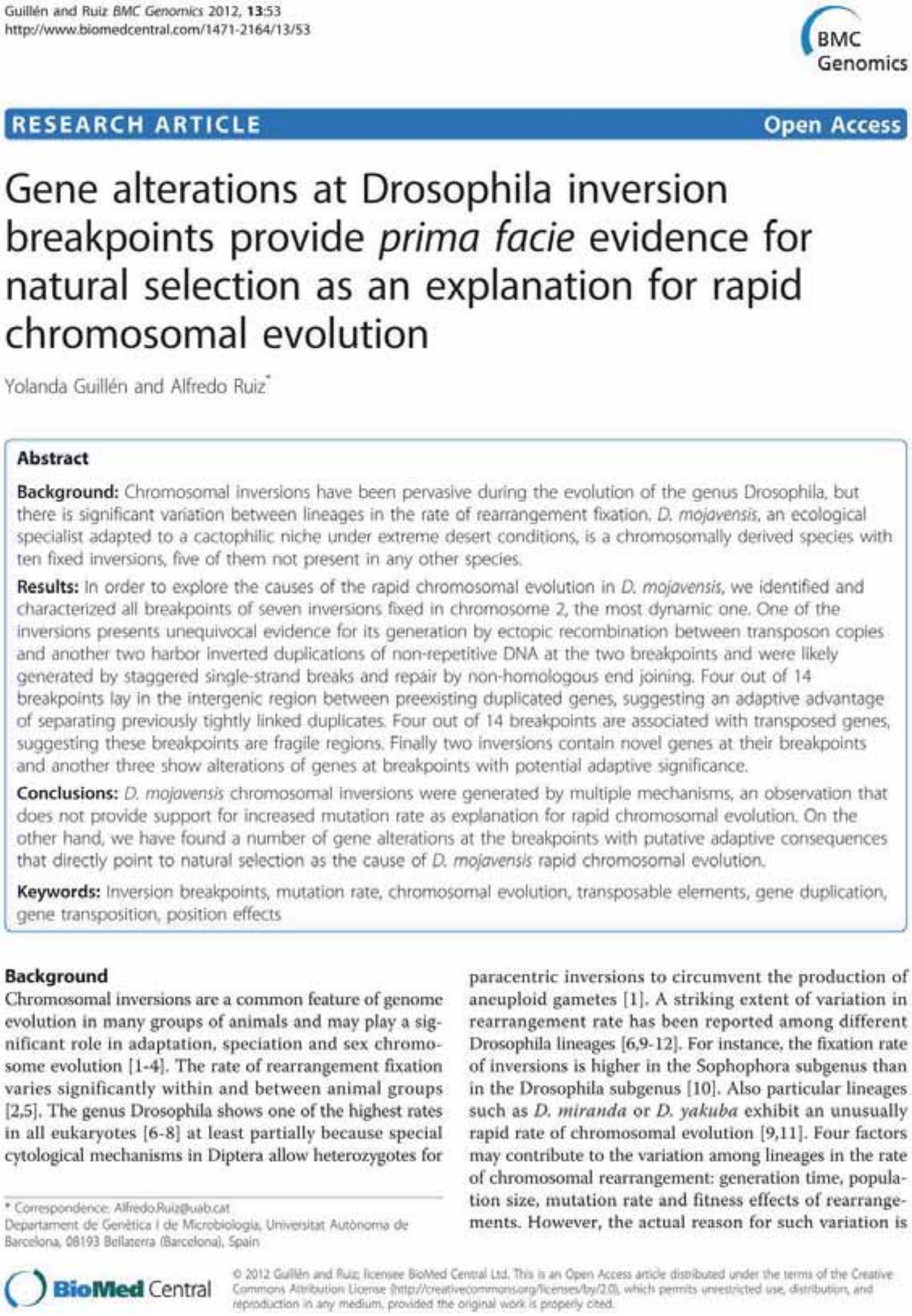
2.3 To compare single copy orthologs between *D. buzzatii* and *D. mojavensis* in order to characterize the patterns of molecular divergence.

2.4. To find genes under positive selection and lineage-exclusive genes in cactophilic flies that might presumably be involved in adaptation to ecological conditions.

4. RESULTS

4.1 Gene alterations at Drosophila inversion breakpoints provide prima facie evidence for natural selection as an explanation for rapid chromosomal evolution

YOLANDA GUILLÉN and ALFREDO RUIZ (2012) Gene alterations at Drosophila inversion breakpoints provide prima facie evidence for natural selection as an explanation for rapid chromosomal evolution. *BMC Genomics* **13**: 53.



Guillén and Ruiz *BMC Genomics* 2012, **13**:53
http://www.biomedcentral.com/1471-2164/13/53

BMC Genomics

RESEARCH ARTICLE **Open Access**

Gene alterations at Drosophila inversion breakpoints provide *prima facie* evidence for natural selection as an explanation for rapid chromosomal evolution

Yolanda Guillén and Alfredo Ruiz*

Abstract

Background: Chromosomal inversions have been pervasive during the evolution of the genus Drosophila, but there is significant variation between lineages in the rate of rearrangement fixation. *D. mojavensis*, an ecological specialist adapted to a cactophilic niche under extreme desert conditions, is a chromosomally derived species with ten fixed inversions, five of them not present in any other species.

Results: In order to explore the causes of the rapid chromosomal evolution in *D. mojavensis*, we identified and characterized all breakpoints of seven inversions fixed in chromosome 2, the most dynamic one. One of the inversions presents unequivocal evidence for its generation by ectopic recombination between transposon copies and another two harbor inverted duplications of non-repetitive DNA at the two breakpoints and were likely generated by staggered single-strand breaks and repair by non-homologous end joining. Four out of 14 breakpoints lay in the intergenic region between preexisting duplicated genes, suggesting an adaptive advantage of separating previously tightly linked duplicates. Four out of 14 breakpoints are associated with transposed genes, suggesting these breakpoints are fragile regions. Finally two inversions contain novel genes at their breakpoints and another three show alterations of genes at breakpoints with potential adaptive significance.

Conclusions: *D. mojavensis* chromosomal inversions were generated by multiple mechanisms, an observation that does not provide support for increased mutation rate as explanation for rapid chromosomal evolution. On the other hand, we have found a number of gene alterations at the breakpoints with putative adaptive consequences that directly point to natural selection as the cause of *D. mojavensis* rapid chromosomal evolution.

Keywords: Inversion breakpoints, mutation rate, chromosomal evolution, transposable elements, gene duplication, gene transposition, position effects

Background

Chromosomal inversions are a common feature of genome evolution in many groups of animals and may play a significant role in adaptation, speciation and sex chromosome evolution [1-4]. The rate of rearrangement fixation varies significantly within and between animal groups [2,5]. The genus Drosophila shows one of the highest rates in all eukaryotes [6-8] at least partially because special cytological mechanisms in Diptera allow heterozygotes for paracentric inversions to circumvent the production of aneuploid gametes [1]. A striking extent of variation in rearrangement rate has been reported among different Drosophila lineages [6,9-12]. For instance, the fixation rate of inversions is higher in the Sophophora subgenus than in the Drosophila subgenus [10]. Also particular lineages such as *D. miranda* or *D. yakuba* exhibit an unusually rapid rate of chromosomal evolution [9,11]. Four factors may contribute to the variation among lineages in the rate of chromosomal rearrangement: generation time, population size, mutation rate and fitness effects of rearrangements. However, the actual reason for such variation is

* Correspondence: Alfredo.Ruiz@uab.cat
Departament de Genètica i de Microbiologia, Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona), Spain

unclear and different studies invoke different explanations [9-12].

Chromosomal inversions can be generated by two major mechanisms. The first of them is ectopic recombination (or non-allelic homologous recombination, NAHR) between transposable elements (TEs) [13-15], segmental duplications [16,17] or short repeat sequences [18]. When ectopic recombination occurs between two copies of a TE inserted in opposite orientation at two different chromosomal sites, the resulting inverted chromosomal segment will be flanked by two chimeric TE copies bounded by exchanged target site duplications (TSD) [14,15]. The second mechanism is chromosomal breakage and erroneous repair of the free ends by non-homologous end-joining (NHEJ) [19]. Breakages can be simple double-strand breaks (DSB) or staggered single-strand breaks (SSB). In the second case, the consequence is the generation of inverted duplications at both sides of the inverted segment [11,20]. Thus, inversions generated in this way can be recognized by duplicated DNA segments (originally single-copy) in inverted orientation flanking the inverted chromosomal segment. The relative contribution of the two mechanisms to the generation of natural Drosophila inversions is not yet clear. In Dipterans, clear-cut evidence for the implication of TEs in their generation has been found for a few polymorphic inversions [15,21-23] but has never been found for fixed inversions [6,11,24,25]. On the other hand, breakage and repair by NHEJ may be the prevalent mechanism in *D. melanogaster* and its close relatives [11].

Several explanations have been put forward for the spread of inversions in populations [3]. Although in principle inversions could be neutral or underdominant and spread by genetic drift, this is probably unusual in Drosophila species given their elevated effective population size, of the order of 10^6 [26,27]. The traditional explanation for the adaptive significance of inversions is based on their recombination-reducing effect [28] that keeps together alleles at loci with epistatic effects on fitness, the "coadaptation" hypothesis [29]. An alternative model proposes that inversions capture a set of locally adapted genes and protect them from recombination with immigrant chromosomes [4,30]. Finally, inversions may spread in populations due to the direct mutational effects associated with their breakpoints, the "position effect" hypothesis [31]. This latter hypothesis has received so far little attention [32] but the relatively high gene density and compact structure of Drosophila genome (> 90% of euchromatin has functional annotations) [33,34] make position effects most likely. Available genomic sequences [35] provide the opportunity to investigate the structure of inversion breakpoints and ascertain their functional consequences.

Drosophila mojavensis has been an excellent model for the study of the genetics of ecological adaptation and speciation for more than fifty years [36-38] and it is now a useful model for genomic studies as the complete genome sequence is available [35,39]. *D. mojavensis* is a cactophilic species in the *repleta* group endemic to the deserts of the Southwestern USA and Northwestern Mexico, chiefly the Sonoran Desert (Arizona, Baja California and Sonora) the Mojave Desert and Santa Catalina Island in southern California. Natural populations are genetically differentiated and use different primary host plants, *Stenocereus gummosus* (pitaya agria) in Baja California, *Stenocereus thurberi* (organ pipe) in Arizona and Sonora, *Ferocactus cylindraceous* (California barrel) in Southern California and Opuntia spp. on Santa Catalina Island [40-42]. The ecological conditions of the Sonoran Desert are extreme (dry, arid and hot according to Köppen classification [43]) as attested by the fact that only four Drosophila species are endemic [41]. Accordingly, *D. mojavensis* is unusually thermotolerant and desiccation resistant [44-47]. In addition, *D. mojavensis* is the exclusive inhabitant of its chief host plants over most of its distribution range, in part because they contain large amounts of unusual lipids and triterpene glycosides that make them unsuitable for other Drosophila species [48,49].

The salivary gland chromosomes of *D. mojavensis* and its close relatives *D. arizonae* and *D. navojoa* were cytologically analyzed and the *D. mojavensis* standard chromosomal arrangement seemingly contain ten fixed inversions compared to Primitive I (the ancestor of the *repleta* group), one in chromosome X (*Xe*), seven in chromosome 2 (*2c, 2f, 2g, 2h, 2q, 2r* and *2s*) and two on chromosome 3 (*3a* and *3d*) [50,51]. Five inversions (*3d, Xe, 2q, 2r* and *2s*) are exclusive to *D. mojavensis* whereas the rest are shared by other cactophilic Drosophila of the *mulleri* complex (see Figure 1). Thus, *D. mojavensis* is a chromosomally derived species that contains the highest number of fixed inversions in the entire *mulleri* complex [52]. Only one of *D. mojavensis* inversions (Xe) has been previously characterized at the molecular level [53]. Here we characterize all inversions fixed in *D. mojavensis* chromosome 2, the most dynamic of the five major chromosomes, and explore the causes of its rapid chromosomal evolution. Using comparative mapping of BAC-end sequences from *D. buzzatii* onto the *D. mojavensis* genome (see Figure 1), we identify the breakpoint regions of all inversions. We then annotate them by comparison with the genome of *D. virilis*, the closest relative with a sequenced genome [35] that represents the ancestral (non-inverted) arrangement. Our results provide information on the multiple causes that generated these inversions, reveal unreported associations of inversion breakpoints with duplicated and transposed genes, and shed light on the functional consequences of *D. mojavensis* inversions. Overall, our results suggest that rapid

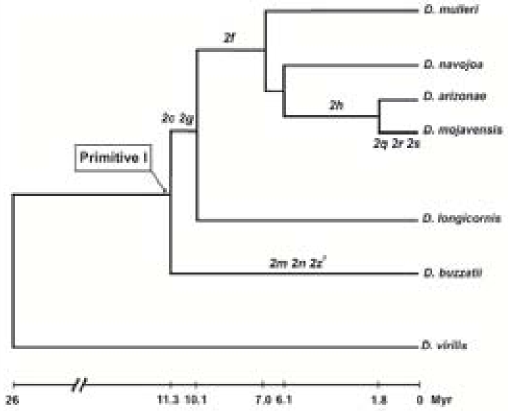


Figure 1 Phylogenetic relationships and divergence times for seven species of the Drosophila subgenus. Six species (*D. buzzatii, D. longicornis, D. mojavensis, D. arizonae, D. navojoa* and *D. mulleri*) belong to the *repleta* species group and the chromosome 2 inversions fixed in the *D. mojavensis* and *D. buzzatii* lineages are indicated [51,52]. Primitive I is the most recent common ancestor of *D. mojavensis* and *D. buzzatii* [52]. *D. virilis* is the outgroup species and belongs to the *virilis* species group. Phylogenetic relationships and divergence times are taken from [38, D.C.S.G. Oliveira, F.C. Almeida, P. O'Grady, W.J. Etges, M.A. Armella and R. DeSalle, personal communication].

chromosomal evolution in *D. mojavensis* is not due to an increase in the rate of inversion generation but to its adaptation to the extremely harsh environment of the Sonoran Desert that was accompanied by strong natural selection.

Results

Identification of syntenic segments and breakpoint regions

We sequenced the ends of 1,152 *D. buzzatii* chromosome 2 BAC clones [54] and 1,870 BAC-end sequences (BES) mapped onto *D. mojavensis* chromosome 2 (see Methods for details). By comparing the chromosomal localization of the markers, we identified 20 syntenic segments (Additional file 1). *D. mojavensis* scaffold 6540, corresponding to chromosome 2 [55], is 34,148,556 bp long (coordinates begin at centromere). The most proximal marker in our map (segment 20) was located at position 1,721,255 bp whereas the most distal marker (segment 1) was located at position 34,039,404, i.e. only 109 kb from the end of the scaffold. The largest segment was number 15 with 5,926.5 kb and 426 markers whereas the smallest one was number 16 with 50.5 kb and 9 markers. The second-smallest segment was number 7 with 80.7 kb and 2 markers. This latter segment was exceptional as it was detected using comparative information from BAC clone 1B03 that has been fully sequenced [56]. In general, the markers were distributed homogeneously along the chromosome as indicated by the highly significant correlation ($r^2 = 0.95$, $P < 0.001$) between segment size and number of markers. The 20 syntenic segments amount to 30,830,590 bp, representing ~90.3% coverage of chromosome 2. The missing 3,317,966 bp are distributed between the endmost chromosomal regions (~5.3%) and the 19 breakpoint regions (~4.4%).

Estimating the genomic distance

The order, size and orientation of the 20 conserved syntenic segments are shown in Figure 2. This breakpoint graph [57] contains nine cycles (represented with different colors), namely eight rectangles and a more complex cycle comprising two concatenated rectangles, suggesting that eight inversions and a more complex rearrangement are fixed in chromosome 2 since the divergence between *D. buzzatii* and *D. mojavensis*. GRIMM software [58] indicated that a minimum of 10 inversions are needed to transform the *D. buzzatii* chromosome 2 into that of *D. mojavensis* (Figure 3). Because there are 20 syntenic segments and 19 breakpoints, this implies one breakpoint reuse. Previous work in our laboratory [12] determined that three inversions, *2m*, *2n* and *2z*^{7}, have been fixed in chromosome 2 of *D. buzzatii* since its divergence from Primitive I, the most recent common ancestor with *D. mojavensis* (Figure 1). Furthermore, the breakpoints of these three inversions have been isolated and sequenced [59]. Inversions *2m* and *2n* are arranged in tandem and share the middle breakpoint. Thus we identified the complex cycle in the breakpoint graph (Figure 2) as corresponding to the *2mn* rearrangement and determined that seven inversions have been fixed in *D. mojavensis* since divergence from Primitive I. These seven inversions entail 14 breakpoints, i.e. they have independent breakpoints. GRIMM software [58] was run again to compare the arrangement of *D. mojavensis* chromosome 2 with that of Primitive I (inferred by subtraction of the three inversions fixed in *D. buzzatii*). The result was the single scenario shown in Figure 3.

In order to compare the inversions proposed by GRIMM with those detected previously using cytological methods (see Introduction), we located on the *D. buzzatii* physical map [54] those clones that mapped on each *D. mojavensis* breakpoint region and identified the chromosomal bands involved in each case. We corroborated the inversion breakpoints identified by cytogenetics and those detected by bioinformatics with an accuracy of three to five bands. Rearrangements detected cytologically and those proposed by GRIMM (Figure 3) did not only match in number but also the regions involved in each of them were in agreement, allowing for the differences between the precision of both techniques. However, three cytological breakpoint coincidences were not corroborated at the

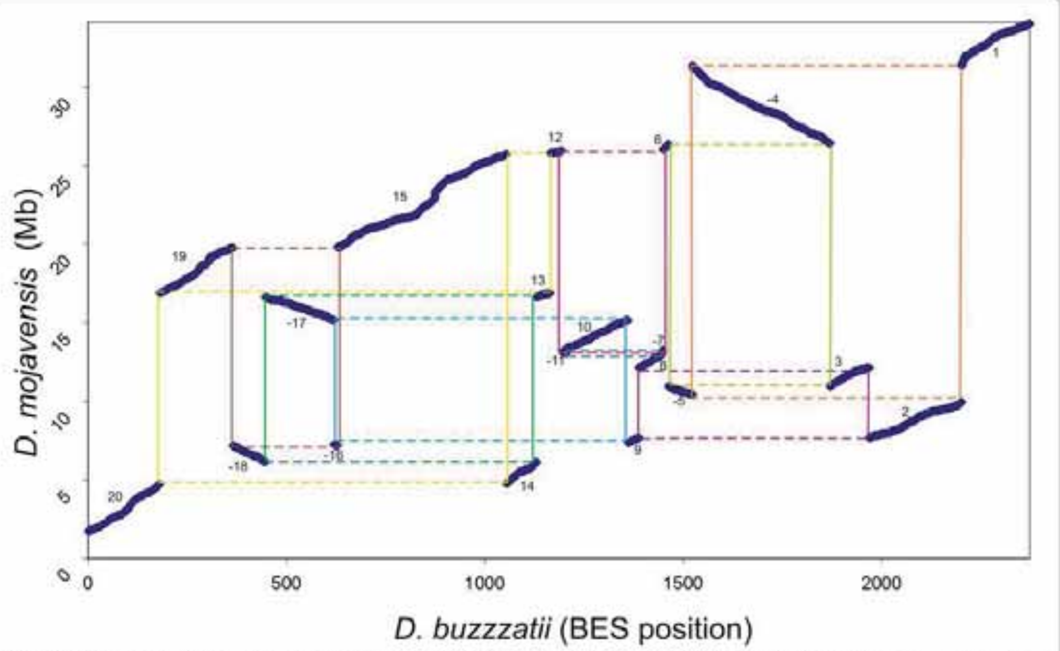


Figure 2 Dot plot comparing the order of chromosome 2 markers between *D. mojavensis* and *D. buzzatii*. The x-axis represents the position of all markers in the physical map of *D. buzzatii* [54], while in the y-axis markers were ordered according to their coordinates in *D. mojavensis* scaffold 6540. Each syntenic segment is identified by a number with inverted segments indicated by a negative sign. Also shown is the breakpoint graph [57] generated by connecting consecutive syntenic segments along *D. buzzatii* chromosome 2 with continuous lines and those along the *D. mojavensis* chromosome 2 with dashed lines. The result is eight rectangles and a more complex cycle comprising two concatenated rectangles (yellow) that correspond to the 10 inversions fixed between the two species.

sequence level. The general agreement between cytogenetics and bioinformatics is remarkable because often these two approaches to chromosomal evolution seem to provide discordant results [60,61]. For instance, in Drosophila, comparative mapping has sometimes revealed fixed inversions overlooked by previous cytological studies [11,62,63].

Delimitation and annotation of breakpoint regions

Among the seven chromosome 2 inversions fixed in the *D. mojavensis* lineage, three (*2f*, *2g* and *2c*) are shared between diverse species of the *mulleri* complex and must be between 7 and 11 myr old (Figure 1); another one (*2h*) is shared between *D. mojavensis* and *D. arizonae* only and should be between 2 and 6 myr old (Figure 1); the remaining three inversions (*2q*, *2r* and *2s*) are exclusive of *D. mojavensis* and thus must be relatively young (less than 2 myr, Figure 1). We initially identified the 14 breakpoint regions of these seven inversions as those sequences between syntenic segments (Additional file 2). These regions varied between 9,776 bp and 480,695 bp. In order to narrow down the size of these regions, the corresponding sequences were blasted against the *D. virilis* genome (see Methods), which represents the parental (non-inverted) chromosome (Figure 1). We expect that breakpoint regions for each inversion will appear in *D. mojavensis* genome as AC (distal) and BD (proximal) but in *D. virilis* genome as AB (distal) and CD (proximal). Similarity comparisons of AC, BD, AB and CD sequences allowed us to reduce the size of the breakpoint regions to between 259 bp and 91,812 bp, on average 8.3% of the original breakpoint regions (Additional file 2). Five breakpoint regions were further reduced to about 71.1% of their previous size (on average) by excluding the coding sequences of orthologous genes. Once the new limits for the 14 breakpoint regions in *D. mojavensis* were established, we analyzed the similarity between the two breakpoint regions of each inversion using BLAST 2 sequences [64]. A summary list of the genes adjacent to the 14 inversion breakpoints is shown in Table 1 and a detailed annotation of the breakpoint regions of each inversion is shown in Figures 4, 5 and 6 for the most recent inversions (*2s*, *2r* and *2q*) and Additional files 3, 4, 5 and 6 for the rest. TE content of all the breakpoint sequences (see

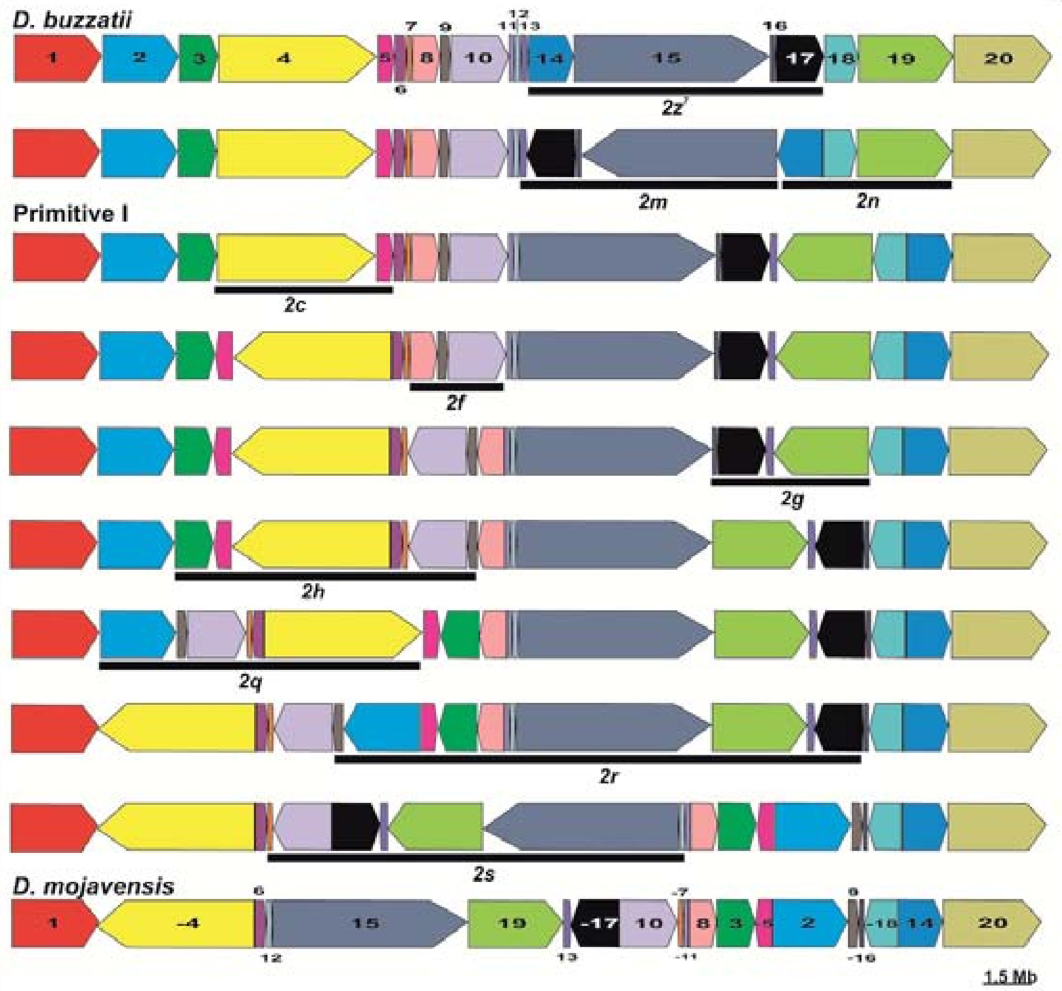


Figure 3 Order and orientation of the 20 chromosome 2 conserved segments during the divergence between *D. buzzatii* and *D. mojavensis*. The genomic distance between the two species calculated using GRIMM software was 10 inversions. When inversions *2m*, *2n* and $2z^7$ fixed in *D. buzzatii* [12,59] were subtracted, GRIMM software yielded a single scenario with seven inversions (*2c*, *2f*, *2g*, *2h*, *2q*, *2r* and *2s*) fixed from Primitive I to *D. mojavensis*.

Methods) is summarized in Additional file 7. Our analysis of the breakpoints provides significant information on the causes and consequences of the seven chromosome 2 inversions fixed in *D. mojavensis* (Table 1) that we present in the following sections.

Generation of chromosomal inversions

In order to test for the implication of TEs in the generation of the seven inversions, we analyzed the TE content of the breakpoint regions and detected copies of a TE at both co-occurrent breakpoints in three inversions: *2s*, *2r* and *2c* (Table 1). One of them, inversion *2s*, provides compelling evidence for the implication of the transposon *BuT5* [65] in its generation. At the distal breakpoint, a 981-bp copy of *BuT5* was found bounded by the 9-bp sequences AAGGCAAGT and CTGTATAAT (Figure 4). At the proximal breakpoint, we uncovered a 27-bp *BuT5* fragment comprising 12 bp identical to one end and 15 bp identical to the other end, and thus resembling the footprints that transposons often leave behind at the donor site following excision [66,67] bounded by the 9-bp sequences ACTTGCCTT and ATTATACAG. These

Table 1 Chief features of inversion breakpoint regions in *D. mojavensis*

Inv	Breakpoints and adjacent protein-coding genes in *D. mojavensis*		TE copies at co-occurrent breakpoints	Inversion-associated inverted duplications	Preexisting duplications in parental genome	Transposition-associated genes and *D. virilis* lineage specific genes (underlined)	Gene gains (bold) and putative position effects
2c	AC	*Ligatin-GstD1a*	*BuT5*		*GstD1a-GstD1b*		*GstD1aGstD1b*
	BD	*Sbp-GstD1b*					
2f	AC	*αTub84B-Pli*				*Lsp1β*	
	BD	*CG1091-Lsp1β*					
2g	AC	*Dmoj\GI22722-CG4511*				*Dmoj\GI22722 CG32344 CG2846 Dvir\GJ23779*	
	BD	*CG32344-spas*					
2h	AC	*pasha-ppk20*		7.1 kb from AB	*ppk20-ppk21*	*Dmoj\GI24456*	***Dmoj\GI23123***
	BD						
2q	AC	*Spargel-CG1213*		1 kb from AB 4.3 kb from CD	*CG1213-CG1208*		***Dmoj\GI22075***
	BD	*CG31528-CG1208*					
2r	AC	*Hsp68a-Hel89B*	*Galileo*		*Hsp68a-Hsp68b*	*Histone clusters*	*Hsp68a* *Hsp68b*
	BD	*Hsp68b-Cad99C*					
2s	AC	*CG9801-CG10214*		*BuT5*			*CG10375*
	BD	*CG34135-CG10375*					

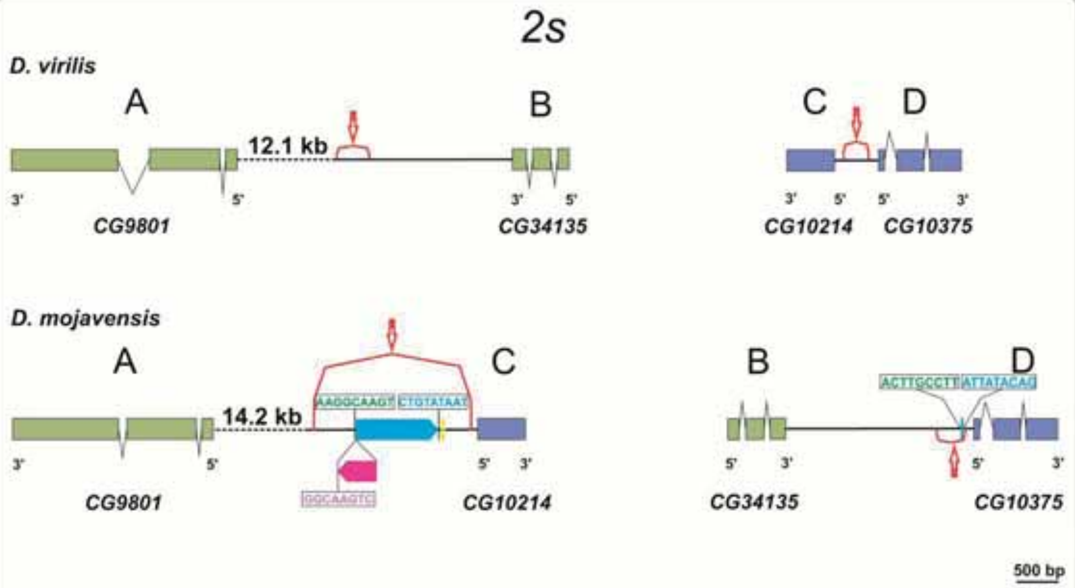


Figure 4 Annotation of inversion *2s* distal and proximal breakpoint regions in *D. virilis* (non-inverted chromosome) and *D. mojavensis* (inverted chromosome). Genes are depicted as solid boxes (exons) linked by polygonal lines (introns) with the 5' and 3' ends showing the direction of transcription. Genes adjacent to the distal (AB) and proximal (CD) breakpoints of *D. virilis* are colored in green and blue, respectively. Orthologs in the distal (AC) and proximal (BD) *D. mojavensis* breakpoint regions are colored accordingly. Red curly brackets with an arrow indicate the breakpoint junctions. TE insertions are shown as solid rectangles: blue (*BuT5*), purple (*Homo3*) or yellow (*Galileo*). Some TE insertions are flanked by TSDs insertions depicted in boxes above (or below) them. Dotted sections are not drawn to scale.

sequences can be interpreted as TSD produced at the time of the transposon insertion and its exchanged arrangement (ACTTGCCTT and CTGTATAAT are the inverted complementary versions of AAGGCAAGT and ATTATACAG, respectively) provides unequivocal evidence for the generation of inversion *2s* by ectopic recombination between *BuT5* copies. Recently work in our laboratory has shown that an inversion fixed in

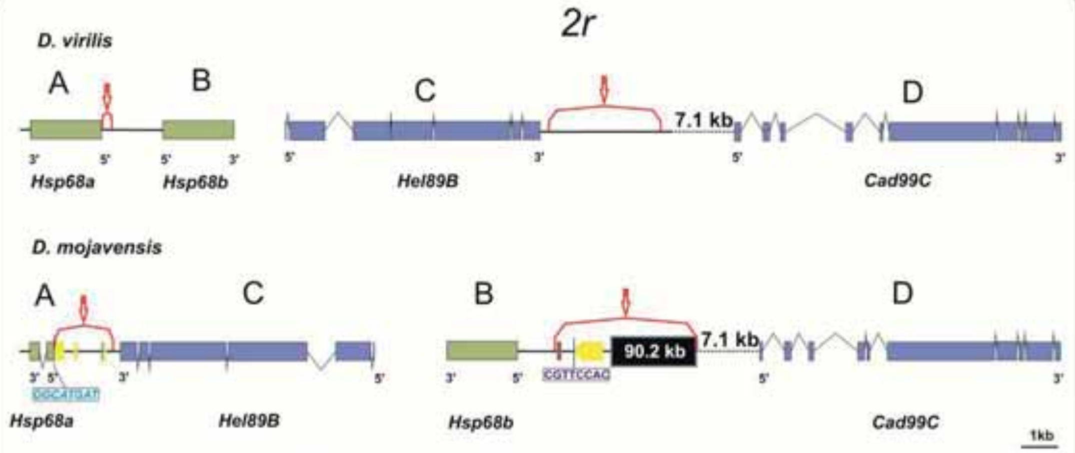


Figure 5 Annotation of inversion *2r* distal and proximal breakpoint regions *in D. virilis* (non-inverted chromosome) and *D. mojavensis* (inverted chromosome). TE insertions shown as solid rectangles: yellow (*Galileo*), green (*Invader*) or brown (*Homo6*). The black box in the *D. mojavensis* BD region represents a 90.2 kb-block containing interspersed histone clusters and TEs. Other symbols as in Figure 4.

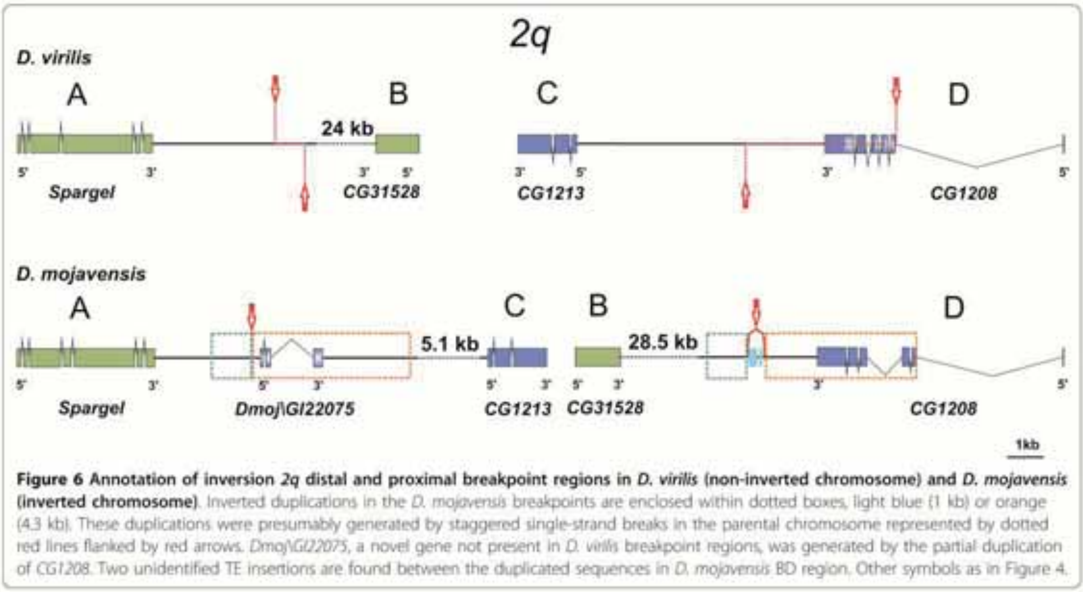


Figure 6 Annotation of inversion *2q* distal and proximal breakpoint regions in *D. virilis* (non-inverted chromosome) and *D. mojavensis* (inverted chromosome). Inverted duplications in the *D. mojavensis* breakpoints are enclosed within dotted boxes, light blue (1 kb) or orange (4.3 kb). These duplications were presumably generated by staggered single-strand breaks in the parental chromosome represented by dotted red lines flanked by red arrows. *Dmoj\GI22075*, a novel gene not present in *D. virilis* breakpoint regions, was generated by the partial duplication of *CG1208*. Two unidentified TE insertions are found between the duplicated sequences in *D. mojavensis* BD region. Other symbols as in Figure 4.

D. uniseta, another related species that belongs to the *buzzatii* complex [56], has also been generated by ectopic recombination between *BuT5* copies.

Two of our inversions provide evidence for generation by chromosomal breakage and erroneous repair by NHEJ [11,19,20] (Table 1). In the breakpoints of inversion *2h* we found evidence for a duplication of a 7.1-kb long segment containing three genes, *CG1792, Dmoj\GI23402* and *pasha* (Additional file 3). When the two *D. mojavensis* breakpoint regions AC and BD were compared, we uncovered 10 blocks of similarity (E-value ≤ 9e-10) 45 to 641-bp long. These blocks are scattered in a 7,151-bp segment containing genes *CG1792, Dmoj\GI23402* and *pasha* in the AC breakpoint, and within a 2,676-bp segment including *Dmoj\GI23123* in the BD breakpoint (Additional file 3). This duplication can be explained by staggered SSB at the distal breakpoint in the parental chromosome. In the distal breakpoint of the derived chromosome (AC) the segment seems intact and the three genes fully functional whereas in the proximal breakpoint (BD), this segment has been reduced to 2.7 kb by several deletions (Additional file 3). Because the duplication was caused by the inversion, we estimated the age of the *2h* inversion using the divergence of those fragments non-coding for proteins and the Drosophila neutral substitution rate of 0.0111 [68] as 4.4 myr, which is in agreement with the phylogenetic distribution of inversion *2h* (Figure 1). In the breakpoints of inversion *2q* there are also inverted duplications. In this case, staggered SSB likely occurred at both breakpoints involving a ~1-kb segment at distal breakpoint (AB) and a 4.3-kb segment at proximal breakpoint (CD) (Figure 6). We estimated the age of inversion *2q* using the same procedure as before as 1.4 myr, a figure fully compatible with the phylogeny (Figure 1). There is a third case, where duplicated genes (*GstD1*) in opposite orientation are found at the two inversion *2c* breakpoints (Additional file 6). However, this observation is best interpreted as a breakage occurring at a preexisting duplication (see below).

Preexisting gene duplications at breakpoints

We found four cases where inversion breakpoints fall between duplicated genes, i.e. there were preexisting gene duplications at the breakpoint regions (Table 1). In order to determine if this is in concordance with the random expectation, we estimated the number of intergenic regions localized between duplicated genes in *D. mojavensis* chromosome 2. Genes in this chromosome encode 3,407 out of the 14,595 *D. mojavensis* predicted proteins (23.34%). Thus there are 3,406 putative intergenic regions in this chromosome. According to the previous established criteria to consider two genes as duplicated copies (see Methods), we detected 215 intergenic regions between duplicated genes along the entire chromosome. We compared the two proportions by three different statistical methods. The χ^2 test with and without Yates correction (χ^2 = 11.526, P = 0.0007 and χ^2 = 8.111, P = 0.0044, respectively) indicated that 4/14 is significantly higher than the proportion expected at random (215/3406). Fisher exact test (P = 0.0098) corroborated this result. It could be argued that breakpoints are distributed at random in non-coding intergenic regions and that duplicated genes accumulate more breakpoints because their mean intergenic distance is longer than that between non-duplicated genes. We tested this possibility by calculating the intergenic distance for both duplicated and non-duplicated genes in *D. mojavensis* chromosome 2. A t-test showed that means are significantly different (t = 3.84, P = 0.0001), but the mean distance between duplicated genes is actually the shortest one. Thus, we conclude that there is an excess of breakpoints localized between duplicated genes with respect to the random expectation. Previously, another *D. mojavensis* inversion (*Xe*) was found to have a breakpoint adjacent to a gene duplication [53] and in primates, rearrangement breakpoints have been sometimes observed in the midst or adjacent to clustered gene families [69,70].

Two explanations can be put forward for these observations. Firstly, duplicated genes might cause instability and increased rate of DBS [71] or might be breakage "permissive" regions [70]. Alternatively, we suggest that the mobilization of a duplicated gene may entail in some cases beneficial position effects that might help the inversion to be fixed within the species. Two duplicated genes may have their evolution constrained because of shared regulatory sequences, their co-location in the same chromatin regulatory domain, or sequence homogenization by frequent conversion and ectopic recombination events. The re-location of one of the copies to a different chromosomal region might produce beneficial changes in the regulation of expression for one or the two copies and/or release them from evolutionary constraints (see below).

Association of inversion breakpoints with gene transposition events

Gene content of chromosomal elements is generally conserved in the genus Drosophila although gene order has been scrambled extensively by fixed paracentric inversions [6]. However, there have been a number of genes that have been relocated between or within chromosomal elements by gene transposition or retroposition [72-74]. We searched the 28 genes adjacent to *D. mojavensis* inversion breakpoints in the 12 Drosophila genomes [35] for evidences of gene transposition and found that there are three genes involved in interchromosomal transposition events (*Lsp1beta*, *Dmoj\GI22722* and *CG32344*) and another one (*Dmoj\GI24456*) is likely involved in a intrachromosomal transposition (Table 1). Furthermore, two genes are present in the *D. virilis* breakpoints regions but not in *D. mojavensis* and thus are likely also transposed genes. Finally, a large DNA block including several clusters of Histone genes have been inserted in the proximal breakpoint of inversion *2r* (Figure 5).

In order to test for an association of inversion breakpoints and gene transpositions, we first determined that 69 out of 514 interchromosomally relocated genes [73] are located in *D. mojavensis* chromosome 2. Then we compared our proportion of interchromosomal gene transpositions (3/28) to the general chromosome 2 proportion (69/3,407). The χ^2 test with and without Yates correction (χ^2 = 6.42, P = 0.0113; χ^2 = 10.22, P = 0.0014), and the Fisher exact test (P = 0.0199) indicated that 3/28 is significantly higher than the relation expected at random. This test is conservative as we did not take into account the putative intrachromosomal transposition of *Dmoj\GI24456*, the two *D. virilis* lineage specific genes or the 90-kb insertion at *2r* proximal breakpoint (see below). The association of inversion breakpoints and transposed genes is likely the result of the "fragility" or "permissivity" of these regions [8]. A clear example is the *2g* distal breakpoint region in *D. virilis* where three genes (*Dvir\GJ23449*, *Dvir\GJ23779* and *CG32344*) have transposed to this region from different sources.

The *2r* proximal breakpoint (BD) harbors a big block of DNA (~90-kb) not found in any of the *D. virilis* breakpoints. This block contains at least five tandemly arranged copies of the Histone gene cluster [75,76]. The exact number of copies cannot be determined due to the presence of a ~10 kb sequence gap bounded by histone genes from different clusters. The block also comprises a large number of fragments annotated as repetitive sequences (110 ReAS elements that amount up to ~45% of the sequence). These elements tend to occur at regular intervals with a periodicity similar to that of the Histone gene clusters. In *D. melanogaster* the Histone complex (*HIS-C*) is located in chromosomal arm 2L (Muller element B) and comprises ~100 tandemly arranged copies of a cluster containing five Histone genes (*His1*, *His2B*, *His2A*, *His4* and *His3*) [75,76]. Histone genes are often involved in transposition events. In the *repleta* group species, the ancestral and chief *HIS-C* (named *HIS-C1*) is likely located at chromosome 3, but there are other derived and probably smaller complexes (named *HIS-C2*) at chromosomes 3 and 4, implying at least two transposition events [62]. The insertion of a ~90-kb block containing several Histone gene clusters in the *2r* proximal breakpoint (BD) seems to represent yet another transposition event, which is probably specific to *D. mojavensis*. This block is not found in *D. virilis* in any of the two breakpoints and was not found in *D. mulleri* or *D. buzzatii* by *in situ* hybridization [62]. We suggest that the occurrence of this ~90-kb block is the result of the reintegration of an extrachromosomal circular DNA fragment (eccDNA) replicated by rolling circle replication [77] perhaps at the time of the inversion generation (when DSBs were available). This hypothesis explains the fact that this large insertion contains tandemly repeated coding (Histone) genes and TEs.

Gene gains and changes in gene structure and/or expression

Two novel genes have been generated at the *D. mojavensis* breakpoints (Table 1). The gene *Dmoj\GI23123*, localized at *2h* proximal breakpoint (BD), comprises two exons encoding a 94-aa protein (Additional file 3). A similarity search indicated that it is related to the gene *pasha* (partner of *drosha, CG1800*), that is found in the distal breakpoint (AC). *pasha* has five exons and encodes a 655-aa protein with a double-strand RNA binding domain that is involved in primary miRNA processing, among other biological processes. Amino acid identity between *Dmoj\GI23123* and *pasha* proteins is 93.5% over a 46-aa segment. Gene *Dmoj\GI23123* has an unknown molecular function but a protein domain *PTHR13482* involved in nucleic acid binding was detected with Interproscan [78]. In addition it is expressed according to modENCODE *D. mojavensis* DB http://www.modencode.org/, suggesting it is fully functional. This gene arose at the time of the inversion generation as a consequence of the duplication of a 7.1-kb segment originally containing three genes: *CG1792, Dvir\GJ23094* and *pasha* (Additional file 3). Seemingly the duplicated copies of *CG1792 and Dvir\GJ23094* were partially lost by deletion whereas the duplicated copy of *pasha* evolved into the novel gene *Dmoj\GI23123*.

Another novel gene, *Dmoj\GI22075*, is found at the distal breakpoint (AC) of inversion *2q* (Figure 6). It arose when this inversion was generated as a consequence of the duplication of a 4.3-kb segment containing a fragment of gene *CG1208*. This gene encodes a 508-aa protein that has glucose transmembrane transporter activity. *Dmoj\GI22075* comprises three exons encoding a 153-aa protein with a 75-aa Major Facilitator Superfamily (MFS) domain [79]. The conservation of this domain indicates that it is a new functional gene and suggests that it has retained a MFS function.

Three inversions entail putative changes in gene structure and/or expression. Two *GstD1* genes in opposite orientation were found at the two *D. mojavensis* breakpoints of inversion *2c* while only one is present in the proximal breakpoint (CD) of the *D. virilis* chromosome (Additional file 6). In order to ascertain the origin of these two genes, a phylogeny of *GstD* genes in *D. mojavensis* and *D. virilis* was built (Additional file 8). The two *Dmoj\GstD1* genes are co-orthologs of the *Dvir\GstD1* gene and we estimated the age of the duplication event that generated them (using divergence at synonymous sites) as 16 myr. Therefore, this duplication event took place before inversion *2c* and the inversion breakpoint occurred between two pre-existing duplicated *GstD1* genes. *GstD1* genes have been associated with the detoxification of insecticides as well as other chemical substances present at larval food sources [80]. Low et al. [81] detected that positive selection has operated on *GstD1* and identified the parallel evolution of a radical glycine to lysine amino acid change (K171) in *D. melanogaster, D. pseudoobscura* and *D. mojavensis*. Matzkin [82] found additional evidence for the adaptive evolution of *Dmoj\GstD1a*, a gene that shows changes of expression level in response to the use of different host plants as larval substrates [83]. Inversion *2c* relocated *GstD1a* to a new chromosomal region and left the other copy *GstD1b* in the original position. This might have triggered changes in their gene expression regulation and/or evolutionary constraints. The two *D. mojavensis GstD1* proteins differ by 14 aa including the critical 171 residue (where *GstD1a* has lysine but *GstD1b* has glutamic acid). In addition, according to *D. mojavensis* modENCODE DB the relocated *GstD1a* gene has seemingly a much higher expression level than the gene in the original location, *GstD1b*. We suggest that the *GstD1* duplication and subsequent separation of the two copies by inversion *2c* may have had significant consequences for the adaptation of the lineage of *D. mojavensis* and related species of the *mulleri* complex to its cactophilic niche (Figure 1).

The *2r* distal breakpoint was localized in *D. virilis* between two *Hsp68* genes oriented head-to-head (Figure 5). These two genes have the same structure and size (a single exon 1,935-bp long encoding a 644-aa protein) and nearly identical sequence (8 mismatches, 99.6% identity). However, in *D. mojavensis Hsp68a* (661 bp) is significantly shorter than *Hsp68b* (1,935 bp) and posses two exons encoding a 152-aa protein (Figure 5). The two genes only show conservation of a segment encoding 90-aa corresponding to a Heat Shock Protein domain (75% aa identity). We built a phylogeny of *Hsp68* in 11 Drosophila genomes (*D. willistoni* is the only of the 12 species lacking *Hsp68*, Additional file 9). While a single *Hsp68* gene is present in the six *melanogaster* group species, two copies oriented head-to-head are found in *D. pseudoobscura, D. persimilis, D. grimshawi* and *D. virilis*. Thus, this is likely to be the ancestral state. Nonetheless the phylogenetic tree shows a high similarity between the two *Hsp68* copies present within each of these four species (Additional file 9) that can be interpreted as the result of concerted evolution by recurrent gene conversion or ectopic recombination [84]. In *D. mojavensis*, inversion *2r* relocated *Hsp68b* to a new chromosomal site along with its upstream regulatory sequences. A detailed sequence analysis confirms that the *Dmoj\Hsp68b* 5' upstream region harbors two *cis*-regulatory motifs called HSEs (heat shock elements) modulating the expression of this gene [85]. But we also detect a third HSE, 683 bp upstream of the *Dmoj\Hsp68b* 5' region, in opposite orientation to the previous two HSEs. This putative *cis*-regulatory motif is likely to correspond to the HSE of *Dmoj\Hsp68a*, apparently dragged by the inversion to the BD region upstream of *Dmoj\Hsp68b*. In addition, only ~2.5 kb upstream of this gene is the ~90-kb block of

Histone genes and TEs (see above). Because TEs may influence chromatin organization [86] and this in turn is a significant determinant of gene expression [87,88], the insertion of this block is likely to have altered the expression level and/or pattern of *Dmoj\Hsp68b*. No promoter or regulatory HSE sequences were detected upstream of *Dmoj\Hsp68a* but according to *D. mojavensis* modENCODE DB this gene is being transcribed. It may be that it has recruited a new promoter (e.g. a fragment of the transposon *Galileo* located 3-bp from the initial codon; see Figure 5) and acquired a new function or it is on the way to becoming a pseudogene. It must be recalled that *Dmoj\Hsp68a* shows an altered structure and a high rate of sequence divergence (Additional file 8). In summary, we found that inversion *2r* has induced significant alterations of this gene in both structure and expression.

A footprint of a *BuT5* was found in the *D. mojavensis* proximal breakpoint of inversion *2s*, 121 bp from the start codon of *CG10375* (Figure 4). We used McPromoter http://tools.igsp.duke.edu/generegulation/McPromoter/[89] to look for the *Dmoj\CG10375* promoter. A unique putative promoter region was located 115-bp 5' from the start codon. This putative promoter region (~100-bp) includes the *BuT5* footprint and has a peak with high score (0.0505) located in region B (across the breakpoint). In addition it corresponds to a model 1 promoter (DNA replication related element). These observations contrast with the promoter region of *Dmel\CG10375* that is model 3 (Motif6/Motif1) and has a narrow peak with a lower score (0.03925) and imply that the *2s* inversion and the *BuT5* element have likely altered the expression of *Dmoj\CG10375*, presumably increasing it. Gene *CG10375* has a single *DnaJ* domain and is the likely orthologous of human *DNAJC8* gene, a member of the Hsp40 family.

Discussion

In this study, we investigated the rapid chromosomal evolution of the *D. mojavensis* lineage that has fixed ten paracentric inversions since the *repleta* group ancestor, ~12 mya (Figure 1). Using *D. buzzatii* BAC-end sequences [54] and the genome sequences of *D. mojavensis* and *D. virilis* [35] we mapped, identified, annotated and analyzed all breakpoints of the seven inversions fixed in *D. mojavensis* chromosome 2, the most dynamic element. The results corroborated previous cytological analyses [51] and allowed us to provide significant information on the causes and consequences of these structural changes.

One hypothesis that may explain an accelerated chromosomal evolution rate is an increased mutation rate that generates more rearrangements per generation. This possibility was invoked to explain the high rate of chromosomal rearrangement between *D. miranda* and *D. pseudoobscura* [9]. Because inversions may be generated by TEs (see Introduction), one possible cause of high mutation rate is an increased transpositional activity. Therefore, it has been suggested that variation in transpositional activity of TEs might contribute to variation in rates of rearrangement fixation [12]. However, an increased mutation rate could also be due to the presence of other causes, both intrinsic and extrinsic (e.g. clastogenic chemicals or ionizing radiation). Overall, our results do not support this hypothesis because the inversions fixed in *D. mojavensis* seem the result of multiple generation mechanisms. We found direct evidence for the implication of transposon *BuT5* in the generation of inversion *2s* and only circumstantial evidence for the implication of the transposons *BuT5* and *Galileo* in inversions *2c* and *2r*, respectively. Inversions *2h* and *2q* harbor inverted duplications of non-repetitive DNA at the two breakpoints and were likely generated by staggered single-strand breaks and repair by non-homologous end joining. Finally, no definitive conclusion can be drawn about the generation of inversions *2f* and *2g*. It could be argued that the latter inversions might have been generated by TEs but subsequent changes in the breakpoint regions hindered our ability to find conclusive evidence for their implication. TE copies might have excised and move to other locations after generating the inversion (a hypothesis known as "hit-and-run" [24]), or be deleted due to the high rate of loss of nonfunctional DNA in Drosophila [90,91]. However, in the absence of supporting evidence we think that such inference is unwarranted.

In any case, the generation of inversion *2s* by transposon *BuT5* is a significant finding because, in Dipterans, the implication of TEs in the generation of chromosomal inversions has been demonstrated for a few polymorphic rearrangements but never for fixed inversions (see Introduction). *BuT5* is a MITE with unusual features [N. Rius, A. Delprat and A. Ruiz, personal communication]. It was discovered in *D. buzzatii* [65] but is present in the genome of most *repleta* group species, implying that it was probably already present in the ancestor ~16 mya [N. Rius, A. Delprat and A. Ruiz, personal communication]. In *D. mojavensis* is relatively abundant and transpositionally active but copy density in the dynamic chromosome 2 is not significantly higher than in the rest of chromosomes. These observations do not support the increased mutation hypothesis.

A second explanation for accelerated chromosomal evolution is an increase of the species' population size because the rate of fixation of selectively advantageous rearrangements is a direct function of population size [26]. The high rate of chromosomal evolution of the *D. yakuba* lineage in comparison with the *D. melanogaster* lineage was attributed to differences in population size [11]. The effective population size of *D. mojavensis* has been estimated as ~10^6 yet there is variation between populations in Baja California and Mainland Sonora [92,93]. However, there is no reason to assume that this is an unusually high figure. Population size of *D. arizonae*, its closest relative (Figure 1), is seemingly higher (or at least not lower) than that of *D. mojavensis* [92,93]. In contrast to *D. mojavensis*, which is fixed for five species-specific inversions, *D. arizonae* has only one [51]. Therefore, population size does not provide an adequate explanation for *D. mojavensis* rapid chromosomal evolution.

The third hypothesis is strong natural selection in a new environment that increases the number of fixed inversions. *D. mojavensis* is the only *mulleri* complex species inhabiting the Sonoran Desert. Other species of this complex, including its closest relatives *D. arizonae* and *D. navojoa* (Figure 1), live in less harsh environments of central and southern Mexico. Thus it must be presumed that adaptation to the extreme conditions of the Sonoran desert and to the exclusive host plants exploited by *D. mojavensis* must have required many adaptive genetic changes. Chromosomal inversions in Drosophila have been considered for decades as adaptive devices that spread in natural populations driven by natural selection (see Introduction). In fact there is ample evidence for the adaptive significance of polymorphic inversions (those that are segregating within species) but no such evidence has been provided for fixed inversions (those that appear as interspecific differences). We have found a variety of gene alterations at the breakpoints of *D. mojavensis* chromosome inversions and propose that these alterations contributed to their adaptive value. Overall, strong natural selection in a new harsh environment seems the most plausible cause for *D. mojavensis* rapid chromosomal evolution.

The alterations associated with the breakpoints of five *D. mojavensis* inversions include two gene gains (*Dmoj\GI23123* and *Dmoj\GI22075*) and three putative alterations of gene structure and/or expression regulation (Table 1). We discuss these effects in turn. In *D. mojavensis* two new genes were generated associated to inversions *2q* and *2h*. As a consequence of the generation mechanism, staggered breakage and NHEJ repair, duplications of single-copy DNA were present at the breakpoints of these inversions at the onset. In the case of inversion *2q* this duplication included gene *CG1208* (except for its first exon and upstream sequences, see Figure 6). The novel gene *Dmoj\GI22075* is shorter than the original gene *CG1208* but retains a MFS domain and could function as a sugar transmembrane transporter (if a new promoter has been recruited). In the case of inversion *2h* the duplicated segment included originally three genes (see Additional file 3). Only one gene (*Dmoj\GI23123*) seems to have survived. This gene is related to *pasha* (a gene involved in primary microRNA processing and gene silencing by miRNA) and according to modENCODE data, it is expressed. We suggest that novel genes might have contributed to the adaptive value of these inversions. Novel genes are widely recognized as a source of new functions [94] but inversion-associated duplication has not been considered a molecular mechanism that can generate new genes until very recently and only in prokaryotes [20].

The two most recent inversions, *2r* and 2s, that are exclusive to *D. mojavensis*, show putative alterations of structure and/or expression of heat shock protein (Hsp) genes. Hsp genes encode intra-cellular chaperones for other proteins and have been established as potential candidates for thermotolerance [95]. Hsp family harbors genes constitutively or inducibley expressed [96]. Heat-inducible genes are regulated by heat shock factor (HSF), which binds to HSE sequences [97] whereas other heat shock genes have an Hsf-independent regulation [98]. The distal breakpoint of inversion *2r* separated two previously linked and very similar Hsp68 genes (Figure 5). One of them, *Hsp68a*, remained in its original location but suffered a radical change in structure and sequence. It may have acquired a new function and expression pattern or may be in the process of becoming a pseudogene. The other gene, *Hsp68b*, apparently kept its HSE regulatory elements but was relocated to a completely new chromatin environment and is now found near a ~90-kb block composed of Histone genes and TEs. It is difficult to imagine that the expression of this gene has not been affected by these changes. Genes of the heat- inducible Hsp70 family (to which *Hsp68* belongs) are positively related to thermotolerance but overexpression has survival costs and it seems that Hsp70 concentration has an intermediate optimum [44,99]. Some African populations of *D. melanogaster* with an exceptional thermotolerance show decreased levels of *Hsp70* expression, caused by the insertion of TEs in one of the promoter regions of the *Hsp70Ba* gene [100]. In *D. mojavensis* an altered expression of Hsp68 genes could contribute to its exceptional thermotolerance. On the other hand, the proximal breakpoint of inversion *2s* was located upstream of *CG10375*, a gene with a DnaJ domain that likely belongs to the constitutively expressed Hsp40 family. In *D. melanogaster, hsp40* is up-regulated in mutants lacking HSF [98] and probably has an essential role in thermotolerance [101]. Thus the changes induced by inversion 2s and *BuT5* insertion in the promoter of *CG10375* likely conferred an adaptive advantage to *D. mojavensis* by increasing its thermotolerance. It can be hypothesized that the alterations of the heat inducible *Hsp68* genes caused by inversion *2r* and the putative positive effect on the expression level of the constitutive gene *CG10375* caused by inversion *2s* were in some way related and jointly contributed to the *D. mojavensis* unusual thermotolerance. This hypothesis might explain the rapid and exclusive fixation of both inversions in the *D. mojavensis* lineage.

By no means do we imply that the alterations unveiled at the breakpoints are the only cause of the *D. mojavensis* inversion adaptive significance. Inversions are not simple point mutations but complex structural changes involving hundreds of loci that may suffer further mutations along their evolutionary trajectory. Therefore we consider that the multiple explanations for the adaptive spread of inversions (see Introduction) are not mutually exclusive alternatives. This means that different inversions may be successful for different reasons but also that a single inversion may increase in frequency for different reasons along its trajectory. For instance, an inversion could gain an initial drive because of the alterations it causes at the breakpoints and incorporate afterwards interacting mutations that led to coadaptation or that increase local adaptation that further propel the inversion towards fixation. The molecular explanations for the role of chromosomal inversions in adaptation and speciation are only beginning to be disentangled.

Conclusions

The breakpoint characterization of seven inversions fixed in *D. mojavensis* has provided significant information on the causes and consequences of these rearrangements. Multiple generation mechanisms seem to have acted in this lineage, an observation that does not support a mutational explanation for *D. mojavensis* rapid chromosomal evolution. On the other hand, we have found a set of alterations at the inversion breakpoints with potential adaptive significance, including novel genes and changes in structure and/or expression of adjacent genes. Overall, our results are consistent with natural selection as an explanation for the rapid chromosomal evolution in this specialist organism living under extreme ecological conditions.

Methods

In order to map and characterize the breakpoints of *D. mojavensis* chromosome 2 inversions we used a three-step approach: (1) End sequencing of a set of BAC clones from *D. buzzatii* chromosome 2; (2) Mapping of the resulting BAC-end sequences (BES) onto the *D. mojavensis* genome in order to determine the number and chromosomal span of the inversions fixed during the divergence of the two lineages; (3) Identification and annotation of the breakpoint regions using the *D. virilis* genome as representative of the parental (non-inverted) genome. Chromosome 2 of *D. mojavensis* differs by 42 chromosomal inversions from the homologous element in *D. virilis* [6]. The use of the *D. buzzatii* BES allowed us to identify and characterize those inversions fixed in the *D. mojavensis* lineage after its divergence from the *repleta* group ancestor (see Figure 1).

BAC end sequencing

We selected 1,152 clones from the *D. buzzatii* BAC library homogenously distributed along the 28 contigs of the chromosome 2 physical map [54]. To minimize redundancy we choose overlapping clones but with different restriction patterns. This was done using the information provided by the fingerprinting analysis of BAC clones that is available at http://www.bcgsc.ca/platform/bioinfo/software/ice. The 1,152 clones were rearrayed into 96 well plates (CHORI, Children's Hospital Oakland Research Institute) and both ends of each clone were sequenced (Macrogen Inc., Seoul, Korea) using the universal T7 primer and the modified universal SP6 primer (ATTTAGGTGACACTATAGAAGG) for PCR amplifications at the forward and reverse ends, respectively. We generated 2,127 reads over 400 bp in length, a success rate of 92.32%. Length distribution of BAC-end sequences (BES) for the two primers were similar with a pronounced mode around 700-800 bp (Additional file 10). If only high-quality BES (Q≥20) are taken into account, 80.82% of all sequences had over 400 bp in length. Our goal was to maximize the number of clones with both ends sequenced (paired BES) to increase coverage and the chances to capture all inversion breakpoints. Thus, a total of 1,004 of the original 1,152 BAC clones (87.2%) produced paired BES, whereas 119 clones (10.3%) produced a single BES. All BES were filtered with *Geneious*® software [102] using *VecScreen* database in order to identify and remove additional plasmidic sequences.

Mapping *D. buzzatii* BES onto the *D. mojavensis* genome

All *D. buzzatii* BES were tested for similarity to the *D. mojavensis* genome by BLASTN. This multiple search was carried out with the parameter set '-e' 1e-20, '-W' 7, '-r' 2, '-q' 3, '-G' 5 and '-E' 2 (e-value, word size, reward for a nucleotide match, penalty for a nucleotide mismatch, gap opening cost and gap extension cost, respectively). The values of the rest of the parameters were assigned by default. A masked CAF1 version of *D. mojavensis* genome, which is available at *FlyBase* website ftp://ftp.flybase.net/genomes/aaa/transposable_elements/ReAS/v1/CAF1_masked/, was used as reference for these blast searches. The use of a masked genome based on the ReAS library [103] allowed us avoiding results with multiples hits due to repetitive sequences, such as TEs or heterochromatic fragments. Only those hits localized at chromosome 2, which is uniquely represented by scaffold 6540 [55], were considered. Of these, we only took into account the hits that had a minimum length of 50 bp (10% of sequence mean length, approximately). The rest of the hits were discarded, including multiple hits for different scaffolds (except BLAST outputs composed by multiple hits in scaffold 6540 only). All validated hits, i.e., those that met the above criteria, were reordered based on the coordinates of *D. mojavensis* genome.

From the initial 2,304 BES, 1,933 (83.9%) matched any region of *D. mojavensis* genome while 1,870 (81.2%) mapped onto chromosome 2 resulting in 2,421 hits (Additional file 10). The number of hits exceeds the number of BES because some BES yielded more than one hit. In most cases the hits produced by a single BES were concatenated, i.e. mapped at adjacent sites in the *D. mojavensis* genome.

We included in our study a number of BES generated in previous works [56,59] reaching a total of 2,456 hits. Assuming that chromosome 2 is ~34 Mb long [55], we estimated an average density of one hit or marker every 13.8 kb. The distributions of hit size, e-value and percent identity are shown in (Additional file 10). Hit size was over 400 bp in 50% of all cases, and we did not obtain hits with a length lower than 50 bp due to filtering restrictions. The distribution of e-value was similar for BES from both primers, T7 and SP6, and shows a prominent peak (18.32% of all hits) at an e-value equal to 0 (Additional file 10). Finally, the distribution of percent identity between the *D. buzzatii* BES and the *D. mojavensis* genome sequences showed a bell-shaped distribution with an average value of 83.1% (Additional file 10).

A revised version of *D. buzzatii* physical map of chromosome 2

The published version of *D. buzzatii* physical map [54] comprises 28 contigs on chromosome 2. Another contig, 1031, has been anchored in chromosome 2 between contigs 1090 and 1181 in a recent mapping work [56]. Here, only four out of the 29 contigs, 1331, 987, 1330 and 1344, were not mapped to chromosome 2 and accordingly are likely to be misassembles or artifacts. We removed them from the revised version. The information provided by the comparative mapping of *D. buzzatii* BES onto the *D. mojavensis* genome allows us to assess the presence or absence of overlaps or gaps between contigs and estimate gap size. Supposing that there are no rearrangements or large indels involving contiguous sequences from adjacent contigs, we expect that contigs overlapping in *D. mojavensis* will also overlap in *D. buzzatii*, and vice versa. Based on this premise and assuming that *D. buzzatii* chromosome 2 has a similar size to that of *D. mojavensis*, we deduce that 15 of the putative gaps between contigs do not exist, i.e. we consider them closed gaps. In addition, we estimated the size of seven gaps between contigs of chromosome 2 as 20-240 kb. Finally, for the remaining two gaps, corresponding to breakpoint regions (see below), we estimated an upper bound for size. In summary, the new version of the map of *D. buzzatii* chromosome 2 comprises 10 contigs covering ~90% of chromosome 2 and contains 9 gaps that amount to ~5%. The remaining 5% correspond to the endmost (proximal and distal) regions that remain unmapped (see below).

Identification of syntenic segments

Each hit in the *D. mojavensis* genome was associated with its corresponding clone in the *D. buzzatii* physical map [54]. In this way, we could infer the number, arrangement and orientation of the conserved segments between *D. buzzatii* and *D. mojavensis*. With a single exception (Additional file 1), no syntenic segments were accepted with less than nine hits. Only 77 markers (3%) were not part of any syntenic segment. We guess that these markers represent common elements scattered throughout the genome, such as structural or functional domains or regulatory sequences, or represent gene transposition events. Some BES did not map to any *D. mojavensis* genome region. This might be caused by incompletely sequenced reads (those which were few bp long), regions with high sequence divergence between *D. mojavensis* and *D. buzzatii* or repetitive fragments. Finally, the centromere was not included in any syntenic segment owing to the lack of markers in this region (a masked *D. mojavensis* genome was used as reference).

Genomic distance

Once established the order and orientation of all syntenic segments in chromosome 2 between *D. buzzatii* and *D. mojavensis*, we estimated the genomic distance between the two species using GRIMM software [58]. The genomic distance is the minimum number of chromosomal rearrangements that differentiate two species [104]. The number of rearrangements estimated in this way was the sum of all inversions that had been fixed in the two lineages, *D. mojavensis* and *D. buzzatii*, since their divergence from Primitive I [12]. The three inversions, *2m*, *2n* and *2z*^{7}, fixed in the *D. buzzatii* lineage have been previously identified [12] and their breakpoints characterized at the molecular level [58]. This allowed us to subtract them from the total and infer the inversions fixed in the other lineage, i.e. from Primitive I to *D. mojavensis*.

Breakpoint analysis

We identified the breakpoint regions as the *D. mojavensis* genome sequences located between each pair of adjacent syntenic segments and estimated their size as the distance from the final marker in one syntenic segment to the initial marker in the next syntenic segment. The two breakpoints belonging to the same inversion were associated with the aid of GRIMM results. Once the two breakpoints of each inversion were identified, we proceeded to confirm these results by comparing the breakpoint regions sequences with *D. virilis* genome using FlyBase GBrowse http://flybase.org/cgi-bin/gbrowse/.

D. virilis is the phylogenetically closest species to *D. mojavensis* whose genome has been sequenced [35]. For this reason it was used as reference for the breakpoint comparative analysis. In order to narrow down the breakpoint regions, we blasted the breakpoint sequences against the *D. virilis* genome CAF1 masked version, also available at FlyBase website ftp://ftp.flybase.net/genomes/aaa/Transposable_ elements/ReAS/v1/CAF1_-masked/. A threshold e-value of 1e-3 was set to take into account the phylogenetic distance between *D. virilis* and *D. mojavensis* (Figure 1). All the BLASTN searches were performed with the parameters '-W' 7, '-r' 2, '-q' 3, '-G' 5 and '-E' 2. All hits for each breakpoint sequence were ordered according to *D. mojavensis* coordinates and the coordinates defined by the similarity loss between *D. mojavensis* and *D. virilis* were the new breakpoint limits. A final refinement of the breakpoint regions was carried out comparing the structures of those genes adjacent to the breakpoints in *D. mojavensis* with their respective orthologs in *D. virilis* (annotations extracted from FlyFase http://flybase.org) [105]. If the exon number and gene size were the same or very similar in the two orthologs, coding sequences still present in the *D. mojavensis* breakpoint regions were excluded from them, confining breakpoints to the intergenic space.

To detect orthologs in the *D. virilis* genome we downloaded from FlyBase [105] all the nucleotide sequences corresponding to the pair of genes adjacent to each *D. mojavensis* breakpoint, and then we used them as queries for BLASTN searches against *D. virilis* genome. We considered as ortholog that gene whose sequence in *D. virilis* was covered by the most significant hit of that search. To ensure the results, each BLAST search was repeated by exchanging the reference genome with *D. mojavensis* using as queries those *D. virilis* genes sequences putatively identified as orthologs in the first BLAST results (the *Reciprocal Best Hit* method, [106]). The function of genes adjacent to the breakpoints was inferred from the function of *D. melanogaster* orthologs in FlyBase and, for those genes without *D. melanogaster* orthologs, by searching for conserved domains using Interproscan [78] or NCBI Conserved Domain Database [107].

Search of transposable elements

We generated a database with all the *D. mojavensis* breakpoint sequences. Then we identified all the TEs present at the breakpoint regions by a set of BLASTN searches against DPDB [108], non redundant nucleotide database [109] and RepBase update [110]. We also used RepeatMasker [111] to detect repeats and TEs. Finally, we performed a set of BLAST searches against the breakpoint database using as queries a group of known TEs: *Galileo* from *D. mojavensis* (BK006357.1) [112], *Newton-1, Newton-2, Kepler-1 and Kepler-5* from *D. buzzatii* [113], and *BuT5* from *D. mojavensis* [N. Rius, A. Delprat and A. Ruiz, personal communication].

Detection of tandemly arranged duplicated genes

A number of the characterized inversion breakpoints were located between tandemly arranged duplicated genes in the parental (*D. virilis* genome). In order to test whether this number was expected under a random breakage model, we analyzed all the intergenic regions between duplicated genes in *D. mojavensis* chromosome 2. We first downloaded a database of predicted proteins for this species available at FlyBase website (version r1.3 of Feb 18, 2010). We extracted from this database all the proteins encoded by genes in scaffold 6540 (chromosome 2) and reordered them based on their gene position on this chromosome. Then, we carried out a search for pairs of similar proteins encoded by adjacent genes using BLAST 2 sequences (*bl2seq*) [64] with a cutoff *e-value* of 1e^{-30}. Based on the characteristics of duplicated genes found at breakpoint regions we considered that a pair of proteins was encoded by duplicated genes when the sequence identity between them was over 33% and at least one of the hits was longer than 57% of the shortest query length. Finally we counted the number of intergenic regions located between duplicated genes according to *bl2seq* results.

Additional material

Additional file 1: Size, coverage and coordinates of syntenic segments between *D. mojavensis* and *D. buzzatii* chromosome 2.

Additional file 2: Data for genome mapping of inversion breakpoint regions in the *D. mojavensis* genome.

Additional file 3: Annotation of inversion *2h* breakpoint regions. Annotation of inversion *2h* distal and proximal breakpoint regions in *D. virilis* (non-inverted chromosome) and *D. mojavensis* (inverted chromosome). Inverted duplications in the *D. mojavensis* breakpoints are enclosed within dotted boxes, orange color. That in region AC (7.1 kb) is intact whereas that in region BD (2.7 kb) has suffered several deletions. These duplications were presumably generated by staggered single-strand breaks in the parental chromosome represented by a dotted red lines flanked by red arrows. A fragment of *BuT3* is shown as a blue rectangle in region BD. Other symbols as in Figure 4.

Additional file 4: Annotation of inversion *2g* breakpoint regions. Annotation of inversion *2g* distal and proximal breakpoint regions in *D. virilis* (non-inverted chromosome) and *D. mojavensis* (inverted chromosome). Two *D. virilis* lineage specific genes are shown as grey rectangles. Other symbols as in Figure 4.

Additional file 5: Annotation of inversion *2f* breakpoint regions. Annotation of inversion *2f* distal and proximal breakpoint regions in *D. virilis* (non-inverted chromosome) and *D. mojavensis* (inverted chromosome). Symbols as in Figure 4.

Additional file 6: Annotation of inversion *2c* breakpoint regions. Annotation of inversion *2c* distal and proximal breakpoint regions in *D. virilis* (non-inverted chromosome) and *D. mojavensis* (inverted chromosome). Phylogenetic analysis of *GstD* genes (Additional file 8) indicates that the *2c* inversion occurred after the duplication of the GstD1 gene in the parental chromosome. The *GstD9* gene has lost its function in *D. mojavensis* becoming a pseudogene. Other symbols as in Figure 4.

Additional file 7: TE content of inversion breakpoint regions in *D. mojavensis*.

Additional file 8: Neighbor-Joining phylogenetic tree of GstD genes in *D mojavensis* and *D virilis*. Neighbor-Joining phylogenetic tree of GstD genes in D *mojavensis* and *D virilis*. Bootstrap values data for all tree nodes are shown. Phylogenetic analysis was conducted with MEGA4 [114]. Evolutionary distances were computed using the Maximum Composite Likelihood method.

Additional file 9: Neighbor-Joining phylogenetic tree of Hsp68 genes of 12 sequenced Drosophila species. Neighbor-Joining phylogenetic tree of Hsp68 genes of 12 sequenced Drosophila species. *D. persimilis, D. pseudoobscura, D. grimshawi, D. virilis* and *D. mojavensis* have two copies of the Hsp68 gene, while *D. sechellia, D. simulans, D. melanogaster, D. erecta, D. yakuba* and *D. ananassae* only one. No Hsp68 gene has been detected in *D. willistoni*. Bootstrap values for all tree nodes are shown. Phylogenetic analysis was carried out using MEGA4 [114]. Evolutionary distances were computed using the Maximum Composite Likelihood method.

Additional file 10: Statistics of *D. buzzatii* BAC end sequences. Description: Size distribution of *D. buzzatii* BAC end sequences (A) and distribution of size (B), E-value (C) and % identity (D) for hits generated blasting them against the *D. mojavensis* genome. See text for details.

Acknowledgements

We thank Esther Betrán, Andrew Clark, Alejandra Delprat, Bill Etges, Nuria Rius and two anonymous reviewers for useful comments on a previous version of the manuscript. This work was supported by a grant (BFU2008-04988) to AR and a FPI doctoral fellowship (BES-2009-021560) to YG from the Ministerio de Ciencia e Innovación, Spain.

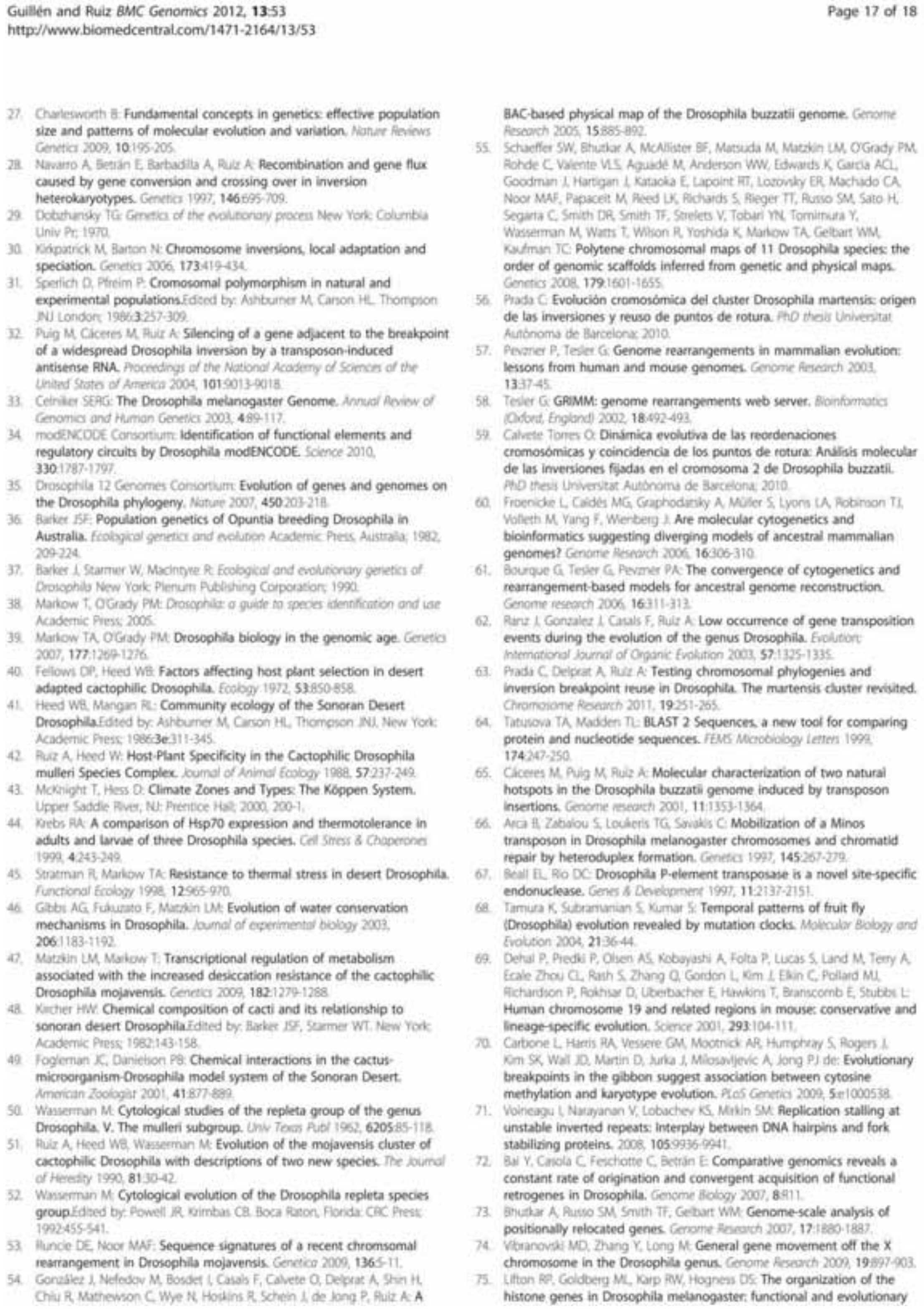
Authors' contributions

YG carried out the computational analysis. AR conceived and coordinated the study. YG and AR wrote the manuscript. All authors read and approved the final manuscript.

Received: 24 August 2011 Accepted: 1 February 2012
Published: 1 February 2012

References

1. White MJD: *Animal Cytology and Evolution*. 3 edition. London: Cambridge University Press; 1973.
2. Coghlan A, Eichler EE, Oliver SG, Paterson AH, Stein L: **Chromosome evolution in eukaryotes: a multi-kingdom perspective.** *TRENDS in Genetics* 2005, **21**:673-682.
3. Hoffmann A, Rieseberg LH: **Revisiting the Impact of Inversions in Evolution: From Population Genetic Markers to Drivers of Adaptive Shifts and Speciation?** *Annual review of ecology, evolution, and systematics* 2008, **39**:21-42.
4. Kirkpatrick M: **How and why chromosome inversions evolve.** *PLoS Biology* 2010, **8**:e1000501.
5. Murphy WJ, Larkin DM, Everts-van der Wind A, Bourque G, Tesler G, Auvil L, Beever JE, Chowdhary BP, Galibert F, Gatzke L, Hitte C, Meyers SN, Milan D, Ostrander EA, Pape G, Parker HG, Raudsepp T, Rogatcheva MB, Schook LB, Skow LC, Welge M, Womack JE, O'brien SJ, Pevzner PA, Lewin HA: **Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps.** *Science* 2005, **309**:613-617.
6. Bhutkar A, Schaeffer SW, Russo SM, Xu M, Smith TF, Gelbart WM: **Chromosomal rearrangement inferred from comparisons of 12 Drosophila genomes.** *Genetics* 2008, **179**:1657-1680.
7. Ranz JM, Casals F, Ruiz A: **How malleable is the eukaryotic genome? Extreme rate of chromosomal rearrangement in the genus Drosophila.** *Genome Research* 2001, **11**:230-239.
8. Grotthuss M von, Ashburner M, Ranz J: **Fragile regions and not functional constraints predominate in shaping gene organization in the genus Drosophila.** *Genome Research* 2010, **20**:1084-1096.
9. Bartolomé C, Charlesworth B: **Rates and patterns of chromosomal evolution in Drosophila pseudoobscura and D. miranda.** *Genetics* 2006, **173**:779-791.
10. Papaceit M, Aguadé M, Segarra C: **Chromosomal evolution of elements B and C in the Sophophora subgenus of Drosophila: evolutionary rate and polymorphism.** *Evolution; International Journal of Organic Evolution* 2006, **60**:768-781.
11. Ranz JM, Maurin D, Chan YS, Grotthuss M von, Hillier LW, Roote J, Ashburner M, Bergman CM: **Principles of genome evolution in the Drosophila melanogaster species group.** *PLoS Biology* 2007, **5**:e152.
12. González J, Casals F, Ruiz A: **Testing chromosomal phylogenies and inversion breakpoint reuse in Drosophila.** *Genetics* 2007, **175**:167-177.
13. Kupiec M, Petes TD: **Allelic and ectopic recombination between Ty elements in yeast.** *Genetics* 1988, **119**:549-559.
14. Lim JK, Simmons MJ: **Gross chromosome rearrangements mediated by transposable elements in Drosophila melanogaster.** *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* 1994, **16**:269-275.
15. Delprat A, Negre B, Puig M, Ruiz A: **The transposon Galileo generates natural chromosomal inversions in Drosophila by ectopic recombination.** *PloS One* 2009, **4**:e7883.
16. Coulibaly MB, Lobo NF, Fitzpatrick MC, Kern M, Grushko O, Thaner DV, Traoré SF, Collins FH, Besansky NJ: **Segmental duplication implicated in the genesis of inversion 2Rj of Anopheles gambiae.** *PloS One* 2007, **2**:e84910.
17. Cáceres M, Sullivan RT, Thomas JW: **A recurrent inversion on the eutherian X chromosome.** *Proceedings of the National Academy of Sciences of the United States of America* 2007, **104**:18571-18576.
18. Richards S, Liu Y, Bettencourt BR, Hradecky P, Letovsky S, Nielsen R, Thornton K, Hubisz MJ, Chen R, Meisel RP, Couronne O, Hua S, Smith MA, Zhang P, Liu J, Bussemaker HJ, Batenburg MF van, Howells SL, Scherer SE, Sodergren E, Matthews BB, Crosby MA, Schroeder AJ, Ortiz-Barrientos D, Rives CM, Metzker ML, Muzny DM, Scott G, Steffen D, Wheeler DA, Worley KC, Havlak P, Durbin KJ, Egan A, Gill R, Hume J, Morgan MB, Miner G, Hamilton C, Huang Y, Waldron L, Verduzco D, Clerc-Blankenburg KP, Dubchak I, Noor MA, Anderson W, White KP, Clark AG, Schaeffer SW, Gelbart W, Weinstock GM, Gibbs RA: **Comparative genome sequencing of Drosophila pseudoobscura: chromosomal, gene, and cis-element evolution.** *Genome Research* 2005, **15**:1-18.
19. Sonoda E, Hochegger H, Saberi A, Taniguchi Y, Takeda S: **Differential usage of non-homologous end-joining and homologous recombination in double strand break repair.** *DNA Repair* 2006, **5**:1021-1029.
20. Furuta Y, Kawai M, Yahara K, Takahashi N, Handa N, Tsuru T, Oshima K, Yoshida M, Azuma T, Hattori M, Uchiyama I, Kobayashi I: **Birth and death of genes linked to chromosomal inversion.** *Proceedings of the National Academy of Sciences of the United States of America* 2011, **108**:1501-1506.
21. Cáceres M, Ranz JM, Barbadilla A, Long M, Ruiz A: **Generation of a widespread Drosophila inversion by a transposable element.** *Science* 1999, **285**:415-418.
22. Casals F, Cáceres M, Ruiz A: **The foldback-like transposon Galileo is involved in the generation of two different natural chromosomal inversions of Drosophila buzzatii.** *Molecular biology and evolution* 2003, **20**:674-685.
23. Evans AL, Mena PA, McAllister BF: **Positive selection near an inversion breakpoint on the neo-X chromosome of Drosophila americana.** *Genetics* 2007, **177**:1303-1319.
24. Bergman CM, Pfeiffer BD, Rincón-Limas DE, Hoskins RA, Gnirke A, Mungall CJ, Wang AM, Kronmiller B, Pacleb J, Park S, Stapleton M, Wan K, George RA, de Jong PJ, Botas J, Rubin GMCS: **Assessing the impact of comparative genomic sequence data on the functional annotation of the Drosophila genome.** *Genome Biology* 2002, **3**:1-20.
25. Prazeres da Costa O, González J, Ruiz A: **Cloning and sequencing of the breakpoint regions of inversion 5g fixed in Drosophila buzzatii.** *Chromosoma* 2009, **118**:349-360.
26. Lande R: **The Expected Fixation Rate of Chromosomal Inversions.** *Evolution* 1984, **38**:743-752.



27. Charlesworth B: **Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation.** *Nature Reviews Genetics* 2009, **10**:195-205.
28. Navarro A, Betrán E, Barbadilla A, Ruiz A: **Recombination and gene flux caused by gene conversion and crossing over in inversion heterokaryotypes.** Genetics 1997, **146**:695-709.
29. Dobzhansky TG: *Genetics of the evolutionary process* New York: Columbia Univ Pr; 1970.
30. Kirkpatrick M, Barton N: **Chromosome inversions, local adaptation and speciation.** Genetics 2006, **173**:419-434.
31. Sperlich D, Pfreim P: **Cromosomal polymorphism in natural and experimental populations.**Edited by: Ashburner M, Carson HL, Thompson JNJ London; 1986:**3**:257-309.
32. Puig M, Cáceres M, Ruiz A: **Silencing of a gene adjacent to the breakpoint of a widespread Drosophila inversion by a transposon-induced antisense RNA.** *Proceedings of the National Academy of Sciences of the United States of America* 2004, **101**:9013-9018.
33. Celniker SERG: **The Drosophila melanogaster Genome.** *Annual Review of Genomics and Human Genetics* 2003, **4**:89-117.
34. modENCODE Consortium: **Identification of functional elements and regulatory circuits by Drosophila modENCODE.** Science 2010, **330**:1787-1797.
35. Drosophila 12 Genomes Consortium: **Evolution of genes and genomes on the Drosophila phylogeny.** Nature 2007, **450**:203-218.
36. Barker JSF: **Population genetics of Opuntia breeding Drosophila in Australia.** *Ecological genetics and evolution* Academic Press, Australia; 1982, 209-224.
37. Barker J, Starmer W, MacIntyre R: *Ecological and evolutionary genetics of Drosophila* New York: Plenum Publishing Corporation; 1990.
38. Markow T, O'Grady PM: *Drosophila: a guide to species identification and use* Academic Press; 2005.
39. Markow TA, O'Grady PM: **Drosophila biology in the genomic age.** Genetics 2007, **177**:1269-1276.
40. Fellows DP, Heed WB: **Factors affecting host plant selection in desert adapted cactophilic Drosophila.** *Ecology* 1972, **53**:850-858.
41. Heed WB, Mangan RL: **Community ecology of the Sonoran Desert Drosophila.**Edited by: Ashburner M, Carson HL, Thompson JNJ. New York: Academic Press; 1986:**3e**:311-345.
42. Ruiz A, Heed W: **Host-Plant Specificity in the Cactophilic Drosophila mulleri Species Complex.** *Journal of Animal Ecology* 1988, **57**:237-249.
43. McKnight T, Hess D: **Climate Zones and Types: The Köppen System.** Upper Saddle River, NJ: Prentice Hall; 2000, 200-1.
44. Krebs RA: **A comparison of Hsp70 expression and thermotolerance in adults and larvae of three Drosophila species.** *Cell Stress & Chaperones* 1999, **4**:243-249.
45. Stratman R, Markow TA: **Resistance to thermal stress in desert Drosophila.** *Functional Ecology* 1998, **12**:965-970.
46. Gibbs AG, Fukuzato F, Matzkin LM: **Evolution of water conservation mechanisms in Drosophila.** *Journal of experimental biology* 2003, **206**:1183-1192.
47. Matzkin LM, Markow T: **Transcriptional regulation of metabolism associated with the increased desiccation resistance of the cactophilic Drosophila mojavensis.** Genetics 2009, **182**:1279-1288.
48. Kircher HW: **Chemical composition of cacti and its relationship to sonoran desert Drosophila.**Edited by: Barker JSF, Starmer WT. New York: Academic Press; 1982:143-158.
49. Fogleman JC, Danielson PB: **Chemical interactions in the cactus-microorganism-Drosophila model system of the Sonoran Desert.** *American Zoologist* 2001, **41**:877-889.
50. Wasserman M: **Cytological studies of the repleta group of the genus Drosophila. V. The mulleri subgroup.** *Univ Texas Publ* 1962, **6205**:85-118.
51. Ruiz A, Heed WB, Wasserman M: **Evolution of the mojavensis cluster of cactophilic Drosophila with descriptions of two new species.** *The Journal of Heredity* 1990, **81**:30-42.
52. Wasserman M: **Cytological evolution of the Drosophila repleta species group.**Edited by: Powell JR, Krimbas CB. Boca Raton, Florida: CRC Press; 1992:455-541.
53. Runcie DE, Noor MAF: **Sequence signatures of a recent chromsomal rearrangement in Drosophila mojavensis.** *Genetica* 2009, **136**:5-11.
54. González J, Nefedov M, Bosdet I, Casals F, Calvete O, Delprat A, Shin H, Chiu R, Mathewson C, Wye N, Hoskins R, Schein J, de Jong P, Ruiz A: **A BAC-based physical map of the Drosophila buzzatii genome.** *Genome Research* 2005, **15**:885-892.
55. Schaeffer SW, Bhutkar A, McAllister BF, Matsuda M, Matzkin LM, O'Grady PM, Rohde C, Valente VLS, Aguadé M, Anderson WW, Edwards K, Garcia ACL, Goodman J, Hartigan J, Kataoka E, Lapoint RT, Lozovsky ER, Machado CA, Noor MAF, Papaceit M, Reed LK, Richards S, Rieger TT, Russo SM, Sato H, Segarra C, Smith DR, Smith TF, Strelets V, Tobari YN, Tomimura Y, Wasserman M, Watts T, Wilson R, Yoshida K, Markow TA, Gelbart WM, Kaufman TC: **Polytene chromosomal maps of 11 Drosophila species: the order of genomic scaffolds inferred from genetic and physical maps.** Genetics 2008, **179**:1601-1655.
56. Prada C: **Evolución cromosómica del cluster Drosophila martensis: origen de las inversiones y reuso de puntos de rotura.** *PhD thesis* Universitat Autònoma de Barcelona; 2010.
57. Pevzner P, Tesler G: **Genome rearrangements in mammalian evolution: lessons from human and mouse genomes.** *Genome Research* 2003, **13**:37-45.
58. Tesler G: **GRIMM: genome rearrangements web server.** *Bioinformatics (Oxford, England)* 2002, **18**:492-493.
59. Calvete Torres O: **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.** *PhD thesis* Universitat Autònoma de Barcelona; 2010.
60. Froenicke L, Caldés MG, Graphodatsky A, Müller S, Lyons LA, Robinson TJ, Volleth M, Yang F, Wienberg J: **Are molecular cytogenetics and bioinformatics suggesting diverging models of ancestral mammalian genomes?** *Genome Research* 2006, **16**:306-310.
61. Bourque G, Tesler G, Pevzner PA: **The convergence of cytogenetics and rearrangement-based models for ancestral genome reconstruction.** *Genome research* 2006, **16**:311-313.
62. Ranz J, Gonzalez J, Casals F, Ruiz A: **Low occurrence of gene transposition events during the evolution of the genus Drosophila.** *Evolution; International Journal of Organic Evolution* 2003, **57**:1325-1335.
63. Prada C, Delprat A, Ruiz A: **Testing chromosomal phylogenies and inversion breakpoint reuse in Drosophila. The martensis cluster revisited.** *Chromosome Research* 2011, **19**:251-265.
64. Tatusova TA, Madden TL: **BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences.** *FEMS Microbiology Letters* 1999, **174**:247-250.
65. Cáceres M, Puig M, Ruiz A: **Molecular characterization of two natural hotspots in the Drosophila buzzatii genome induced by transposon insertions.** *Genome research* 2001, **11**:1353-1364.
66. Arca B, Zabalou S, Loukeris TG, Savakis C: **Mobilization of a Minos transposon in Drosophila melanogaster chromosomes and chromatid repair by heteroduplex formation.** Genetics 1997, **145**:267-279.
67. Beall EL, Rio DC: **Drosophila P-element transposase is a novel site-specific endonuclease.** *Genes & Development* 1997, **11**:2137-2151.
68. Tamura K, Subramanian S, Kumar S: **Temporal patterns of fruit fly (Drosophila) evolution revealed by mutation clocks.** *Molecular Biology and Evolution* 2004, **21**:36-44.
69. Dehal P, Predki P, Olsen AS, Kobayashi A, Folta P, Lucas S, Land M, Terry A, Ecale Zhou CL, Rash S, Zhang Q, Gordon L, Kim J, Elkin C, Pollard MJ, Richardson P, Rokhsar D, Uberbacher E, Hawkins T, Branscomb E, Stubbs L: **Human chromosome 19 and related regions in mouse: conservative and lineage-specific evolution.** Science 2001, **293**:104-111.
70. Carbone L, Harris RA, Vessere GM, Mootnick AR, Humphray S, Rogers J, Kim SK, Wall JD, Martin D, Jurka J, Milosavljevic A, Jong PJ de: **Evolutionary breakpoints in the gibbon suggest association between cytosine methylation and karyotype evolution.** *PLoS Genetics* 2009, **5**:e1000538.
71. Voineagu I, Narayanan V, Lobachev KS, Mirkin SM: **Replication stalling at unstable inverted repeats: Interplay between DNA hairpins and fork stabilizing proteins.** 2008, **105**:9936-9941.
72. Bai Y, Casola C, Feschotte C, Betrán E: **Comparative genomics reveals a constant rate of origination and convergent acquisition of functional retrogenes in Drosophila.** *Genome Biology* 2007, **8**:R11.
73. Bhutkar A, Russo SM, Smith TF, Gelbart WM: **Genome-scale analysis of positionally relocated genes.** *Genome Research* 2007, **17**:1880-1887.
74. Vibranovski MD, Zhang Y, Long M: **General gene movement off the X chromosome in the Drosophila genus.** *Genome Research* 2009, **19**:897-903.
75. Lifton RP, Goldberg ML, Karp RW, Hogness DS: **The organization of the histone genes in Drosophila melanogaster: functional and evolutionary**

implications. *Cold Spring Harbor Symposia on Quantitative Biology* 1978, , **42 Pt 2:** 1047-1051.

76. Kremer H, Hennig W: **Isolation and characterization of a Drosophila hydei histone DNA repeat unit.** *Nucleic Acids Research* 1990, **18**:1573-1580.
77. Cohen S, Agmon N, Yacobi K, Mislovati M, Segal D: **Evidence for rolling circle replication of tandem genes in Drosophila.** *Nucleic acids research* 2005, **33**:4519-4526.
78. Zdobnov EM, Apweiler R: **InterProScan–an integration platform for the signature-recognition methods in InterPro.** *Bioinformatics (Oxford, England)* 2001, **17**:847-848.
79. Marger MD, Saier MH: **A major superfamily of transmembrane facilitators that catalyse uniport, symport and antiport.** *Trends in Biochemical Sciences* 1993, **18**:13-20.
80. Salinas AE, Wong MG: **Glutathione S-transferases–a review.** *Current Medicinal Chemistry* 1999, **6**:279-309.
81. Low WY, Ng HL, Morton CJ, Parker MW, Batterham P, Robin C: **Molecular evolution of glutathione S-transferases in the genus Drosophila.** *Genetics* 2007, **177**:1363-1375.
82. Matzkin LM: **The molecular basis of host adaptation in cactophilic Drosophila: molecular evolution of a glutathione S-transferase gene (GstD1) in Drosophila mojavensis.** *Genetics* 2008, **178**:1073-1083.
83. Matzkin LM, Watts TD, Bitler BG, Machado CA, Markow T: **Functional genomics of cactus host shifts in Drosophila mojavensis.** *Molecular Ecology* 2006, **15**:4635-4643.
84. Bettencourt BR, Feder ME: **Rapid concerted evolution via gene conversion at the Drosophila hsp70 genes.** *Journal of Molecular Evolution* 2002, **54**:569-586.
85. Tian S, Haney RA, Feder ME: **Phylogeny Disambiguates the Evolution of Heat-Shock cis-Regulatory Elements in Drosophila.** *PLoS One* 2010, **5**: e10669.
86. Zhang Z, Pugh BF: **Genomic Organization of H2Av Containing Nucleosomes in Drosophila Heterochromatin.** *PloS One* 2011, **6**:e20511.
87. Bell O, Tiwari VK, Thomä NH, Schübeler D: **Determinants and dynamics of genome accessibility.** *Nature Reviews Genetics* 2011, **12**:554-564.
88. Li G, Reinberg D: **Chromatin higher-order structures and gene regulation.** *Current Opinion in Genetics & Development* 2011, **21**:175-186.
89. Ohler U: **Identification of core promoter modules in Drosophila and their application in accurate transcription start site prediction.** *Nucleic Acids Research* 2006, **34**:5943-5950.
90. Petrov DA, Lozovskaya ER, Hartl DL: **High intrinsic rate of DNA loss in Drosophila.** *Nature* 1996, **384**:346-349.
91. Petrov DA, Hartl DL: **High rate of DNA loss in the Drosophila melanogaster and Drosophila virilis species groups.** *Molecular Biology and Evolution* 1998, **15**:293-302.
92. Matzkin LM, Eanes WF: **Sequence variation of alcohol dehydrogenase (Adh) paralogs in cactophilic Drosophila.** *Genetics* 2003, **163**:181-194.
93. Matzkin LM: **Population genetics and geographic variation of alcohol dehydrogenase (Adh) paralogs and glucose-6-phosphate dehydrogenase (G6pd) in Drosophila mojavensis.** *Molecular Biology and Evolution* 2004, **21**:276-285.
94. Kaessmann H: **Origins, evolution, and phenotypic impact of new genes.** *Genome Research* 2010, **20**:1313-1326.
95. Hoffmann A, Sørensen JG, Loeschcke V: **Adaptation of Drosophila to temperature extremes: bringing together quantitative and molecular approaches.** *Journal of Thermal Biology* 2003, **28**:175-216.
96. McColl G, Hoffmann A, McKechnie SW: **Response of two heat shock genes to selection for knockdown heat resistance in Drosophila melanogaster.** *Genetics* 1996, **143**:1615-1627.
97. Parker CS, Topol J: **A Drosophila RNA polymerase II transcription factor binds to the regulatory site of an hsp 70 gene.** *Cell* 1984, **37**:273-283.
98. Neal SJ, Karunanithi S, Best A, So AK-C, Tanguay RM, Atwood HL, Westwood JT: **Thermoprotection of synaptic transmission in a Drosophila heat shock factor mutant is accompanied by increased expression of Hsp83 and DnaJ-1.** *Physiological Genomics* 2006, **25**:493-501.
99. Krebs RA, Feder ME: **Hsp70 and larval thermotolerance in Drosophila melanogaster: how much is enough and when is more too much?** *Journal of Insect Physiology* 1998, **44**:1091-1101.
100. Lerman DN, Feder ME: **Naturally occurring transposable elements disrupt hsp70 promoter function in Drosophila melanogaster.** *Molecular Biology and Evolution* 2005, **22**:776-783.
101. Carmel J, Rashkovetsky E, Nevo E, Korol A: **Differential Expression of Small Heat Shock Protein Genes in Fruit Flies (Drosophila melanogaster) along a Microclimatic Gradient.** *Journal of Heredity* 2011, 10.1093/jhered/esr027.
102. Drummond A, Ashton B, Cheung M, Cooper A, Heled J, Kearse M, Moir R, Stones-Havas S, Sturrock S, Thierer TWA: **Geneious v5.1.** 2010, Available from http://www.geneious.com.
103. Li , Ye J, Li S, Wang J, Han Y, Ye C, Wang J, Yang H, Yu J, Wong GK-S, Wang J: **ReAS: Recovery of ancestral sequences for transposable elements from the unassembled reads of a whole genome shotgun.** *PLoS Computational Biology* 2005, **1**:e43.
104. Hannenhalli S, Pevzner P: **Transforming men into mice (polynomial algorithm for genomic distance problem).** *Proceedings of the 36th Annual IEEE Symposium on Foundations of Computer Science* 1995, 581-592.
105. Drysdale R: **FlyBase: a database for the Drosophila research community.** *Methods in Molecular Biology* 2008, **420**:45-59.
106. Moreno-Hagelsieb G, Latimer K: **Choosing BLAST options for better detection of orthologs as reciprocal best hits.** *Bioinformatics* 2008, **24**:319-324.
107. Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Jackson JD, Ke Z, Lanczycki CJ, Lu F, Marchler GH, Mullokandov M, Omelchenko MV, Robertson CL, Song JS, Thanki N, Yamashita RA, Zhang D, Zhang N, Zheng C, Bryant SH: **CDD: a Conserved Domain Database for the functional annotation of proteins.** *Nucleic Acids Research* 2011, **39**: D225-229.
108. Casillas S, Egea R, Petit N, Bergman CM, Barbadilla A: **Drosophila polymorphism database (DPDB): a portal for nucleotide polymorphism in Drosophila.** *Fly* 2007, **1**:205-211.
109. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW: **GenBank.** *Nucleic Acids Research* 2011, **39**:D32-37.
110. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J: **Repbase Update, a database of eukaryotic repetitive elements.** *Cytogenetic and Genome Research* 2005, **110**:462-467.
111. Chen N: **Using RepeatMasker to identify repetitive elements in genomic sequences.** *Current Protocols in Bioinformatics* 2004, **Chapter 4, Unit 4.10**.
112. Marzo M, Puig M, Ruiz A: **The Foldback-like element Galileo belongs to the P superfamily of DNA transposons and is widespread within the Drosophila genus.** *Proceedings of the National Academy of Sciences of the United States of America* 2008, **105**:2957-2962.
113. Casals F, Cáceres M, Manfrin MH, González J, Ruiz A: **Molecular characterization and chromosomal distribution of Galileo, Kepler and Newton, three foldback transposable elements of the Drosophila buzzatii species complex.** *Genetics* 2005, **169**:2047-2059.
114. Tamura K, Dudley J, Nei M, Kumar S: **MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0.** *Molecular Biology and Evolution* 2007, **24**:1596-1599.

doi:10.1186/1471-2164-13-53
Cite this article as: Guillén and Ruiz: Gene alterations at Drosophila inversion breakpoints provide *prima facie* evidence for natural selection as an explanation for rapid chromosomal evolution. *BMC Genomics* 2012 13:53.

Additional file 1. Size, coverage and coordinates of syntenic segments between *D. mojavensis* and *D. buzzatii* chromosome 2.

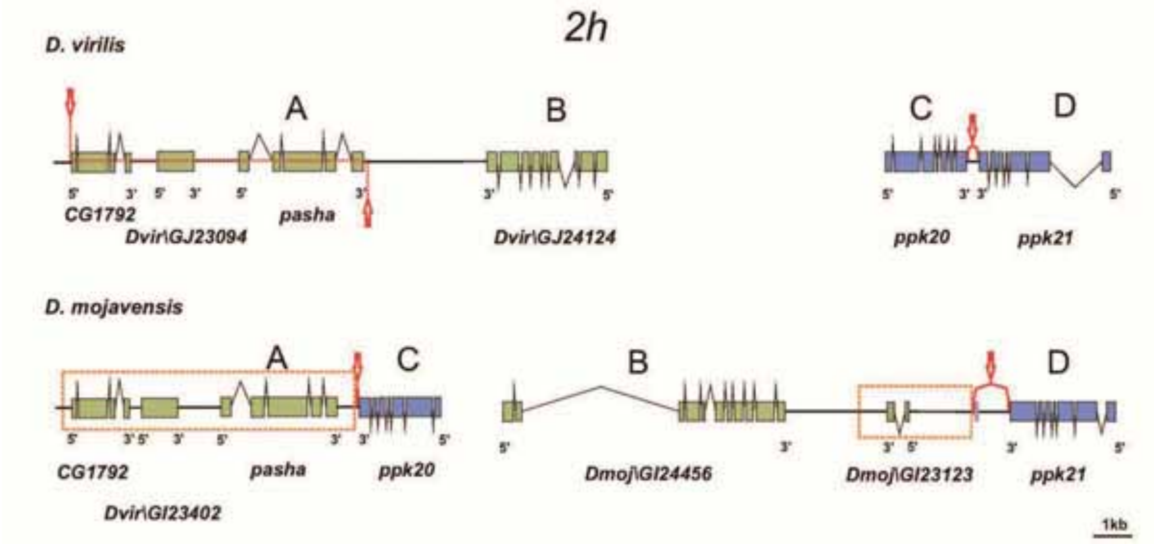
Syntenic segment	Begin	End	Size (bp)	Coverage (number of markers)
20	1721255	4692600	2971346	183
14	4743675	6104645	1360971	75
18	6137184	7154445	1017262	82
16	7172282	7222783	50502	9
9	7365221	7654616	289396	28
2	7664393	9955684	2291292	233
5	10436380	10941168	504789	61
3	10957988	12125979	1167992	98
8	12137327	12970351	833025	57
11	13067258	13124282	57025	10
7	13151145	13231800	80656	2*
10	13381003	15145288	1764286	155
17	15167727	16621615	1453889	173
13	16659223	16888133	228911	34
19	16903388	19774789	2871402	184
15	19825375	25751837	5926463	426
12	25824411	25953117	128707	30
6	25968812	26375571	406760	13*
4	26441888	31225471	4783584	350
1	31397073	34039404	2642332	172

*The complete sequence of the clone 01B03 was used as a marker (Prada 2010). This sequence mapped in two different regions of the chromosome 2, one belonging to the syntenic segment 6 and the other to the syntenic segment 7.

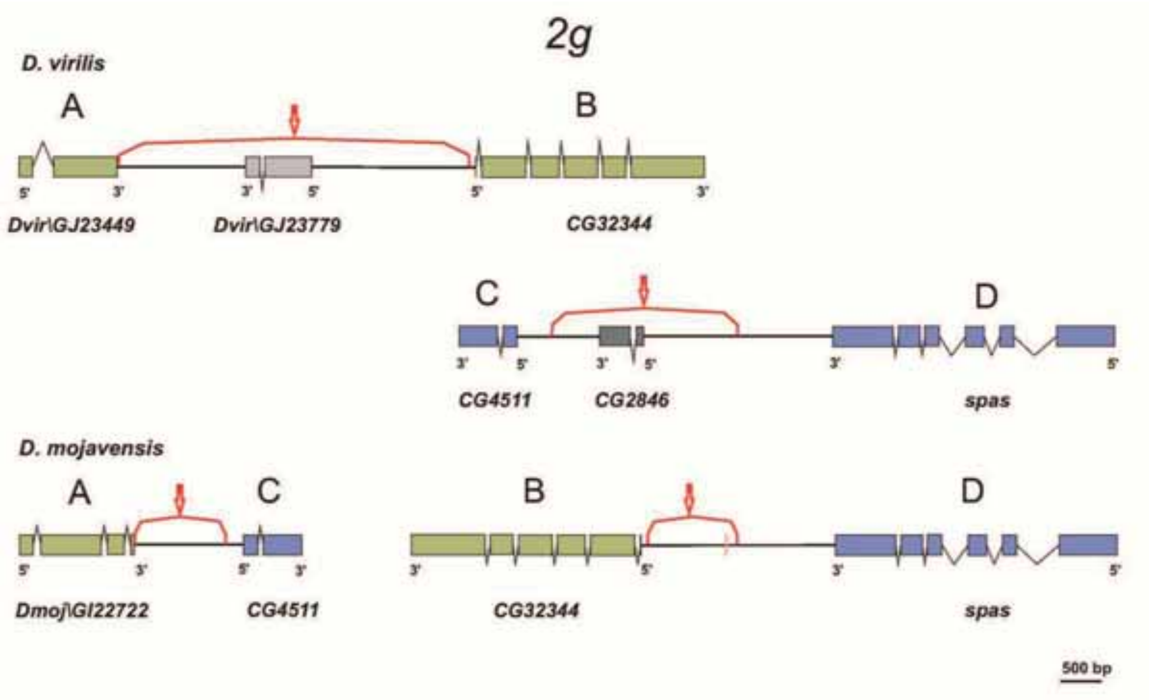
Additional file 2. Genome mapping of inversion breakpoint regions in the *D. mojavensis* genome.

Inversion	BP	Neighboring syntenic segments	Initial BES mapping			Similarity to *D. virilis* genome			CDS of neighboring genes		
			D. mojavensis coordinates		BP region (bp)	*D. mojavensis* coordinates		BP region (bp)	*D. mojavensis* coordinates		BP region (bp)
			Begin	End		Begin	End		Begin	End	
2c	Distal	3 – 5	10941169	10957987	16819	10951558	10952204	647			
	Proximal	4 – 6	26375572	26441887	66316	26378790	26379233	444			
2f	Proximal	11 – 8	12970352	13067257	96906	13059356	13061415	2060	13060199	13061415	1217
	Distal	10 – 7	13231801	13381002	149202	13376979	13377791	813			
2g	Proximal	16 – 18	7154446	7172281	17836	7159934	7161052	1119			
	Distal	15 -19	19774790	19825374	50585	19804465	19805612	1148	19804465	19805311	847
2h	Distal	2 – 9	7654617	7664392	9776	7664068	7664784	717	7664342	7664784	443
	Proximal	8 – 3	12125980	12137326	11348	12128366	12129507	1142	12128366	12129293	928
2q	Proximal	5 – 2	9955685	10436379	480695	10420224	10422204	1981			
	Distal	1 – 4	31225472	31397072	171601	31254883	31255399	517			
2r	Proximal	9 – 16	7222784	7365220	142437	7230145	7321956	91812			
	Distal	17 – 10	15145289	15167726	22438	15160462	15162581	2120	15160909	15162581	1673
2s	Proximal	7 – 11	13124283	13151144	26862	13149238	13149496	259			
	Distal	6 – 12	25953118	25968811	15694	25966954	25968814	1861			

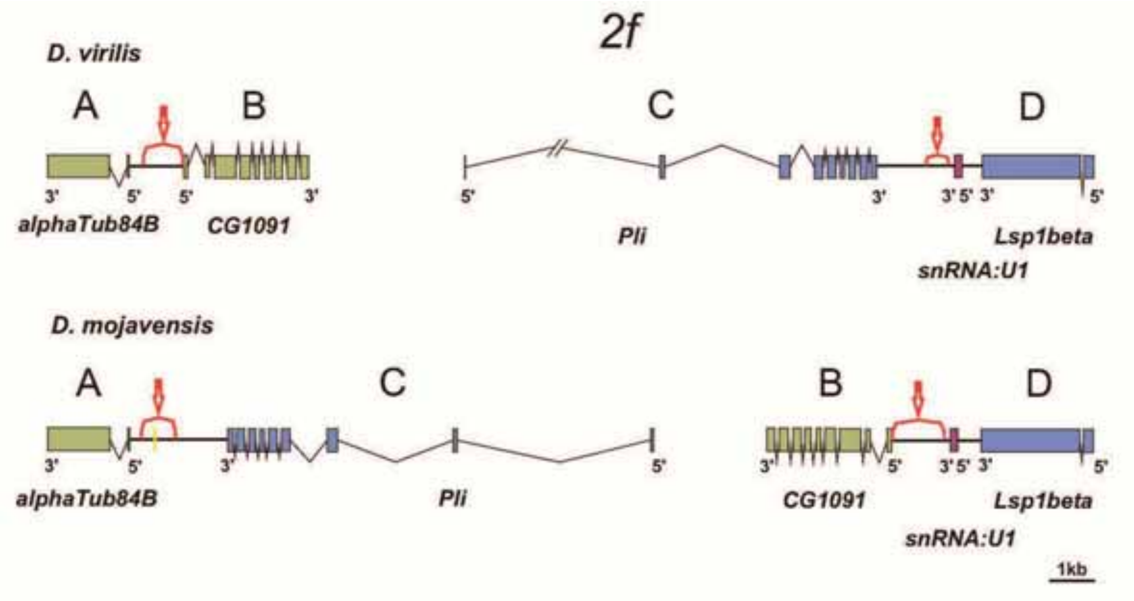
Additional file 3. Annotation of inversion *2h* breakpoint regions. Annotation of inversion *2h* distal and proximal breakpoint regions in*D. virilis* (non-inverted chromosome) and *D. mojavensis* (inverted chromosome). Inverted duplications in the *D. mojavensis* breakpoints are enclosed within dotted boxes, orange color. That in region AC (7.1 kb) is intact whereas that in region BD (2.7 kb) has suffered several deletions. These duplications were presumably generated by staggered single-strand breaks in the parental chromosome represented by a dotted red lines flanked by red arrows. A fragment of *Bu*T3 is shown as a blue rectangle in region BD. Other symbols as in Figure 4.



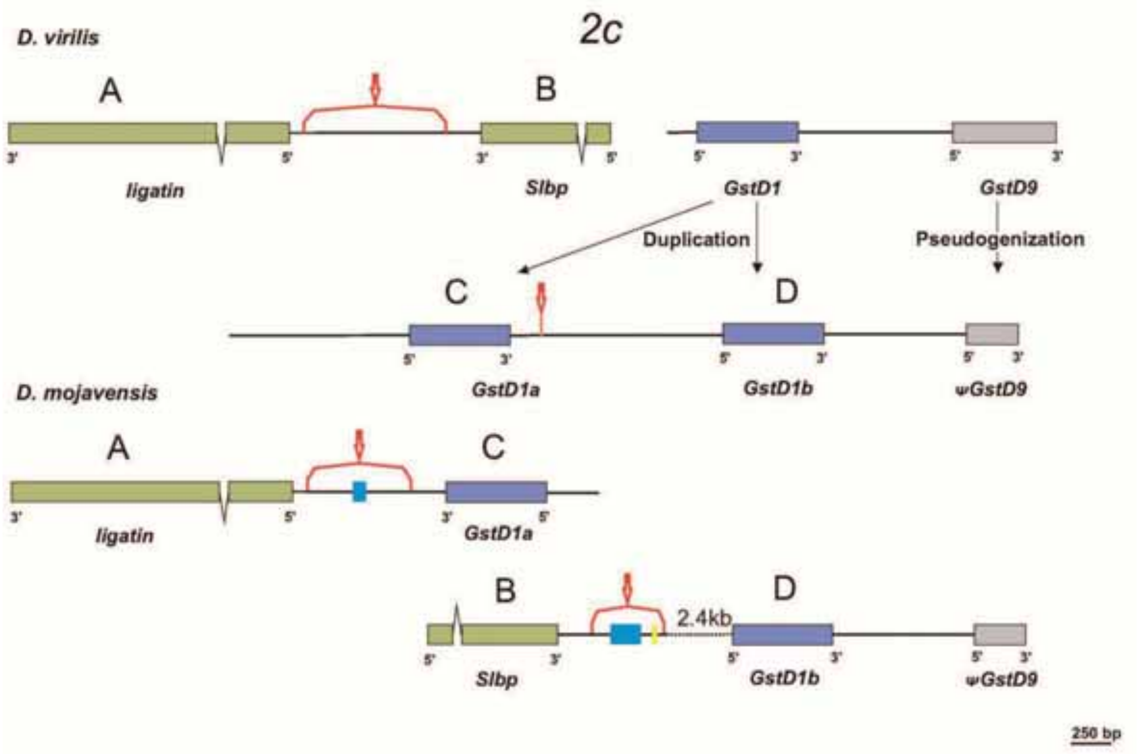
Additional file 4. Annotation of inversion *2g* breakpoint regions. Annotation of inversion *2g* distal and proximal breakpoint regions in *D. virilis* (non-inverted chromosome) and *D. mojavensis* (inverted chromosome). Two *D. virilis* lineage specific genes are shown as grey rectangles. Other symbols as in Figure 4.



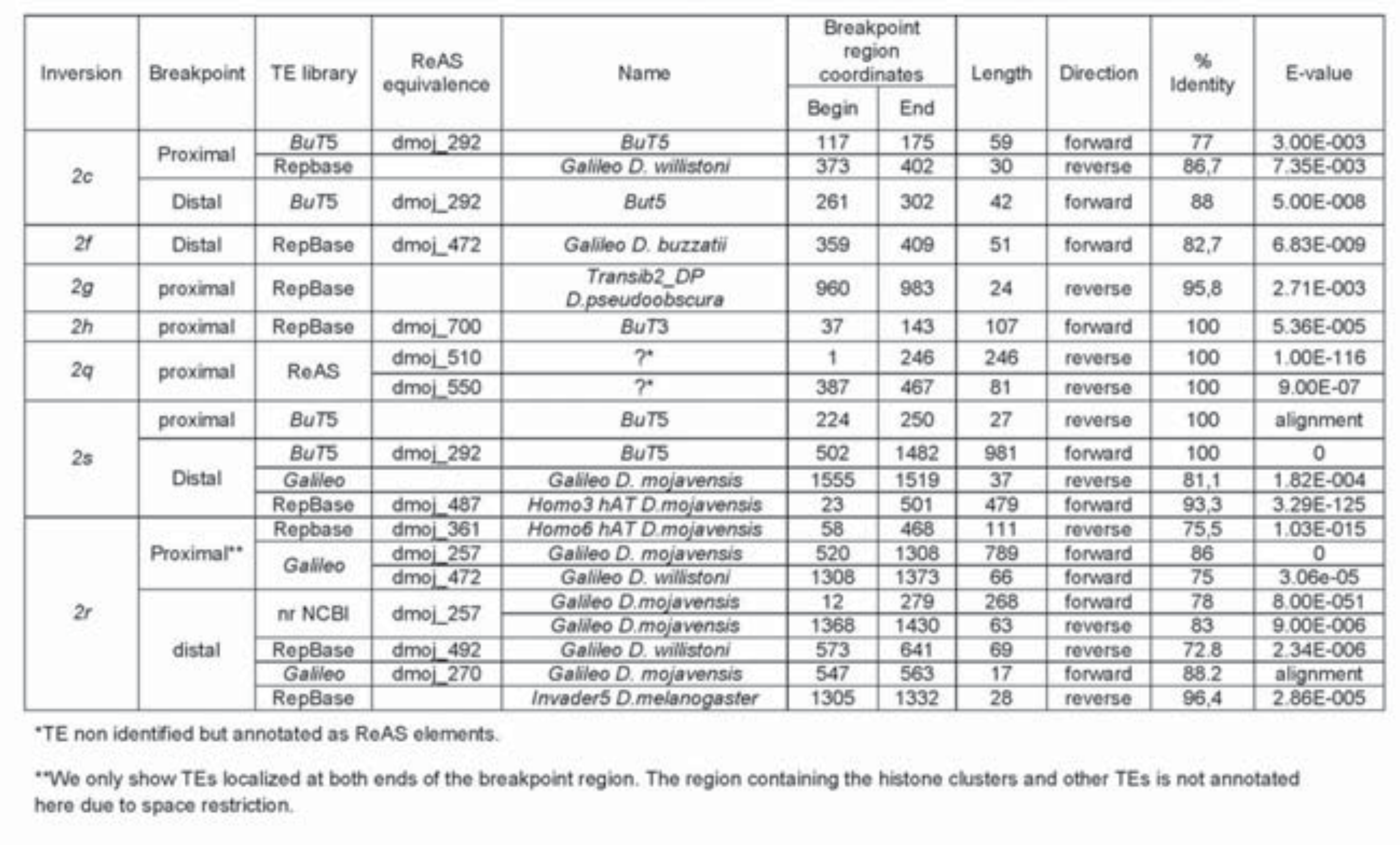
Additional file 5. Annotation of inversion *2f* breakpoint regions. Annotation of inversion *2f* distal and proximal breakpoint regions in *D. virilis* (non-inverted chromosome) and *D. mojavensis* (inverted chromosome). Symbols as in Figure 4.



Additional file 6. Annotation of inversion *2c* breakpoint regions. Annotation of inversion *2c* distal and proximal breakpoint regions in *D. virilis* (non-inverted chromosome) and *D. mojavensis* (inverted chromosome). Phylogenetic analysis of GstD genes (Additional file 8) indicates that the *2c* inversion occurred after the duplication of the GstD1 gene in the parental chromosome. The GstD9 gene has lost its function in *D. mojavensis* becoming a pseudogene. Other symbols as in Figure 4.



Additional file 7. TE content of inversion breakpoint regions in *D. mojavensis*.

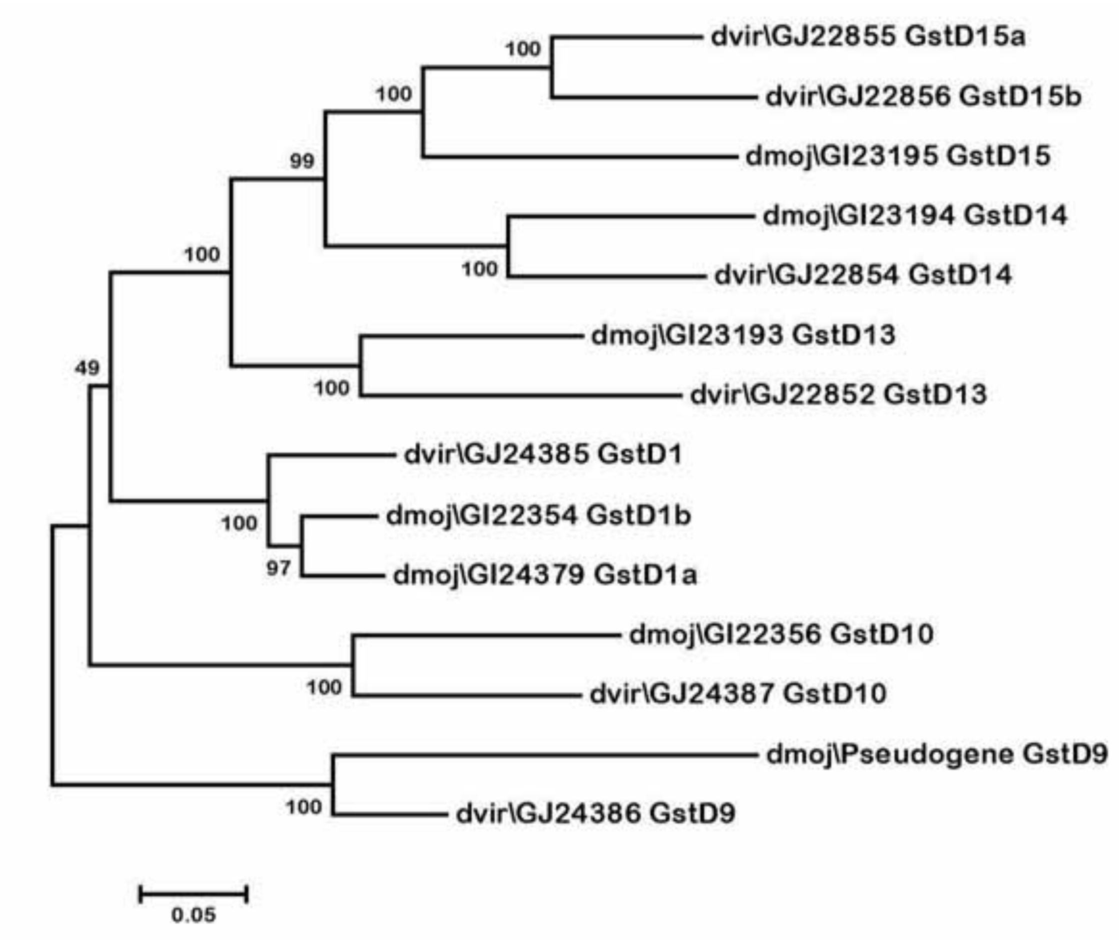


Inversion	Breakpoint	TE library	ReAS equivalence	Name	Breakpoint region coordinates		Length	Direction	% Identity	E-value
					Begin	End				
2c	Proximal	*BuT5*	dmoj_292	*BuT5*	117	175	59	forward	77	3.00E-003
		Repbase		*Galileo D. willistoni*	373	402	30	reverse	86,7	7.35E-003
	Distal	*BuT5*	dmoj_292	*But5*	261	302	42	forward	88	5.00E-008
2f	Distal	RepBase	dmoj_472	*Galileo D. buzzatii*	359	409	51	forward	82,7	6.83E-009
2g	proximal	RepBase		*Transib2_DP D.pseudoobscura*	960	983	24	reverse	95,8	2.71E-003
2h	proximal	RepBase	dmoj_700	*BuT3*	37	143	107	forward	100	5.36E-005
2q	proximal	ReAS	dmoj_510	?*	1	246	246	reverse	100	1.00E-116
			dmoj_550	?*	387	467	81	reverse	100	9.00E-07
2s	proximal	*BuT5*		*BuT5*	224	250	27	reverse	100	alignment
	Distal	*BuT5*	dmoj_292	*BuT5*	502	1482	981	forward	100	0
		Galileo		*Galileo D. mojavensis*	1555	1519	37	reverse	81,1	1.82E-004
		RepBase	dmoj_487	*Homo3 hAT D.mojavensis*	23	501	479	forward	93,3	3.29E-125
2r	Proximal**	Repbase	dmoj_361	*Homo6 hAT D.mojavensis*	58	468	111	reverse	75,5	1.03E-015
		Galileo	dmoj_257	*Galileo D. mojavensis*	520	1308	789	forward	86	0
			dmoj_472	*Galileo D. willistoni*	1308	1373	66	forward	75	3.06e-05
	distal	nr NCBI	dmoj_257	*Galileo D.mojavensis*	12	279	268	forward	78	8.00E-051
				Galileo D.mojavensis	1368	1430	63	reverse	83	9.00E-006
		RepBase	dmoj_492	*Galileo D. willistoni*	573	641	69	reverse	72.8	2.34E-006
		Galileo	dmoj_270	*Galileo D. mojavensis*	547	563	17	forward	88.2	alignment
		RepBase		*Invader5 D.melanogaster*	1305	1332	28	reverse	96,4	2.86E-005

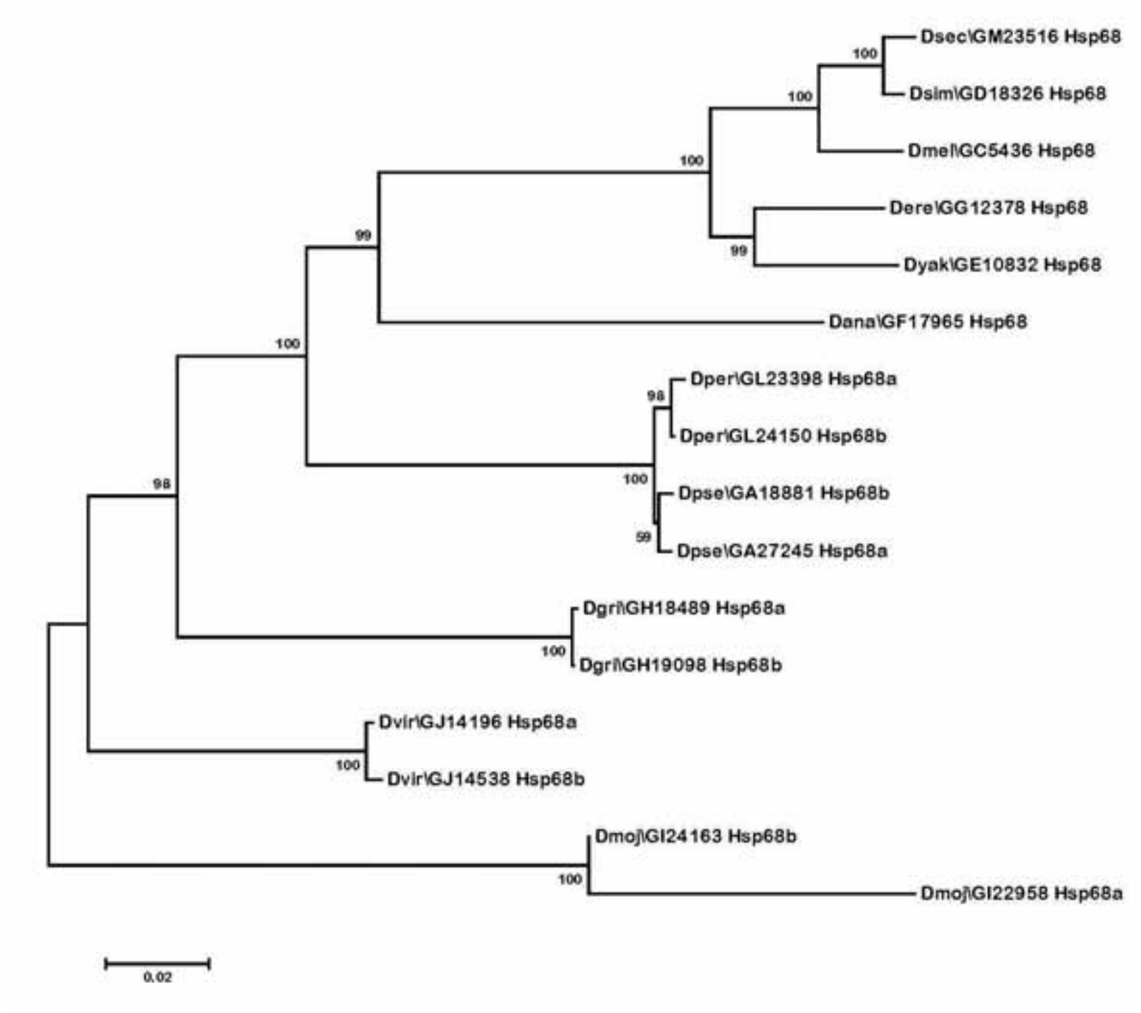
*TE non identified but annotated as ReAS elements.

**We only show TEs localized at both ends of the breakpoint region. The region containing the histone clusters and other TEs is not annotated here due to space restriction.

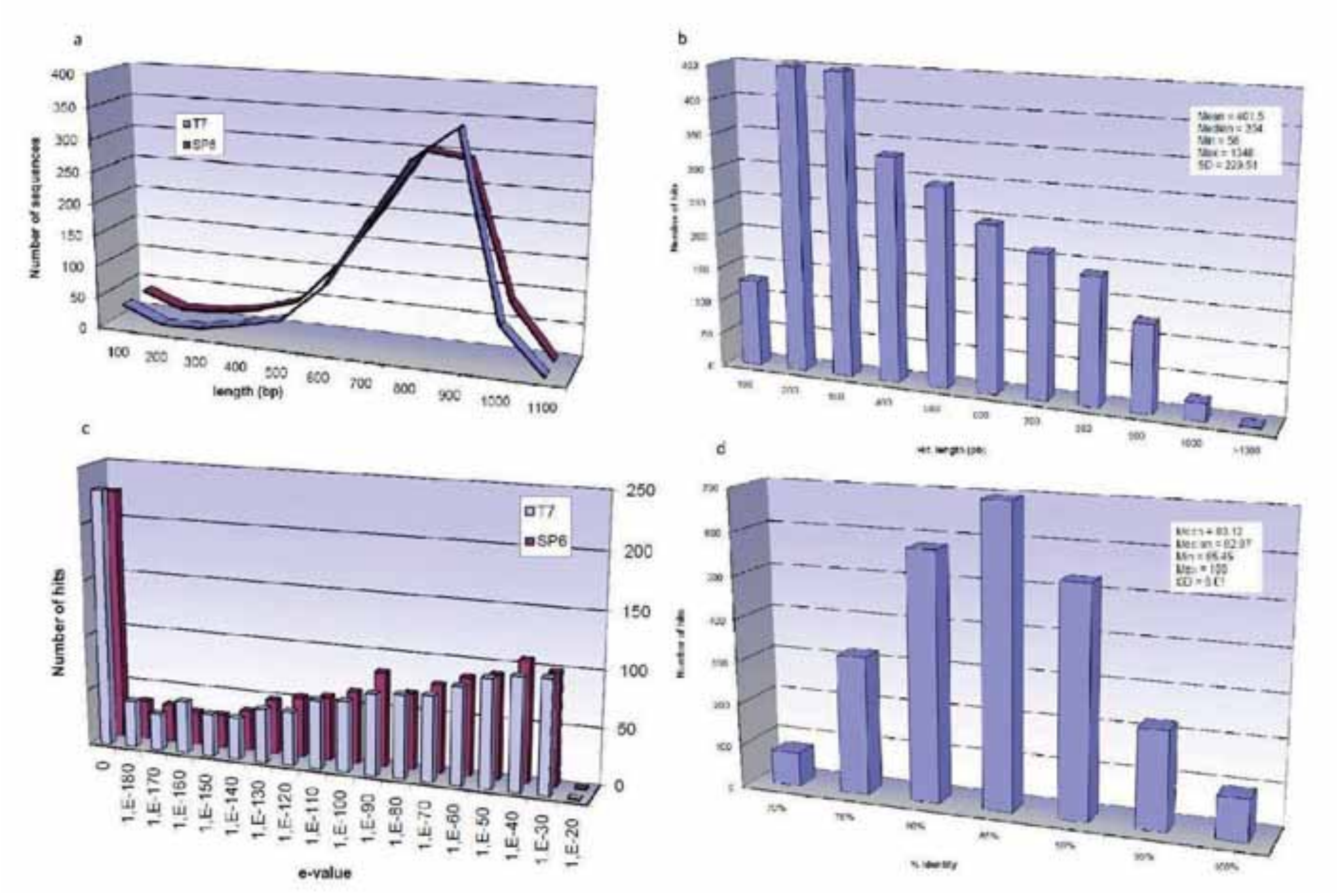
Additional file 8. Neighbor-Joining phylogenetic tree of GstD genes in *D. mojavensis* and *D. virilis*. Neighbor-Joining phylogenetic tree of GstD genes in *D. mojavensis* and D. *virilis*. Bootstrap values data for all tree nodes are shown. Phylogenetic analysis was conducted with MEGA4 [114]. Evolutionary distances were computed using the Maximum Composite Likelihood method.



Additional file 9. Neighbor-Joining phylogenetic tree of Hsp68 genes of 12 sequenced Drosophila species. Neighbor-Joining phylogenetic tree of Hsp68 genes of 12 sequenced Drosophila species. *D. persimilis*, *D. pseudoobscura*, *D. grimshawi*, *D. virilis* and *D. mojavensis* have two copies of the Hsp68 gene, while D. sechellia, D. simulans, D. melanogaster, D. erecta, *D. yakuba* and *D. ananassae* only one. No Hsp68 gene has been detected in *D. willistoni*. Bootstrap values for all tree nodes are shown. Phylogenetic analysis was carried out using MEGA4 [114]. Evolutionary distances were computed using the Maximum Composite Likelihood method.



Additional file 10. Statistics of *D. buzzatii* BAC end sequences. Description: Size distribution of *D. buzzatii* BAC end sequences (A) and distribution of size (B), E-value (C) and % identity (D) for hits generated blasting them against the *D. mojavensis* genome. See text for details.



4.2 Genomics of ecological adaptation in cactophilic Drosophila: hundreds of genes under positive selection in the *D. buzzatii* and *D. mojavensis* lineages

YOLANDA GUILLÉN et al. (2014) Genomics of ecological adaptation in cactophilic Drosophila: hundreds of gene under positive selection in the *D. buzzatii* and *D. mojavensis* lineages. *Manuscript submitted*.

Genomics of ecological adaptation in cactophilic Drosophila: hundreds of genes under positive selection in the *D. buzzatii* and *D. mojavensis* lineages

Yolanda Guillén[1], Núria Rius[1], Alejandra Delprat[1], Francesc Muyas[1], Marta Puig[1], Sònia Casillas[2], Miquel Ràmia[2], Raquel Egea[2], Gisela Mir[3], Jordi Camps[4], Valentí Moncunill[5], Robert L. Unckless[6], Aurelie Kapusta[7], Francisco J. Ruiz-Ruano[8], Josefa Cabrero[8], Guilherme B. Dias[9], Leonardo G. de Lima[9], Jeronimo Ruiz[9], Marta Gut[4], Ivo G. Gut[4], Jordi Garcia-Mas[3], David Torrents[5], Juan Pedro Camacho[8], Gustavo C.S. Kuhn[9], Andrew G. Clark[6], Cedric Feschotte[7], Antonio Barbadilla[2] and Alfredo Ruiz[1]

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

2 Plataforma Bioinformàtica de la UAB, Institut de Biotecnologia i de Biomedicina, Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona), Spain.

3 Centre for Research in Agricultural Genomics (CRAG), Campus UAB, Edifici CRAG, 08193 Bellaterra (Barcelona), Spain.

4 Parc Científic de Barcelona, Centro Nacional de Análisis Genómico (CNAG), Torre I, Baldiri Reixac 4, 08028 Barcelona, Spain.

5 Barcelona Supercomputing Center (BSC), Edifici TG (Torre Girona), Jordi Girona 31, 08034 Barcelona, Spain.

6 Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York, USA.

7 Department of Biology, University of Texas at Arlington, Arlington, TX 76019, USA.

8 Departamento de Genética,Universidad de Granada, Granada, Spain

9 Instituto de Ciências Biológicas, Departamento de Biologia Geral, Universidade Federal de Minas Gerais, Belo Horizontte (MG, Brazil)

ABSTRACT

We have sequenced the genome and developmental transcriptome of *D. buzzatii* using second-generation sequencing platforms to analyze the genomic basis of ecological adaptation in cactophilic Drosophila. *D. buzzatii* and *D. mojavensis*, its closest relative with a genome sequence, belong to the *repleta* group of the Drosophila subgenus, and both species feed and breed on decaying cactus tissues. The assembly (Freeze 1) of the *D. buzzatii* genome (~160 Mb) comprises 826 scaffolds (< 3 kb) with N50 and N90 indexes 30 and 158, respectively. The 158 N90 scaffolds were assigned to chromosomes X (48), 2 (7), 3 (38), 4 (26), 5 (35), and 6 (4), as well as ordered and oriented by conserved synteny and additional information. Transposable elements account for at least 8% of the *D. buzzatii* genome. Protein-coding genes (13,657, Annotation release 1) were annotated using *ab initio* and homology based algorithms. Using RNA-seq of five life-stages (embryos, larvae, pupae, adult females and males) we detected expression of 15026 genes, 80% protein-coding genes and 20% ncRNA genes. Comparison of single-copy orthologs between *D. buzzatii* and *D. mojavensis* revealed an influence of chromosome type, recombination and fixed inversions on synonymous (ds) and non-synonymous (dn) divergence. In addition, protein length, exon number, expression breadth and maximum expression level have a significant effect on ds whereas exon number and expression breadth are predictors for dn. Using maximum likelihood models implemented in PAML, we detected in cactophilic flies 1294 genes putatively under positive selection. Besides we found in cactophilic flies 117 orphan genes coding for proteins with no similarity to any predicted Drosophila protein. These genes are clear candidates for involvement in adaptation of these flies to their ecological conditions.

INTRODUCTION

Comparative genomics provides us with the opportunity to investigate the evolution of genes and genomes at an unprecedented scale. The sequencing and *de novo* assembly of eukaryotic genomes is a feasible, although by no means easy, task with second-generation sequencing platforms such as Roche 454 or Illumina (Mardis 2008; Shendure and Ji 2008; Baker 2012). With the genomes of two or more related species in hand, an opportunity is open to investigate questions on the evolution of chromosomes or particular chromosome regions, protein-coding genes (PCG) and gene families, non-coding RNA (ncRNA) genes, transposable elements (TE), regulatory sequences, and so forth. Furthermore, several comparative genomic methods have been developed to carry out genome-wide scans for genes evolving under positive selection (Yang and Bielawski 2000; Nielsen et al. 2005; Anisimova and Liberles 2007). These methods are usually based on the comparison of the nonsynonymous substitution rate (dN) with the synonymous substitution rate (dS), which under neutrality should be equal. The ratio ω = dN/dS is a measure of selection pressure at the protein level and a ratio $\omega < 1$ indicates purifying selection whereas $\omega > 1$ is usually taken as indication of positive selection. This test to detect positively selected genes is manifestly conservative at the gene level because different sites can evolve under different selection pressures or neutrally and therefore will cancel each other out. However, site models and branch-site models implemented in PAML allow carrying out the analysis at the codon level thus increasing power (Wong et al. 2004; Zhang et al. 2005; Yang 2007). Positively selected genes are likely to be responsible for the adaptation of species to their ecological conditions, yet some of them may be responsible to internal adaptations or to intraspecific or sex interactions.

Drosophila is a leading model for comparative genomics (Drosophila 12 Genomes Consortium et al. 2007; Singh et al. 2009). The Drosophila genus is large and diverse with > 2,000 known species. Phylogenetic analyses indicate that two main lineages exist, which diverged ~60 myr ago (Tamura et al. 2004). One lineage led to the *Sophophora* subgenus comprising more than 300 species, whereas the other one led to the subgenus *Drosophila*, with about 1700 species. *D. melanogaster*, a species belonging to Sophophora subgenus, is a centenary model species for studies in genetics and development with one of the first sequenced and best annotated eukaryotic genomes (Adams et al. 2000; Rubin and Lewis 2000, Celniker and Rubin 2003). Furthermore, the genomes of another 23 Drosophila species have already been sequenced and annotated, providing a valuable resource for comparative genomics. These species are: *D. simulans*, *D. sechellia*, *D. yakuba*, *D. erecta*, *D. ficusphila*, *D. eugracilis*, *D. biarmipes*, *D. takahashii*, *D. elegans*, *D. rhopaloa*, *D. kikkawai*, *D. ananassae*, *D. bipectinata*, *D. suzukii*, *D*, *pseudoobscura*, *D. persimilis*, *D. miranda* and *D. willistoni* in the Sophophora subgenus; *D. mojavensis*, *D. virilis*, *D. americana*, *D. grimshawi* and *D. albomicans* in the Drosophila subgenus (Drosophila 12 Genomes Consortium et al. 2007, 12; Zhou and Bachtrog 2012; Zhou et al. 2012; Ometto et al. 2013; Fonseca et al. 2013). The ecological diversity of the completely sequenced Drosophila genomes is considerable including species inhabiting different geographical locations separated by a wide range of evolutionary distances (Drosophila 12 Genomes Consortium et al. 2007; Markow and O'Grady 2007; Singh et al. 2009). This genomic data will make possible to better understand the patterns of ecological adaptation and genome evolution in a fine-scale approach.

The *repleta* species group of the Drosophila subgenus comprises >100 species living in the deserts and arid zones of the American continent (Wasserman 1982, 1992). Many of them are cactophilic species that use as feeding and breeding substrates the decaying stems and fruits of different cacti. The cactus-yeast-Drosophila system in arid zones provides a valuable model to investigate gene-environment interactions and ecological adaptation from a genetic and evolutionary perspective (Barker and Starmer 1982; Barker et al. 1990, Etges et al. 1999; Fogleman and Danielson 2001). Some

Drosophila species are able to colonize cactus widely distributed along different geographical areas. In contrast, specialist species are restricted to certain environments and have limited growing conditions (Patterson and Stone 1953; Wasserman 1982, 1992; Vilela 1983). Niche specificity depends on a variety of ecological factors like the availability of nutrition resources or tolerance to toxic compounds present in the host plant (Heed 1978; Kircher 1982; Ruiz and Heed 1988). For instance, senita cactus (*Lophocereus schottii*) is the unique host plant of *D. pachea*, one of the four endemic Drosophila species inhabiting the Sonora Desert (Heed and Mangan 1986). This plant has a characteristic chemical composition (unique sterols and toxic alkaloids) that make it unsuitable for other Drosophila species (Kircher et al. 1967). Seemingly a few positive selected changes in the gene *Neverland* turned *D. pachea* into an obligate specialist (Lang et al. 2012). These results evidenced that the ecological niche can be determined by few but crucial mutations.

We have sequenced the genome and developmental transcriptome of *D. buzzatii* to carry out a comparative analysis with those of *D. mojavensis*, its closest relative with a sequenced genome, and other species. *D. buzzatii* and *D. mojavensis* belong to the *repleta* group of the Drosophila subgenus and diverged ~12 mya (Figure 1). However, they have different geographical distributions and hostplants. *D. buzzatii* is a subcosmopolitan species which is found in four out of the six major biogeographic regions associated with prickly pear and other cacti (David and Tsacas 1980). This species is original from Argentina and Bolivia but has now a wide geographical distribution that includes other regions of South America (Uruguay, Paraguay, Brazil, Peru, and Chile) and the Old World (Iberian Peninsula and Mediterranean Basin) and Australia (Carson and Wasserman 1965; Fontdevila et al. 1981; Hasson et al. 1995; Manfrin and Sene 2006). It chiefly feeds and breeds in rotting tissues of cactus from Opuntia genus (*O. ficus-indica*, *O. quimilo*, *O. monacantha*, *O. sulphurea*, *O. pampeana*, *O. aurantiaca*) but can also use occasionally columnar cacti (*Echinopsis terschekii*, *Cereus hildmannianus*) (Hasson et al. 1992; Ruiz et al. 2000). The geographical diffusion of Opuntia by humans in historical times is considered the main cause of *D. buzzatii* world-wide colonization (Fontdevila et al. 1981; Hasson et al. 1995).

D. mojavensis is endemic to the deserts of the Southwestern USA and Northwestern Mexico, chiefly the Sonoran Desert (Arizona, Baja California and Sonora), the Mojave Desert and Santa Catalina Island in southern California. Its primary host plants are *Stenocereus gummosus* (pitaya agria) in Baja California, *Stenocereus thurberi* (organ pipe) in Arizona and Sonora, *Ferocactus cylindraceous* (California barrel) in Southern California and *Opuntia demissa* in Santa Catalina Island (Fellows and Heed 1972; Heed and Mangan 1986; Ruiz and Heed 1988; Etges et al. 1999). The ecological conditions of the Sonoran Desert are extreme as attested by the fact that only four Drosophila species are endemic (Heed and Mangan 1986). The analysis of the chemical composition of pitaya agria and organ pipe revealed that they contain large quantities of triterpene glycosids as well as unusual medium-chain fatty acids and sterol diols (Kircher 1982; Fogleman and Danielson 2001). These natural organic allelochemicals have been related to important biological activities in animals and plants (Natori et al. 1981; Fogleman and Armstrong 1989). Even though it has been proposed that both chemical and physical aspects of these plants affect the host specificity of *D. mojavensis*, there is no clear evidence of this relationship from a genetic point of view (Kircher 1982; Matzkin et al. 2006).

Here we seek to understand the genetic bases of ecological adaptation by comparing the genomes of the two Drosophila cactophilic species and another two non-cactophilic species of the Drosophila subgenus, *D. virilis* and *D. grimshawi* (Figure 1). We estimated the divergence at synonymous and nonsynonymous sites in 9017 orthologous protein-coding genes between *D. buzzatii* and *D. mojavensis* and tested for the effect on divergence of seven genomic variables. In addition, using maximum likelihood methods, we carried out a genome-wide scan for genes under positive selection in the *D. buzzatii* and *D. mojavensis* lineages as well as the shared cactophilic lineage of the Drosophila subgenus (Figure 1*)*. We postulated that positive selected loci are the main candidates involved in specific environment adaptation (Lang et al. 2012; Amemiya et al. 2013). Based on our comparative analyses results we propose that candidate genes under positive selection likely play a meaningful role in the chemistry of the interactions between the fruit flies and their host plants.

RESULTS

Genome sequencing and assembly

We sequenced and assembled *de novo* the genome of *D. buzzatii* line st-1 using shotgun and paired-end reads from 454/Roche, mate-pair and paired-end reads from Illumina, and Sanger BAC-end sequences (~22x total expected coverage; see Materials and Methods for details). The resulting assembly (Freeze 1) is considered the reference *D. buzzatii* genome sequence (Table 1). This assembly comprises 826 scaffolds >3 kb long with a total size of 161.5 Mb. Scaffold N50 and N90 indexes are 30 and 158, respectively whereas scaffold N50 and N90 lengths are 1.38 and 0.16 Mb, respectively (Table 1). Quality controls performed comparing the reference genome sequence with five BACs sequenced previously using Sanger and with genomic and RNA-seq reads generated with Illumina (see Materials and Methods) yielded a relatively low error rate of ~ 0.0005 (Q33). For comparison, we also assembled the genome of the same line (st-1) with the SOAPdenovo software (Luo et al. 2012) using only four lanes of short (100 bp) Illumina paired-end reads (~76x expected coverage). This resulted in 10949 scaffolds >3 kb long with a total size of 144.2 Mb (Table 1). All scaffolds are available for download from the *Drosophila buzzatii* Genome Project web page (http://dbuz.uab.cat). This site also displays all the information generated in this project (see below).

Genome size estimation

The genome sizes of two *D. buzzatii* strains, st-1 and j-19, were estimated by Feulgen Image Analysis Densitometry on testis cells (Ruiz-Ruano et al. 2011) using *D. mojavensis* as reference. Integrative Optical Density (IOD) values were 21% (st-1) and 25% (j-19) smaller than those for *D. mojavensis*. Thus, taking 194 Mb (total assembly size) as the genome size of *D. mojavensis* (Drosophila 12 Genomes Consortium et al. 2007) we estimated the genome size for *D. buzzatii* st-1 and j-19 lines as 153 and 146 Mb, respectively.

Chromosome organization and evolution

The basic karyotype of *D. buzzatii* is similar to that of the Drosophila ancestor and consists of six chromosome pairs four pairs of equal-length acrocentric autosomes, one pair of dot autosomes, a long acrocentric X and a mall acrocentric Y (Ruiz and Wasserman 1993). Because interchromosomal reorganizations between *D. buzzatii* and *D. mojavensis* are not expected (Ruiz et al. 1990; Ruiz and Wasserman 1993) the 158 scaffolds in the N90 index were assigned to chromosomes by blastn against the *D. mojavensis* genome using MUMmer (Delcher et al. 2003). The number of scaffolds in chromosomes X, 2, 3, 4, 5, and 6 were 48, 7, 38, 26, 35 and 4, respectively (Figure 2). The seven scaffolds corresponding to chromosome 2 were ordered and oriented using *D. buzzatii* BAC-based physical map and BAC-end sequences (Gonzalez et al. 2005, Guillén and Ruiz 2012). Following Schaeffer et al. (2008), the scaffolds corresponding to the remaining chromosomes were ordered and oriented using a combination of conserved linkage and in situ hybridizations (Delprat et al. in preparation). A comparison of *D. buzzatii* and *D. mojavensis* chromosomes using MUMmer (Delcher et al. 2003) and GRIMM (Tesler 2002) confirmed that chromosome 2 differs between the two species by 10 inversions (2m, 2n, $2z^7$, 2c, 2f, 2g, 2h, 2q, 2r, 2s), chromosomes X and 5 differ by one inversion each (Xe and 5g, respectively) and chromosome 4 is homosequential (Ruiz et al. 1990; Ruiz and Wasserman 1993, Guillén and Ruiz 2012). By contrast, chromosome 3 showed six inversions of difference instead of the two inversions expected by previous cytological analyses, 3a and 3d (Ruiz et al. 1990). The four additional chromosome 3 inversions seem to have been fixed not in the *D. buzzatii* lineage but in the *D.*

mojavensis lineage. One of them is inversion 3f^2, polymorphic in *D. mojavensis*, which is seemingly fixed in the sequenced strain (in contrast to previous reports; Ruiz et al. 1990, Schaeffer et al. 2008).

Hox genes were arranged in a single complex in the Drosophila ancestor. However, this HOM-C suffered two splits in the lineage leading to the repleta species group (Negre et al. 2005). We previously characterized three of the eight Drosophila Hox genes in *D. buzzatii*, *labial* (*lab*), *proboscipedia* (*pb*) and *abdominal* (*abdA*) (Negre et al. 2005). In order to fully characterize HOM-C organization in *D. buzzatii*, we manually annotated all Hox genes using EVM and Exonerate predictions (see below) as well as RNA-seq information (see below) and available information for *D. buzzatii*, *D. mojavensis* and *D. melanogaster* (Supplemental Table S1). Hox genes are distributed into three scaffolds (2, 5 and 229) of chromosome 2 (Figure 3). However, our analysis revealed that the gene *Deformed* (*Dfd*) belongs to scaffold 2 although it has been misassembled into a separate scaffold (229). Thus only two clusters of genes are present (Figure 3). The distal one contains *pb*, *Dfd*, *Sex combs reduced* (*Scr*), *Antennapedia* (*Antp*) and *Ultrabithorax* (*Ubx*) whereas the proximal one contains *lab*, *abdA* and *Abdominal B* (*AbdB*). This is precisely the same HOM-C organization observed in *D. mojavensis* (Negre and Ruiz 2007). Therefore there seem to be no additional rearrangements of the HOM-C in *D. buzzatii* besides those already described in the genus Drosophila (Negre and Ruiz 2007).

Repeat content

To assess the transposable element (TE) content of the *D. buzzatii* genome we masked the 826 scaffolds of Freeze 1 assembly using a library of TEs compiled from several sources (see Materials and Methods). We detected a total of 57109 TE copies covering ~8% of the genome (Table 2). The most abundant TEs seem to be rolling-circle

Helitrons that cover 3.2% of the genome and the less abundant TIR transposons that comprise 1.2%. LINEs and LTR retrotransposons represent 1.5% and 1.4%, respectively (Table 2). In addition, we identified tandemly repeated satellite DNAs (satDNA) with repeat units longer than 50 bp (Melters et al. 2013) using Tandem Repeats Finder (TRF) program (see Materials and Methods). The pBuM189 satellite (Kuhn et al. 2008), with repeat units 189 bp long, was identified as the most abundant tandem repeat family, covering 0,039% of the genome (Table 3). The second most abundant tandem repeat family (DbuTR198) is novel, showed repeat units 198 bp long and covers 0,027% of the genome (Table 3). The remaining tandem repeats had sequence similarity to integral parts of TEs, such as the internal tandem repeats of the Galileo transposon (data not shown) (Casals et al. 2006).

Protein-coding gene content

We used different *ab initio* and homology-based algorithms (NSCAN, SNAP, Augustus and Exonerate) to annotate protein-coding genes (PCG) in the *D. buzzatii* reference genome. Predictions were combined with EVidence Modeler generating 12,102 gene models. We noticed that orthologs for a considerable number of *D. mojavensis* PCG were absent from this data set. Thus, we used the homology-based method Exonerate to detect another 1,555 PCG (Poptsova and Gogarten 2010). Therefore, we predicted a total of 13,657 PCG models in the *D. buzzatii* reference genome (Annotation Release 1). These PCG models contain a total of 52,250 exons with an average of 3.8 exons per gene. Gene expression analyses (see below) provided transcriptional evidence for 88.4% of these gene models.

The number of PCG in the *D. buzzatii* genome is lower than that in the genome of *D. mojavensis* (the closest relative) but similar to that in the genome of *D. melanogaster* (one of the best annotated eukaryotic genomes) (Supplemental Table S2).

However PCG in both *D. buzzatii* and *D. mojavensis* genomes tend to be smaller and contain less exons than those in the *D. melanogaster* genome which suggests that the annotation in the two cactophilic species might be incomplete. After performing multiple quality controls on the *D. buzzatii* PCG set, a total of 12,977 putatively well annotated coding sequences (CDS) were selected for further analysis (see Material and Methods).

Developmental transcriptome

To characterize the expression profile along *D. buzzatii* development we performed RNA-seq experiments by collecting samples from five different stages: embryo, larvae, pupae, adult female and adult male. We used Illumina sequencing platform to generate non-strand-specific paired-end ~100 bp reads from poly(A)+ RNA. A total of ~286 million filtered reads were mapped to Freeze 1 with Tophat representing ~180 x coverage of the total genome size (see Materials and Methods).

Transcripts were assembled with Cufflinks using the Annotation Release 1 as reference (see Materials and Methods). PCG models that did not show evidence of transcription by RNAseq were classified as non expressed PCG. Transcribed regions that did not overlap to any annotated PCG model were considered non-coding RNA (ncRNA) genes (Figure 4a). Gene expression levels were calculated based on FPKM values. We detected expression (FPKM > 1) of 26,455 transcripts and 15,026 genes, 12,066 (80%) are PCG and 2,960 (20%) are ncRNA genes. The number of expressed genes is highest in pupae and male adults (12,059 and 12,171 genes respectively) whereas it is much lower in embryos and larvae (9,760 and 9,519 genes respectively) (Figure 4a). Adult males express 1,824 more genes than adult females.

Expression breadth is radically different for PCG and ncRNA genes (Figure 4b). A total of 6,546 expressed PCG (54.2%) are constitutively expressed (i.e. we observed expression in the five stages) but only 260 of ncRNA genes (8.8%) are constitutively expressed. In contrast, 925 expressed PCG (7.7%) and 1,292 ncRNA genes (43.6%) are expressed only in one stage (Figure 4b). These differences are highly significant (P< 0.0001). Mean expression breadth was 3.9 for PCG and 2.2 for ncRNA genes. Adult males show more stage-exclusive expressed genes (844 genes) compared to adult females (137 genes), the group with less number of stage-exclusive expressed genes.

Protein coding gene evolution

A total of 11,154 single-copy orthologs between *D. buzzatii* and *D. mojavensis* were detected (see Materials and Methods). Orthologous proteins usually showed a similar size in *D. buzzatii* and *D. mojavensis* (median sizes 406 and 407 aa, respectively). However, there were a number of orthologous genes coding for proteins with a length difference >20%. Because this protein length difference might be due to incompletely or incorrectly annotated genes (see Materials and Methods), these PCG were discarded for subsequent analyses to avoid biases in the results, leaving a set of 9,114 orthologs between *D. buzzatii* and *D. mojavensis*. Furthermore, in order to correlate divergence estimates with seven genomic variables (see below), we restricted the analysis of divergence to a complete data set of 9,017 orthologs with information for all seven variables.

Overall median estimates for the number of non-synonymous (dn) and synonymous (ds) substitutions were 0.0343 and 0.4043, respectively (Table 4). The median estimate for the ratio ω = dn/ds was 0.0895 that indicates a relatively high level of functional constrain in most genes. However, divergence estimates show a considerable variation among and within the six chromosomes (Figure 3). Median divergence rates dn and ds vary significantly among all chromosomes (dn: X^2=21.38, P=0.0007; ds: X^2=60.79, P=8e-12); among-chromosome variation was non-significant for ω. In addition, dn and ds are higher for genes located in chromosome X than for those in the autosomes (dn: X^2=8.36, P=0.0038; ds: X^2=21.61, P=3e-6). The ratio w is also higher but nonsignificant (Table 4).

We also found that all three divergence parameters are significantly higher for genes in the non–recombining chromosome 6 (dot) than for those in the rest of autosomes (dn: X^2=8.10, P=0.0044; ds: X^2=15.45, P=8.5e-5; ω: X^2=3.96, P=0.0466). Finally, we tested for a correlation between nucleotide and structural divergences by comparing divergence estimates for genes in chromosomes 2 and 3 that harbor 10 and 6 fixed chromosomal inversions, respectively, between *D. mojavensis* and *D. buzzatii* (see above) with those for genes in chromosomes 4 and 5, with 0 and 1 fixed inversion, respectively. The results indicate that ds is significantly higher in genes located in chromosomes with more fixed inversions (X^2=22.87, P=2e-06) but dn and ω are not significantly different.

We used multiple linear models to test the dependence of divergence rates (dn, ds and ω) on seven genomic factors (Table 5). These factors are: chromosome type (X versus autosomes), recombination (non-recombining versus recombining regions), state (inverted versus non-inverted regions), protein length, exon number, expression breadth and maximum expression level. Some of these variables show significant pairwise correlations (see Materials and Methods and Table S13) and the joint analysis using linear models intended to disentangle their effects. The determination coefficients (Multiple R^2) of the three linear models (one for each independent variable, dn, ds and ⍰) are highly significant ($P < 2.2e-16$) (Table 5). All seven regressors have a significant effect on ds. Chromosome type, recombination, exon number and expression breadth are statistically significant as predictors for dn, whereas chromosome type, protein length, exon number and expression breadth have a significant effect on ω. The estimation of the relative importance of each variable in the linear models revealed that the contribution of each genomic factor varies among dn, ds and ω. Expression breadth is the variable with the more relative importance in dn and ω linear models. In the case of ds, exon number is the genomic factor that has more importance in the proposed model.

Genes under positive selection

We first identified genes that evolved under positive selection during the divergence between *D. buzzatii* and *D. mojavensis* using codon substitution models implemented in PAML 4 package (Yang 2007). Two pairs of different site models (SM) were compared by LRT, M1a vs. M2a and M7 vs. M8 (see Materials and Methods). In each case, a model that does allow for sites with $\omega > 1$ (positive selection) is compared with a null model that considers only sites with $\omega < 1$ and $\omega = 1$. The first comparison (M1a vs M2a) detected 915 genes while the second comparison (M7 vs M8) detected 802 genes, in both cases under the rather strict criterion of $P < 0.001$. Comparison of the two gene sets allowed us to detect 772 genes present in both, and this was taken as the final list of genes putatively under positive selection using SM (see Supplemental Table S4 for the list of genes).

We tested for a random distribution among chromosomes of the 772 genes under positive selection detected with SM. A highly significant departure was found ($X^2 = 32.28$, P=2e-6). The main cause is a significant excess of genes under selection in the X chromosome in comparison with the autosomes ($X^2 = 23.80$, P=e-6). When chromosome 6 (dot) was compared with the rest of autosomes, no significant departure was found. However we did detect a significant lower number of genes under selection in rearranged chromosomes 2 and 3 when compared with chromosomes 4 and 5 with few or no fixed inversions ($X^2 = 6.39$, P=0.01). A linear model with the same seven variables used to analyze divergence (see above) was used to analyze the distribution of genes under selection. Although Multiple R^2 was low (0.05), it was highly significant (P < 2.2e-16). This analysis It corroborated a positive effect of the X chromosome on the number of genes under selection (P = 1e-8) and a negative effect of recombination, i.e. less genes under selection in non-recombining regions (P = 0.02). The effect of inversions, however, although negative, was non-significant.

In addition, we found a negative effect of expression breadth (P = 7e-10) and a positive effect of protein length (P = 1.8e-8) and exon number (P < 2e-16).

Next, we used branch-site models (BSM) from PAML 4 package (Yang 2007) to identify genes under natural selection in a phylogeny with four Drosophila subgenus species, *D. buzzatii*, *D. mojavensis*, *D. virilis* and *D. grimshawi* (Figure 1). Orthology relationships among the four species were inferred from *D. buzzatii-D. mojavensis* list of orthologs and the OrthoDB catalog (version 6). A total of 8,328 unequivocal 1:1:1:1 orthologs were included in the comparison of a branch-site model allowing sites with ⍰ > 1 (positive selection) and a null model that does not. We selected three branches to test for positive selection (the foreground branches): *D. buzzatii* lineage, *D. mojavensis* lineage and cactophilic lineage (denoted as #1, #2 and #3 in Figure 1). The number of genes under positive selection detected in the three branches was 350, 172 and 458, respectively (see Supplemental Table S4 for the list of genes). These genes only partially overlap those previously detected in the *D. buzzatii-D. mojavensis* comparison using SM (Figure 6). While 69.4% and 55.8% of the genes selected in the *D. buzzatii* and *D. mojavensis* lineages had already been detected in the *D. buzzatii-D. mojavensis* comparison, only 22.3% of the genes detected in the cactophilic lineage were present in the previous list (Figure 6). Thus the total number of genes under positive selection is 1,294.

The main candidate genes involved in specific environment adaptation are those considered under positive selection. To understand patterns of adaptation we looked for functional categories overrepresented among the selected candidates reported by both site and branch-site models (Table 6).

We first performed a GO analysis on the 772 positive selected genes obtained by site models comparing *D. mojavensis* and *D. buzzatii* orthologs using DAVID tools (Huang et al. 2007). Two molecular functions show higher proportion within the candidate genes list than expected by random: antiporter activity and transcription factor activity. With respect to the biological process, regulation of transcription is the only overrepresented category. A significant enrichment in Src Homology-3 domain has been observed. This domain is commonly found within proteins with enzymatic activity and it is associated to protein binding function.

A similar GO analysis was carried out for candidate genes obtained in each of the three targeted branches when performing branch site models. Positive selected candidate genes in *D. buzzatii* lineage show a significant enrichment in DNA-binding function. DNA-dependent regulation of transcription and phosphate metabolic processes were overrepresented in the list of 350 genes. We also found a significant enrichment in a domain involved in functions related to cell-cell recognition and immune system, the Ig-like domain.

The 172 positively selected genes in *D. mojavensis* lineage show a significant excess of genes related to heterocycle catabolic process (P=5.9e-04). As we mentioned in the introduction, columnar cacti, the main host of *D. mojavensis*, contain large quantities of tryterpene glycosids, an heterocyclic compound. These results will be discussed below.

Among the positive selected genes in the branch that lead to cactophilic species, there are three overrepresented molecular functions related to both metal and DNA binding. The GO terms with the highest significance in biological process category are cytoskeleton organization and once again regulation of transcription.

We tested for a random distribution of positively selected genes among chromosomes. A highly significant departure was found when the total number of 1294 genes was tested (X^2 = 39.13, P=7e-07) and also when the 772 genes detected by using site models between *D. mojavensis* and *D. buzzatii* were tested (X^2 = 32.28, P=0.00001). In both cases there is a significant excess of genes in the X chromosome in comparison with the autosomes (57 and 47 genes respectively). On the other hand, there is a higher proportion of positively selected genes in the *D. buzzatii* branch located at chromosome 5 than expected by chance (X^2 = 6.69, P=0.01).

Using the RNAseq data we were able to determine the expression profile of all the 1,294 PCG under positive selection. A total of 1,213 (93.7%) of these genes are expressed in at least one developmental stage. A comparison of expression level and breadth between putative positively and non-positively selected genes revealed that genes showing evidence of positive selection are expressed at a lower level (X^2=84.96, P<2e-16) and in less stages (X^2=26.99, P<2e-6) than the rest.

Orphan genes

To detect orphan genes we blasted the aminoacid sequences encoded by 9114 *D. buzzatii* genes with *D. mojavensis* 1:1 orthologs against all proteins from the 11 Drosophila protein database available in Flybase (that correspond to the 12 Drosophila genomes other than *D. mojavensis*). We found 117 proteins that showed no similarity with any predicted Drosophila protein (cutoff value of 1e-05) and were considered to be encoded by putative orphan genes. We focused on the evolutionary dynamics of these orphan genes by studying their properties in comparison to the remaining 8,997 1:1 orthologs (Figure 7). We observed that median dn of orphan genes was significantly higher than that of non-orphan genes (dn_{orphan} = 0.1291; $dn_{non\text{-}orphan}$ = 0.0341; W=846254, P < 2.2e-16) and the same pattern was observed for ω (ω_{orphan} = 0.4253, $\omega_{no\ orphan}$ = 0.0887, W=951117, P < 2.2e-16). However median ds of orphan genes is somewhat lower than that for the rest of genes (ds_{orphan}=0.3000, $ds_{no\ orphan}$ = 0.4056, W=406799, P=2.4e-05).

We found 19 out of the 117 orphan genes in the list of positively selected genes detected in the *D. buzzatii-D. mojavensis* comparison (see above). This proportion (16.3%) was significantly higher than that found in non-orphan 1:1 orthologs (753/8997 = 8.4%), which indicates an association between gene lineage specificity and positive selection (Fischer exact test, two tailed, P < 0.0001). The 19 orphan genes included in the positively selected candidate group are not associated to any GO category. As a matter of fact, information about protein domains was found for only two of these genes (GYR and YLP motifs in both cases: FBgn10143727 and FBgn0143728). We also compared the protein length between orphan and non-orphan gene products. Our results showed that orphan genes are shorter (W=68825.5, P<2.2e-16) and have less exons than non lineage specific genes (W=201068, P<2.2e-16). Orphan genes seem to be randomly distributed among chromosomes.

RNAseq data allowed us to test for expression of orphan genes. From the 117 gene candidates, 82 (70%) are expressed at least in one of the five analyzed developmental stages. A comparison of the expression profile between orphan and the rest of 1:1 orthologous genes showed that the expression breadth of orphans is different to that of non-orphans (X^2=101.4, P=0). Thus, the orphan set contains more exclusive-stage expressed genes (29) and less constitutive genes (16) than non-orphan genes and mean expression breadth is 2.56 for orphans versus 3.94 for non-orphans.

DISCUSSION

The *D. buzzatii* genome

Drosophila is a leading model for comparative genomics, with 24 genomes of different species already sequenced (see Introduction). However only five of these species belong to the Drosophila subgenus, the most numerous one, and only one, *D. mojavensis*, belongs to the large repleta species group and is cactophilic. Here we sequenced the genome and transcriptome of *D. buzzatii*, another cactophilic member of the repleta group, to investigate the genomic basis of adaptation to this distinct ecological niche. Using different sequencing platforms (454 Roche, Illumina and Sanger) and a three-stage de novo assembly, we generated a high quality genome sequence contained in 826 scaffolds >3 kb (Freeze 1). A large portion (>90%) of the genome is represented by 158 scaffolds with a minimum size of 160 kb that have been assigned, ordered and oriented in the six chromosomes of the *D. buzzatii* karyotype. As expected the assembly is best for chromosome 2 (because of the use of Sanger generated BAC-end sequences) and worst for chromosome X (because of the ¾ representation of this chromosome in adults of both sexes). The quality of our Freeze 1 assembly compares favorably with the assembly generated by us using only Illumina reads and the SOAPdenovo assembler, and with those of other Drosophila genomes generated using second-generation sequencing platforms (Zhou and Bachtrog 2012; Zhou et al. 2012; Ometto et al. 2013; Fonseca et al. 2013) although does not reach the quality of the 12 Drosophila genomes generated using Sanger only (Drosophila 12 Genomes Consortium et al. 2007).

D. buzzatii is a subcosmopolitan species that has been able to colonize four of the six major biogeographical regions (David and Tsacas 1980). Only two other repleta group species (*D. repleta* and *D. hydei*) have reached such widespread distribution.

Invasive species are likely to share special genetic traits that enhance their colonizing ability (Parsons 1983; Lee 2002). From an ecological point of view we would expect colonizing species to be r-strategists with a short developmental time (Lewontin 1965). Because there is a correlation between developmental time and genome size (Gregory and Johnston 2008), they are also expected to have a small genome size (Lavergne et al. 2010). The genome size of *D. buzzatii* was estimated in our assembly as 161 Mb and by cytological techniques as 153 Mb, ~20% smaller than the *D. mojavensis* genome. The genome size of a second *D. buzzatii* strain, estimated by cytological techniques, is even smaller, 146 Mb. However, the relationship between genome size and colonizing ability does not hold in the Drosophila genus at large. Although colonizing species such as *D. melanogaster* and *D. simulans* have relatively small genomes, specialist species with a narrow distribution such as *D. sechelia* and *D. erecta* also have small genomes. On the other hand, *D. ananassae*, *D. malerkotliana*, *D. suzuki*, *D. virilis*, and *Zaprionus indianus* are also colonizing Drosophila species but have relatively large genomes. Further, there seem to be little difference in genome size between original and colonized populations within species (Nardon et al. 2005; Drosophila 12 Genomes Consortium et al. 2007). Seemingly, other factors such as historical or chance events, niche dispersion, genetic variability or behavioral shifts are more significant than genome size in determining the current distribution of colonizing species.

Repeat content

The TE content in *D. buzzatii* was estimated as 8% (Table 2), a relatively low value compared with that of *D. mojavensis*, 10-14% (Ometto et al. 2013, Rius et al. in preparation). Because genome size is positively correlated with the contribution of TEs (Kidwell 2002; Feschotte and Pritham 2007), these data agree well with the smaller genome size of *D. buzzatii* (see above). However, copy number and coverage estimated in *D. buzzatii* (Table 2) must be taken cautiously. Coverage is surely underestimated due to the difficulties in assembling repeats, in particular with short sequence reads, whereas the number of copies may be overestimated due to copy fragmentation (Rius et al. in preparation).

We identified the pBuM189 satDNA as the most abundant tandem repeat of *D. buzzatii*. Previous *in situ* hybridization experiments revealed that pBuM189 copies are located in the centromeric region of all chromosomes, except chromosome X (Kuhn et al. 2008). Thus pBuM189 satellite is likely the main component of the *D. buzzatii* centromere. Interestingly, a pBuM189 homologous sequence has recently been identified as the most abundant tandem repeat of *D. mojavensis* (Melters et al. 2013). Although the chromosome location in *D. mojavensis* has not been determined, the persistence of pBuM189 as the major satellite DNA in *D. buzzatii* and *D. mojavensis* may reflect a possible role for these sequences in centromere function (Ugarković 2009).

Chromosome evolution

The chromosomal evolution of *D. buzzatii* and *D. mojavensis* has been previously studied by comparing the banding pattern of the salivary gland chromosomes (Ruiz et al. 1990; Ruiz and Wasserman 1993). *D. buzzatii* has few fixed inversions (2m, 2n, 2z^7, 5g) when compared with the ancestor of the repleta group. In contrast, *D. mojavensis* showed ten fixed inversions (Xe, 2c, 2f, 2g, 2h, 2q, 2r, 2s, 3a, 3d), five of them (Xe, 2q, 2r, 2s and 3d) exclusive to *D. mojavensis* whereas the rest shared by other cactophilic Drosophila (Guillén and Ruiz 2012). Thus the *D. mojavensis* lineage appeared as a derived lineage with a relatively high rate of rearrangement fixation. Here we compared the organization of both genomes corroborating all known inversions in chromosomes X, 2, 4 and 5. In *D. mojavensis* chromosome 3, however, we found six inversions fixed instead of the two expected. One of the four additional inversions is the polymorphic inversions 3f^2 (Ruiz et al. 1990). This inversion has previously been found segregating in Baja California and Sonora (Mexico) and is seemingly fixed in the strain of Santa Catalina Island (California) that was used to generate the *D. mojavensis* genome sequence (Drosophila 12 Genomes Consortium et al. 2007). Previously, the Santa Catalina Island population was thought to have the standard (ancestral) arrangements in all chromosomes, like the populations in Southern California and Arizona (Ruiz et al. 1990; Etges et al. 1999). The presence of inversion 3f^2 in Santa Catalina Island is significant because it indicates that the flies that colonized this island came from Baja California and are derived instead of ancestral with regard to the rest of *D. mojavensis* populations. The other three additional chromosome 3 inversions are fixed in the *D. mojavensis* lineage and emphasize its rapid chromosomal evolution. Guillén and Ruiz (2012) analyzed the breakpoint of all chromosome 2 inversions fixed in *D. mojavensis* and concluded that the numerous gene alterations at the breakpoints with putative adaptive consequences directly point to natural selection as the cause of *D. mojavensis* rapid chromosomal evolution. The five fixed chromosome 3 inversions provide an opportunity for further testing this hypothesis.

Drosophila has a partially disassembled Hox gene complex (HOM-C) with at least three major splits, five microinversions and six gene transpositions fixed in diverse species of the genus (Negre et al. 2005; Negre and Ruiz 2007). Here we localized and annotated the eight Hox genes present in the *D. buzzatii* genome, corroborating information for three of them reported previously (Negre et al. 2005). The organization of the *D. buzzatii* HOM-C is similar to that observed in *D. mojavensis* (Negre and Ruiz 2007). Thus no rearrangements were found in *D. buzzatii* in addition to those already reported.

Gene content and developmental transcriptome

A total of 13,657 protein-coding genes were annotated in *D. buzzatii* genome using *ab initio* and homology-based predictors (Annotation Release 1). This number is lower than the number of PCG predicted in *D. mojavensis* (14,595, Release 1.3) but quite close to the number annotated in *D. melanogaster* (13955, Release 5.56), one of the best known eukaryotic genomes (The FlyBase Consortium 2002). The combination of *ab initio* and homology-based algorithms attempted to reduce the high false-positive rate associated to *de novo* gene prediction (Wang et al. 2003; Misawa and Kikuno 2010) as well as to avoid the propagation of wrong predicted gene models in close species used as references (Poptsova and Gogarten 2010). Regardless the efforts to obtain a proper set of reliable PCG models, subsequent quality filters were performed in order to avoid artifacts and biased results in posterior analyses.

We analyzed gene expression through the development by sequencing poly(A)+ RNA samples from five life-stages (embryos, larvae, pupae, adult males and adult females). We found evidence of expression for approximately 92.4% (12614) of the 13,657 PCG models predicted in Annotation Release 1. PCG models that did not show transcriptional evidence can be expressed at very low level (FPKM < 1) in the tissues analyzed here but at a higher level in other tissues or times, can be inducible (expressed only under particular environmental conditions; Weake and Workman 2010) or can be false positives (Wang et al. 2003). However, because we used a combination of different annotation methods to reduce the proportion of false-positives, we expect this proportion to be very small. On the other hand, we found expression evidence for 2959 genes not present in the Annotation Release 1. These genes are likely ncRNA genes although we cannot discard that some of them might be false negatives, i.e. genes that went undetected by our annotation methods perhaps because they contain small open reading frames (Ladoukakis et al. 2011). One observation supporting that most of them are in fact ncRNA genes is that their expression breadth is quite different from that of

PCG and a high fraction of them are stage-exclusive genes. In most Drosophila species, with limited analyses of the transcriptome (Celniker et al. 2009), few ncRNA genes have been annotated. For instance, in *D. mojavensis* 30 snRNA, 139 snoRNA, 71 miRNA and 3 miscellaneous ncRNA genes have been identified (Release 3.1, FlyBase). By contrast, in *D. melanogaster* that has a very well annotated genome, 31 snRNA, 288 snoRNA, 238 miRNA and 2096 miscellaneous ncRNA genes have been found (Release 5.56, FlyBase). Thus, the number of ncRNA found in *D. buzzatii* is significantly higher than that of *D. mojavensis* but much close to that of *D. melanogaster*.

D. buzzatii is the second Drosophila species whose-genome expression profile has been analyzed throughout its life cycle and the pattern is similar to that of *D. melanogaster* (Graveley et al. 2011). The number of expressed genes (PCG + ncRNA) increases through the life cycle with a maximum of 12171 in male adults. In addition, we observed a clear sex-biased expression in adults. This pattern cannot be attributed to other stages as we did not have sex differentiation in the rest of life cycle samples. Previous studies have attributed this sex differential gene expression mainly to the germ cells, indicating that the differences between ovary and testis are comparable to that between germ and somatic cells (Parisi et al. 2004; Graveley et al. 2011).

Patterns of divergence

Genome-wide gene molecular evolution has been previously analyzed in the 12 Drosophila genomes with special emphasis on the *melanogaster* species group of the Sophophora subgenus (Drosophila 12 Genomes Consortium et al. 2007; Heger and Ponting 2007; Larracuente et al. 2008). In addition, detailed analyses of genome-wide divergence and polymorphism patterns have been carried out using many *D. melanogaster* lines (Mackay et al. 2012; Langley et al. 2012). Here we focused on the two cactophilic species, *D. buzzatii* and *D. mojavensis*, to look for patterns of divergence. We did not include paralogs in our analysis because approaches for automating their detection yield sub-standard quality output. In addition, we filtered single copy orthologous using several criteria (Materials and Methods) to retain a set of 9017 high-quality reliable single-copy orthologs. We found expression evidence for the vast majority of them (94.7%) in our transcriptome analysis. In addition they were mapped to chromosomes and had complete values for seven genomic variables. Therefore, we used this PCG set for investigating patterns of divergence. The median estimate for the ratio ω = dn/ds was 0.0895, a similar value to that estimated in the *D. mojavensis* branch using a significantly lower number of orthologs (Heger and Ponting 2007).

Firstly, we tested for the effect of the type of chromosome (X vs autosomes) because X chromosome has been predicted to evolve at a faster rate (Charlesworth et al. 1987). We find that X-linked genes showed higher divergence rates (dn, ds and ⍰) than autosomal genes (Table 4 and 5), a pattern consistent with previous observations in the *D. melanogaster* and *D. simulans* lineages (Mackay et al. 2012; Langley et al. 2012; Campos et al. 2014) and other lineages (Meisel and Connallon 2013). In addition, we found a significant excess of genes under positive selection on the X, pointing to a faster rate of adaptive evolution (see above). The faster rate of adaptive evolution of chromosome X may be due to two reasons: (i) Exposure of recessive or partially recessive favorable X-linked mutations to selection in hemyzygous males (Charlesworth et al. 1987; Meisel and Connallon 2013); (ii) Higher effective recombination rate that reduces Hill-Robertson interference (see below); because males are hemyzygous and do not recombine, effective recombination rate on the X chromosome is 2/3 the recombination rate in females (against ½ in the autosomes). In a thorough analysis of the two hypotheses, Campos et al. (2014) concluded that the dominance level of favorable mutations is the chief factor although recombination and hitchhiking may play some role.

The faster-X pattern for synonymous sites does not conform with the expectation of stronger codon usage bias reported in other lineages (Campos et al. 2012; Meisel and Connallon 2013; Campos et al. 2014). This observation could be consistent with the hypothesis that the mutation rate associated to X-linked genes is greater than that of autosomes (Begun et al. 2007; Meisel et al. 2012; Hu et al. 2013). The dosage compensation effect resulting in the hypertranscription of X-linked genes in males (Conrad and Akhtar 2012) could lead to higher mutation rates.

We also tested for an effect of recombination on rates on divergence. The efficacy of selection acting simultaneously at linked sites is expected to be reduced in regions of low recombination. This is so because, due to linkage disequilibrium, selection at one locus will interfere with selection at linked loci (Hill and Robertson 1966). This interference may be caused by selective sweeps of beneficial mutations spreading through the population to fixation, or by the pervasive elimination of deleterious mutations, i.e. background selection (Charlesworth 1994). Interference between weakly selected mutations is expected to increase that rate of interspecific divergence (McVean and Charlesworth 1999). Because detailed recombination estimates for *D. buzzatii* or *D. mojavensis* chromosomes are not available (Schafer et al. 1993; Staten et al. 2004) and genome-wide recombination varies substantially among Drosophila species (True et al. 1996; Cáceres et al. 1999), we used a rather conservative approach. We compared the dot chromosome with the rest of autosomes and also pericentromeric regions of all chromosomes (including the entire dot) against the rest of chromosome regions. The *D. buzzatii* chromosome 6 (dot) and the pericentromeric regions likely have a reduced or nearly null rate of recombination, as in *D. melanogaster* (Arguello et al. 2010; Comeron et al. 2012). The accumulation of TE insertions in both the dot chromosome and pericentromeric regions of *D. melanogaster* (Kaminker et al.

2002; Slawson et al. 2006) and *D. buzzatii* (Casals et al. 2006) is an indirect support for their reduced recombination rate.

We found a significantly increased rate of divergence (dn, ds and ⍰) in the dot chromosome than in the rest of autosomes (Table 4). A similar pattern, although less marked, is found when we consider the reduced-recombination pericentromeric regions of all autosomes, yet only dn and ds are statistically significant (Table 5). These observations agree well with previous observations in *Drosophila* (Haddrill et al. 2007; Larracuente et al. 2008; Leung et al. 2010; Arguello et al. 2010; Campos et al. 2012, 2014). Besides, we find a lower number of genes under positive selection in non-recombining regions. Thus our results support the hypothesis that accelerated rate of evolution is not due to beneficial mutations but to the fixation of slightly or mildly deleterious mutations, a notion supported by the measurements of divergence and polymorphism in several studies.

Thirdly, we tested for an effect on divergence of chromosomal inversions. Inversions segregating in natural populations reduce recombination in the inverted segment in heterokaryotypes yet not in homokaryotypes (Navarro et al. 1997). Inversions than have been fixed in a lineage have all passed through a more or less long phase of polymorphism. Thus historical recombination rates in rearranged chromosomal regions must be reduced to some extent in comparison with collinear chromosomal regions. This reduced recombination rate in regions rearranged by chromosomal inversions might imply a relaxation of the efficacy of selection due to Hill-Robertson interference and thus a higher fixation rate for slightly or mildly deleterious mutations (see above). On the other hand, inversions might facilitate speciation by protecting population specific adaptations from recombination (Rieseberg 2001; Navarro and Barton 2003). This hypothesis predicts an accumulation of positively selected alleles in rearranged chromosomal regions in comparison with collinear chromosomal regions.

Natural populations of *D. buzzatii* and *D. mojavensis* are polymorphic for inversions in chromosomes 2 and 4 (Hasson et al. 1995) and chromosomes 2 and 3 (Ruiz et al. 1990; Etges et al. 1999), respectively. The reference *D. buzzatti* genome comes from a line standard for all chromosomes (st-1) but the *D. mojavensis* genome was generated from a line (Santa Catalina Island) with the polymorphic inversion $3f^2$ fixed (see above). In addition, both species differ by 10 and 5 inversions fixed in chromosome 2 and 3 while only one inversion is fixed in each of chromosomes X and 5. We compared the divergence parameters between the rearranged autosomes 2 and 3 and the nearly collinear chromosomes 4 and 5. Although the pattern resembles that of non-recombining regions, the increases of dn and ds are modest and only the latter is significant (Table 5). When all rearranged chromosomal regions were considered together in a multiple linear model, ds increase although slight was again statistically significant (Table 5). Rearranged chromosomal regions did not show an increased number of positively selected genes (as a matter of fact they showed a slightly and nonsignificant lower number). Although rearranged chromosomal regions may contain both positively selected genes and mildly deleterious mutations, we consider that overall their molecular evolution pattern resembles more that of reduced-recombination regions with relaxed selective constraints than that of the X chromosome with its faster adaptive rate. It is perhaps worth recalling that chromosome X, with a significant excess of positively selected genes, has few fixed chromosomal inversions in comparison with autosomes 2 and 3.

Finally our results indicate that divergence rates are simultaneously influenced by multiple genomic factors (Table 5). The negative correlation between breadth expression and rates of protein evolution indicates that genes that are expressed in more life stages do not evolve as fast as genes with higher bias expression. In Drosophila it has been previously reported that narrowly expressed genes evolve faster as showed by higher rates of divergence (Drosophila 12 Genomes Consortium et al. 2007;

Larracuente et al. 2008). Thus, it seems that genes that are expressed in more stages tend to evolve slowly due to the high evolutionary constraint derived from gene pleiotropy (Fischer 1930; Larracuente et al. 2008; Singh et al. 2009). According to our results expression breadth, rather than expression level, is the major contributor to gene evolution.

We also show that exon number is negatively correlated with dn, ds and ω. This observation is consistent with the influence of the sequences responsible for a correct introns excision (Exonic splite site enhancers, ESEs) on evolutionary constrainment (Warnecke et al. 2008; Larracuente et al. 2008; Cáceres and Hurst 2013). Furthermore, we observe that protein length is positively correlated with ds (Table 5). The degree of codon bias is positively correlated with the rate of synonymous substitutions. In turn, we expect a significant positive correlation between the expression level of a gene and its degree of codon bias (Bulmer 1991; Plotkin and Kudla 2011). Accordingly, the correlation between ds and protein length could be a consequence of a smaller coding sequence size of highly expressed genes. We tested for a correlation between these two parameters and corroborated that highly expressed genes encode for shorter proteins (Pearson test, $P < 2.2$ e-16). Comeron et al. (1999) hypothesized with the possibility that highly expressed genes shortening their length by eliminating nonessential amino acids from their sequence supporting a length-dependent selection coefficient model (LdSC) affected by translational efficiency, i.e. the shorter the coding sequence, the stronger the relative effects in translational efficiency.

Genes under positive selection and orphan genes

We used *D. buzzatii* and *D. mojavensis* for detecting genes under positive selection using site models (SM). In addition, we used four species of the Drosophila subgenus (Figure 1) to find genes under positive selection using branch-site models

(BSM). We restricted the analysis to this subset of the Drosophila phylogeny to avoid the saturation of synonymous substitutions expected with phylogenetically very distant species (Bergman et al. 2002; Larracuente et al. 2008) and also because these are the genomes with the highest quality available (Schneider et al. 2009). We considered positively selected genes those with statistical evidence for a subset of codons where replacement mutations were fixed faster than mutation at silent sites (Yang et al. 2000; Yang 2007). A total of 1294 genes positively selected were detected both SM and BSM, which represents ~14% of the total set of 1:1 orthologs accurately detected between *D. mojavensis* and *D. buzzatii*. The number of positive selected genes is likely underestimated because (i) we are not able to detect orthology relationships between genes that evolve too fast (Bierne and Eyre-Walker 2004) and (ii) only orthologs 1:1 are included in the analyses.

Branch-site models allowed us to identify positively selected genes in the three targeted lineages (*D. buzzatii*, *D. mojavensis* and cactophilic branch). A GO enrichment analysis was performed on the resulting positively selected genes dataset in order to identify good candidates for environment adaptation given the ecological properties of both cactophilic species (Table 6). The most important point in our results is that genes that evolved under positive selection in *D. mojavensis* branch are enriched in heterocycle catabolic processes, which involve functions strongly linked to the characteristic adaptation of *D. mojavensis* to columnar cacti, which are plants showing particularly large quantities of heterocyclic compounds (see Introduction). We suggested that there exists a causal link between adaptation to columnar cacti and the molecular evolution of these candidate genes. Even the reference genome of *D. mojavensis* used herein (Drosophila 12 Genomes Consortium et al. 2007) was obtained by sequencing individuals from Catalina Island (the only one of the four subpopulations that inhabit cactus of Opuntia genus), two evidences suggest that the common ancestor of the four subpopulations (Figure 1) adapted to columnar cacti rather than Opuntia.

First, the presence of the inversion $3f^2$ in the sequenced strain from Catalina Island indicates that the flies that colonized this region came from populations that feed from columnar cacti in Baja California, where the inversion is segregating. And second, the study of the transcriptional dynamics along the four *D. mojavensis* subpopulations revealed that the minor gene expression differences are showed between individuals from Catalina Island and Baja California (Matzkin and Markow 2013).

Orphan genes are genes that have no homologues in any other known lineage. It has been reported that orphans or also called taxonomically restricted genes, play an important role in adaptive evolution on multiple species (Domazet-Lošo and Tautz 2003; Khalturin et al. 2009). The detection of orphan genes is highly dependent on the availability of sequenced and well annotated genomes of closely related species, consequently the total number of lineage specific genes tend to be overestimated (Khalturin et al. 2009). We were as conservative as possible when filtering data to detect the final dataset of 117 orphan genes, trying to optimize the fidelity of orphans identification. For that reason, some particular orphan genes (including in-paralogs not considered in 1:1 orthologs dataset) are missing and we are likely underestimating the abundance of orphans.

Even though previous studies have focused on the evolution of orphan genes in different species, little is known about the evolution of orphans along short phylogenetics distances as that separating cactophilic species.

We observed that orphan genes clearly show a different molecular evolution pattern compared to that of older conserved genes. Our results reveal that they exhibit a higher rate of dn, indicating that the number of fixated adaptive mutations is greater or they have fixated more deleterious mutations by hitchhiking. However, since the number of positive selected genes within orphan genes dataset is much higher than expected by chance, we assume that they experience adaptive evolution more frequently (Cai and Petrov 2010; Palmieri et al. 2014). Orphans also showed a lower rate of ds suggesting a higher codon usage efficacy, which has been evidenced in recent studies focused on Drosophila orphan genes (Palmieri et al. 2014). Orphans also have less exons and encode shorter proteins than non orphans. This observation has been reported in multiple eukaryotic organisms like yeasts (Carvunis et al. 2012), fruitflies (Domazet-Lošo and Tautz 2003) and primates (Cai and Petrov 2010), and it is evidencing a positive correlation between protein length and sequence conservation (Lipman et al. 2002) (see above). We did not find expression support for all the orphan genes detected. This is indicated us that either orphans are more tissue-stage specific than non-orphans or we are actually detecting spurious CDSs not expressed. However, given the divergence rate pattern of orphan's dataset, evidencing positive selection, the first explanation is the most plausible. Collectively, all these results are evidencing that orphans evolve faster than older genes, experiencing lower levels of purifying selection and higher rates of adaptive evolution.

It has been widely reported that genes that evolve faster show lower expression levels than older genes on average (Cai and Petrov 2010; Tautz and Domazet-Lošo 2011). Here we observe that orphan genes that are being transcribed are less expressed than non-orphans (Kruskal test, X^2 = 9.370, P=0.0022). One of the proposed hypothesis to explain these observations is that genes that are more conserved are indeed implicated in more functions (Pál et al. 2006; Tautz and Domazet-Lošo 2011).

Different studies have demonstrated that newer genes are more likely to have a stage-specific expression than older genes. Here we show that the number of stage-specific expressed orphans is significantly higher than that of older genes. It has been proposed that newer genes tend to be more developmentally regulated than conserved genes. This means that they contribute most to the ontogenic differentiation between taxa (Tautz and Domazet-Lošo 2011). In *D. buzzatii* the vast majority of stage-specific orphan genes are expressed in larvae (15/29), indicating that expression of younger genes is mostly related to stages in which *D. buzzatii* and *D. mojavensis* lineages most diverge from each other.

MATERIALS AND METHODS

See Supplemental Material.

REFERENCES

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, et al. 2000. The genome sequence of Drosophila melanogaster. *Science* **287**: 2185–2195.

Amemiya CT, Alföldi J, Lee AP, Fan S, Philippe H, MacCallum I, Braasch I, Manousaki T, Schneider I, Rohner N, et al. 2013. The African coelacanth genome provides insights into tetrapod evolution. *Nature* **496**: 311–316.

Anisimova M, Liberles DA. 2007. The quest for natural selection in the age of comparative genomics. *Heredity* **99**: 567–579.

Arguello JR, Zhang Y, Kado T, Fan C, Zhao R, Innan H, Wang W, Long M. 2010. Recombination yet inefficient selection along the Drosophila melanogaster subgroup's fourth chromosome. *Mol Biol Evol* **27**: 848–861.

Baker M. 2012. De novo genome assembly: what every biologist should know. *Nat Methods* **9**: 333–337.

Barker JSF, Starmer WT. 1982. *The Cactus-Yeast-Drosophila Model System*. Academic Press, Sidney, Australia.

Begun DJ, Holloway AK, Stevens K, Hillier LW, Poh Y-P, Hahn MW, Nista PM, Jones CD, Kern AD, Dewey CN, et al. 2007. Population Genomics: Whole-Genome Analysis of Polymorphism and Divergence in Drosophila simulans. *PLoS Biol* **5**: e310.

Bergman CM, Pfeiffer BD, Rincón-Limas DE, Hoskins RA, Gnirke A, Mungall CJ, Wang AM, Kronmiller B, Pacleb J, Park S, et al. 2002. Assessing the impact of comparative genomic sequence data on the functional annotation of the Drosophila genome. *Genome Biol* **3**: research0086.

Bierne N, Eyre-Walker A. 2004. The Genomic Rate of Adaptive Amino Acid Substitution in Drosophila. *Mol Biol Evol* **21**: 1350–1360.

Bulmer M. 1991. The selection-mutation-drift theory of synonymous codon usage. *Genetics* **129**: 897–907.

Cáceres EF, Hurst LD. 2013. The evolution, impact and properties of exonic splice enhancers. *Genome Biol* **14**: R143.

Cáceres M, Barbadilla A, Ruiz A. 1999. Recombination rate predicts inversion size in Diptera. *Genetics* **153**: 251–259.

Cai JJ, Petrov DA. 2010. Relaxed purifying selection and possibly high rate of adaptation in primate lineage-specific genes. *Genome Biol Evol* **2**: 393–409.

Campos JL, Halligan DL, Haddrill PR, Charlesworth B. 2014. The Relation between Recombination Rate and Patterns of Molecular Evolution and Variation in Drosophila melanogaster. *Mol Biol Evol* **31**: 1010–1028.

Campos JL, Zeng K, Parker DJ, Charlesworth B, Haddrill PR. 2012. Codon usage bias and effective population sizes on the X chromosome versus the autosomes in Drosophila melanogaster. *Mol Biol Evol* **4**: 278–288.

Carson HL, Wasserman M. 1965. A widespread chromosomal polymorphism in a widespread species, Drosophila buzzatii. *Am Nat* **99**: 111–115.

Carvunis A-R, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, Charloteaux B, Hidalgo CA, Barbette J, Santhanam B, et al. 2012. Proto-genes and de novo gene birth. *Nature* **487**: 370–374.

Casals F, González J, Ruiz A. 2006. Abundance and chromosomal distribution of six Drosophila buzzatii transposons: BuT1, BuT2, BuT3, BuT4, BuT5, and BuT6. *Chromosoma* **115**: 403–412.

Celniker SE, Dillon LAL, Gerstein MB, Gunsalus KC, Henikoff S, Karpen GH, Kellis M, Lai EC, Lieb JD, MacAlpine DM, et al. 2009. Unlocking the secrets of the genome. *Nature* **459**: 927–930.

Charlesworth B. 1994. The effect of background selection against deleterious mutations on weakly selected, linked variants. *Genet Res* **63**: 213–227.

Charlesworth B, Coyne JA, Barton NH. 1987. The Relative Rates of Evolution of Sex Chromosomes and Autosomes. *Am Nat* **130**: 113–46.

Comeron JM, Ratnappan R, Bailin S. 2012. The Many Landscapes of Recombination in Drosophila melanogaster. *PLoS Genet* **8**: e1002905.

Conrad T, Akhtar A. 2012. Dosage compensation in Drosophila melanogaster: epigenetic fine-tuning of chromosome-wide transcription. *Nat Rev Genet* **13**: 123–134.

David J, Tsacas L. 1980. Cosmopolitan, subcosmopolitan and widespread species: different strategies within the Drosophilid family (Diptera). *C R Soc Biogéogr* **57**: 11–26.

Delcher AL, Salzberg SL, Phillippy AM. 2003. Using MUMmer to identify similar regions in large sequence sets. *Curr Protoc Bioinforma Ed Board Andreas Baxevanis Al* **Chapter 10**: Unit 10.3.

Domazet-Lošo T, Tautz D. 2003. An evolutionary analysis of orphan genes in Drosophila. *Genome Res* **13**: 2213–2219.

Drosophila 12 Genomes Consortium, Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA, Kaufman TC, Kellis M, Gelbart W, et al. 2007. Evolution of genes and genomes on the Drosophila phylogeny. *Nature* **450**: 203–218.

Etges WJ, Johnson WR, Duncan GA, Huckins G, Heed WB. 1999. Ecological Genetics of Cactophilic Drosophila. In *Ecology of Sonoran Desert plants and plant communities*, pp. 164–214, University of Arizona Press.

Fellows DP, Heed WB. 1972. Factors Affecting Host Plant Selection in Desert-Adapted Cactiphilic Drosophila. *Ecology* **53**: 850–858.

Feschotte C, Pritham EJ. 2007. DNA Transposons and the Evolution of Eukaryotic Genomes. *Annu Rev Genet* **41**: 331–368.

Fischer RA. 1930. *The Genetical Theory of Natural Selection*. Oxford University Press.

Fogleman JC, Armstrong L. 1989. Ecological aspects of cactus triterpene glycosides I. Their effect on fitness components ofDrosophila mojavensis. *J Chem Ecol* **15**: 663–676.

Fogleman JC, Danielson PB. 2001. Chemical Interactions in the Cactus-Microorganism-Drosophila Model System of the Sonoran Desert1. *Am Zool* **41**: 877–889.

Fogleman JC, Kircher HW. 1986. Differential effects of fatty acid chain length on the viability of two species of cactophilic Drosophila. *Comp Biochem Physiol A Physiol* **83**: 761–764.

Fonseca NA, Morales-Hojas R, Reis M, Rocha H, Vieira CP, Nolte V, Schlötterer C, Vieira J. 2013. Drosophila americana as a model species for comparative studies on the molecular basis of phenotypic variation. *Genome Biol Evol* **5**: 661–679.

Fontdevila A, Ruiz A, Alonso G, Ocana J. 1981. Evolutionary History of Drosophila buzzatii. I. Natural Chromosomal Polymorphism in Colonized Populations of the Old World. *Evolution* **35**: 148.

Gonzalez 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**: 885–889.

Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin JM, Yang L, Artieri CG, van Baren MJ, Boley N, Booth BW, et al. 2011. The developmental transcriptome of Drosophila melanogaster. *Nature* **471**: 473–479.

Gregory TR, Johnston JS. 2008. Genome size diversity in the family Drosophilidae. *Heredity* **101**: 228–238.

Guillén Y, Ruiz A. 2012. Gene alterations at Drosophila inversion breakpoints provide prima facie evidence for natural selection as an explanation for rapid chromosomal evolution. *BMC Genomics* **13**: 53.

Haddrill PR, Halligan DL, Tomaras D, Charlesworth B. 2007. Reduced efficacy of selection in regions of the Drosophila genome that lack crossing over. *Genome Biol* **8**: R18.

Hasson E, Naveira H, Fontdevila A. 1992. The breeding sites of Argentinian cactophilic species of the Drosophila mulleri complex (subgenus Drosophila-repleta group). *Rev Chilena de Hist Nat* **65**: 319–326.

Hasson E, Rodríguez C, Fanara JJ, Naveira H, Reig O, 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.

Heed WB. 1978. Ecology and Genetics of Sonoran Desert Drosophila. In *Ecological Genetics: The Interface* (ed. P.F. Brussard), *Proceedings in Life Sciences*, pp. 109–126, Springer New York.

Heed WB, Mangan RL. 1986. Community ecology of the Sonoran Desert Drosophila. In *The genetics and biology of Drosophila*, Vol. 3e of, Academic Press, London.

Heger A, Ponting CP. 2007. Evolutionary rate analyses of orthologs and paralogs from 12 Drosophila genomes. *Genome Res* **17**: 1837–1849.

Hill WG, Robertson A. 1966. The effect of linkage on limits to artificial selection. *Genet Res* **8**: 269–294.

Hu TT, Eisen MB, Thornton KR, Andolfatto P. 2013. A second-generation assembly of the Drosophila simulans genome provides new insights into patterns of lineage-specific divergence. *Genome Res* **23**: 89–98.

Huang DW, Sherman BT, Tan Q, Kir J, Liu D, Bryant D, Guo Y, Stephens R, Baseler MW, Lane HC, et al. 2007. DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Res* **35**: W169–W175.

Kaminker JS, Bergman CM, Kronmiller B, Carlson J, Svirskas R, Patel S, Frise E, Wheeler DA, Lewis SE, Rubin GM, et al. 2002. The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective. *Genome Biol* **3**: research0084.

Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TCG. 2009. More than just orphans: are taxonomically-restricted genes important in evolution? *Trends Genet* **25**: 404–413.

Kidwell MG. 2002. Transposable elements and the evolution of genome size in eukaryotes. *Genetica* **115**: 49–63.

Kircher HW. 1982. Chemical composition of cacti and its relationship to Sonoran Desert Drosophila. In *Ecological Genetics and Evolution: The Cactus-Yeast-Drosophila Model System*, pp. 143–158, Academic Press, Sydney, Australia.

Kircher HW, Heed WB, Russell JS, Grove J. 1967. Senita cactus alkaloids: their significance to Sonoran Desert ecology. *J Insect Physiol* **13**: 1869–1874.

Kuhn GCS, Sene FM, Moreira-Filho O, Schwarzacher T, Heslop-Harrison JS. 2008. Sequence analysis, chromosomal distribution and long-range organization show that rapid turnover of new and old pBuM satellite DNA repeats leads to different patterns of variation in seven species of the Drosophila buzzatii cluster. *Chromosome Res* **16**: 307–324.

Ladoukakis E, Pereira V, Magny EG, Eyre-Walker A, Couso JP. 2011. Hundreds of putatively functional small open reading frames in Drosophila. *Genome Biol* **12**: R118.

Lang M, Murat S, Clark AG, Gouppil G, Blais C, Matzkin LM, Guittard E, Yoshiyama-Yanagawa T, Kataoka H, Niwa R, et al. 2012. Mutations in the neverland gene turned Drosophila pachea into an obligate specialist species. *Science* **337**: 1658–1661.

Langley CH, Stevens K, Cardeno C, Lee YCG, Schrider DR, Pool JE, Langley SA, Suarez C, Corbett-Detig RB, Kolaczkowski B, et al. 2012. Genomic variation in natural populations of Drosophila melanogaster. *Genetics* **192**: 533–598.

Larracuente AM, Sackton TB, Greenberg AJ, Wong A, Singh ND, Sturgill D, Zhang Y, Oliver B, Clark AG. 2008. Evolution of protein-coding genes in Drosophila. *Trends Genet* **24**: 114–123.

Lavergne S, Muenke NJ, Molofsky J. 2010. Genome size reduction can trigger rapid phenotypic evolution in invasive plants. *Ann Bot* **105**: 109–116.

Lee CE. 2002. Evolutionary genetics of invasive species. *Trends Ecol Evol* **17**: 386–91.

Leung W, Shaffer CD, Cordonnier T, Wong J, Itano MS, Slawson Tempel EE, Kellmann E, Desruisseau DM, Cain C, Carrasquillo R, et al. 2010. Evolution of a distinct genomic domain in Drosophila: comparative analysis of the dot chromosome in Drosophila melanogaster and Drosophila virilis. *Genetics* **185**: 1519–1534.

Lewontin RC. 1965. Selection for colonizing ability. In *The genetics of colonizing species* (eds. H.G. Baker and Stebbins), Academic Press, New York.

Lipman DJ, Souvorov A, Koonin EV, Panchenko AR, Tatusova TA. 2002. The relationship of protein conservation and sequence length. *BMC Evol Biol* **2**: 20.

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, et al. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* **1**: 18.

Mackay TFC, Richards S, Stone EA, Barbadilla A, Ayroles JF, Zhu D, Casillas S, Han Y, Magwire MM, Cridland JM, et al. 2012. The Drosophila melanogaster Genetic Reference Panel. *Nature* **482**: 173–178.

Manfrin MH, Sene FM. 2006. Cactophilic Drosophila in South America: a model for evolutionary studies. *Genetica* **126**: 57–75.

Mardis ER. 2008. The impact of next-generation sequencing technology on genetics. *Trends Genet* **24**: 133–141.

Markow TA, O'Grady PM. 2007. Drosophila biology in the genomic age. *Genetics* **177**: 1269–1276.

Matzkin LM, Markow TA. 2013. Transcriptional differentiation across the four subspecies of drosopihla mojavensis. In *Speciation: Natural Processes, Genetics and Biodiversity*, Nova Scientific Publishers, New York.

Matzkin LM, Watts TD, Bitler BG, Machado CA, Markow TA. 2006. Functional genomics of cactus host shifts in Drosophila mojavensis. *Mol Ecol* **15**: 4635–4643.

McVean G a. T, Charlesworth B. 1999. A population genetic model for the evolution of synonymous codon usage: patterns and predictions. *Genet Res* **74**: 145–158.

Meisel RP, Connallon T. 2013. The faster-X effect: integrating theory and data. *Trends Genet TIG* **29**: 537–544.

Meisel RP, Malone JH, Clark AG. 2012. Faster-X Evolution of Gene Expression in Drosophila. *PLoS Genet* **8**.

Melters DP, Bradnam KR, Young HA, Telis N, May MR, Ruby JG, Sebra R, Peluso P, Eid J, Rank D, et al. 2013. Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biol* **14**: R10.

Misawa K, Kikuno RF. 2010. GeneWaltz--A new method for reducing the false positives of gene finding. *BioData Min* **3**: 6.

Nardon C, Deceliere G, Loevenbruck C, Weiss M, Vieira C, Biémont C. 2005. Is genome size influenced by colonization of new environments in dipteran species? *Mol Ecol* **14**: 869–878.

Natori S, Ikekawa N, Suzuki M. 1981. *Advances in natural products chemistry: extraction and isolation of biologically active compounds*. Kodansha ; Wiley, Tokyo; New York.

Navarro A, Barton NH. 2003. Chromosomal Speciation and Molecular Divergence--Accelerated Evolution in Rearranged Chromosomes. *Science* **300**: 321–324.

Navarro A, Betrán E, Barbadilla A, Ruiz A. 1997. Recombination and Gene Flux Caused by Gene Conversion and Crossing Over in Inversion Heterokaryotypes. *Genetics* **146**: 695–709.

Negre B, Casillas S, Suzanne M, Sánchez-Herrero E, Akam M, Nefedov M, Barbadilla A, de Jong P, Ruiz A. 2005. Conservation of regulatory sequences and gene expression patterns in the disintegrating Drosophila Hox gene complex. *Genome Res* **15**: 692–700.

Negre B, Ruiz A. 2007. HOM-C evolution in Drosophila: is there a need for Hox gene clustering? *Trends Genet* **23**: 55–59.

Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, Hubisz MJ, Fledel-Alon A, Tanenbaum DM, Civello D, White TJ, et al. 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol* **3**: e170.

Ometto L, Cestaro A, Ramasamy S, Grassi A, Revadi S, Siozios S, Moretto M, Fontana P, Varotto C, Pisani D, et al. 2013. Linking genomics and ecology to investigate the complex evolution of an invasive Drosophila pest. *Genome Biol Evol* **5**: 745–757.

Pál C, Papp B, Lercher MJ. 2006. An integrated view of protein evolution. *Nat Rev Genet* **7**: 337–348.

Palmieri N, Kosiol C, Schlötterer C. 2014. The life cycle of Drosophila orphan genes. *eLife* **3**: e01311.

Parisi M, Nuttall R, Edwards P, Minor J, Naiman D, Lü J, Doctolero M, Vainer M, Chan C, Malley J, et al. 2004. A survey of ovary-, testis-, and soma-biased gene expression in Drosophila melanogaster adults. *Genome Biol* **5**: R40.

Parsons P. 1983. *The Evolutionary Biology of Colonizing Species*. Cambridge University Press, New York.

Patterson JT, Stone WS. 1953. *Evolution in the Genus Drosophila*. MacMillan Co., New York.

Plotkin JB, Kudla G. 2011. Synonymous but not the same: the causes and consequences of codon bias. *Nat Rev Genet* **12**: 32–42.

Poptsova MS, Gogarten JP. 2010. Using comparative genome analysis to identify problems in annotated microbial genomes. *Microbiol Read Engl* **156**: 1909–1917.

Rieseberg LH. 2001. Chromosomal rearrangements and speciation. *Trends Ecol Evol* **16**: 351–358.

Rubin GM, Lewis EB. 2000. A Brief History of Drosophila's Contributions to Genome Research. *Science* **287**: 2216–2218.

Ruiz A, Cansian AM, Kuhn GC, Alves MA, Sene FM. 2000. The Drosophila serido speciation puzzle: putting new pieces together. *Genetica* **108**: 217–227.

Ruiz A, Heed WB. 1988. Host-Plant Specificity in the Cactophilic Drosophila mulleri Species Complex. *J Anim Ecol* **57**: 237–249.

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.

Ruiz A, Wasserman M. 1993. Evolutionary cytogenetics of the Drosophila buzzatii species complex. *Heredity* **70**: 582–596.

Ruiz-Ruano FJ, Ruiz-Estévez M, Rodríguez-Pérez J, López-Pino JL, Cabrero J, Camacho JPM. 2011. DNA amount of X and B chromosomes in the grasshoppers Eyprepocnemis plorans and Locusta migratoria. *Cytogenet Genome Res* **134**: 120–126.

Schafer DJ, Fredline DK, Knibb WR, Green MM, Barker JSF. 1993. Genetics and Linkage Mapping of Drosophila buzzatii. *J Hered* **84**: 188–194.

Schneider A, Souvorov A, Sabath N, Landan G, Gonnet GH, Graur D. 2009. Estimates of positive Darwinian selection are inflated by errors in sequencing, annotation, and alignment. *Genome Biol Evol* **1**: 114–118.

Shendure J, Ji H. 2008. Next-generation DNA sequencing. *Nat Biotechnol* **26**: 1135–1145.

Singh ND, Larracuente AM, Sackton TB, Clark AG. 2009. Comparative Genomics on the Drosophila Phylogenetic Tree. *Annu Rev Ecol Evol Syst* **40**: 459–480.

Slawson EE, Shaffer CD, Malone CD, Leung W, Kellmann E, Shevchek RB, Craig CA, Bloom SM, Bogenpohl J 2nd, Dee J, et al. 2006. Comparison of dot chromosome sequences from D. melanogaster and D. virilis reveals an enrichment of DNA transposon sequences in heterochromatic domains. *Genome Biol* **7**: R15.

Staten R, Schully SD, Noor MA. 2004. A microsatellite linkage map of Drosophila mojavensis. *BMC Genet* **5**: 12.

Tamura K, Subramanian S, Kumar S. 2004. Temporal patterns of fruit fly (Drosophila) evolution revealed by mutation clocks. *Mol Biol Evol* **21**: 36–44.

Tautz D, Domazet-Lošo T. 2011. The evolutionary origin of orphan genes. *Nat Rev Genet* **12**: 692–702.

Tesler G. 2002. GRIMM: genome rearrangements web server. *Bioinformatics* **18**: 492–493.

The FlyBase Consortium. 2002. The FlyBase database of the Drosophila genome projects and community literature. *Nucleic Acids Res* **30**: 106–108.

True JR, Mercer JM, Laurie CC. 1996. Differences in crossover frequency and distribution among three sibling species of Drosophila. *Genetics* **142**: 507–523.

Ugarković Đ. 2009. Centromere-Competent DNA: Structure and Evolution. In *Centromere* (ed. D. Ugarkovic), *Progress in Molecular and Subcellular Biology*, pp. 53–76, Springer Berlin Heidelberg.

Vilela CR. 1983. A revision of the Drosophila repleta species group (Diptera, Drosophilidae). *Revta Bras Ent* **27**: 1–114.

Wang J, Li S, Zhang Y, Zheng H, Xu Z, Ye J, Yu J, Wong GK-S. 2003. Vertebrate gene predictions and the problem of large genes. *Nat Rev Genet* **4**: 741–749.

Warnecke T, Parmley JL, Hurst LD. 2008. Finding exonic islands in a sea of non-coding sequence: splicing related constraints on protein composition and evolution are common in intron-rich genomes. *Genome Biol* **9**: R29.

Wasserman M. 1992. Cytological evolution of the Drosophila repleta species group. In *Drosophila inversion polymorphism*, pp. 455–552, CRC Press, Boca Raton, FL.

Wasserman M. 1982. Evolution of the repleta group. In *The genetics and biology of Drosophila*, Vol. 3b of, pp. 61–139, Academic Press, London.

Wong WSW, Yang Z, Goldman N, Nielsen R. 2004. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. *Genetics* **168**: 1041–1051.

Yang, Bielawski. 2000. Statistical methods for detecting molecular adaptation. *Trends Ecol Evol* **15**: 496–503.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**: 1586–1591.

Yang Z, Nielsen R, Goldman N, Pedersen AM. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155**: 431–449.

Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol* **22**: 2472–2479.

Zhou Q, Bachtrog D. 2012. Sex-specific adaptation drives early sex chromosome evolution in Drosophila. *Science* **337**: 341–345.

Zhou Q, Zhu H, Huang Q, Zhao L, Zhang G, Roy SW, Vicoso B, Xuan Z, Ruan J, Zhang Y, et al. 2012. Deciphering neo-sex and B chromosome evolution by the draft genome of Drosophila albomicans. *BMC Genomics* **13**: 109.

TABLES

Table 1. Summary of assembly statistics for the genome of *Drosophila buzzatii* (strain st-1).

Assembly	Freeze 1	SOAPdenovo
Number of scaffolds (>3kb)	826	10949
Coverage	~22x	~76x
Assembly size (bp)	161490851	144184967
Scaffold N50 index	30	2035
Scaffold N50 length (bp)	1380942	18900
Scaffold N90 index	158	7509
Scaffold N90 length (bp)	161757	5703
Contig N50 index	1895	2820
Contig N50 length (bp)	17678	3101

Table 2. Transposable element content of *D. buzzatii* genome (Freeze 1 assembly).

Order	Superfamily	Copy number	bp Masked	% Masked
LTR	Gypsy	7548	1541621	0.95
	BEL	1407	429740	0.27
	Copia	1102	304433	0.19
	ERVK	121	9900	0.01
	Total	**10178**	**2285694**	**1.42**
DIRS	DIRS	1	38	0.00
LINE	R1	7522	1312191	0.81
	Jockey	1953	450561	0.28
	CR1	770	384683	0.24
	L2	1938	180881	0.11
	I	140	74216	0.05
	Other LINE	61	13931	0.01
	RTE	17	6763	0.00
	L1	94	4878	0.00
	R4	23	1504	0.00
	R2	2	1491	0.00
	LOA	2	1175	0.00
	Total	**12522**	**2432274**	**1.50**
DNA-TIR	P	2471	669565	0.41
	hAT	2255	417862	0.26
	Tc1Mariner	1443	391936	0.24
	Transib	1917	273248	0.17
	Other DNA	690	113444	0.07
	MULE-MuDR	168	19955	0.01
	PiggyBac	36	18647	0.01
	Novosib	226	16909	0.01
	PIF-Harbinger	18	3803	0.00
	Sola	2	183	0.00
	Total	**8926**	**1925552**	**1.18**
Helitron	Helitron	16256	5153798	3.19
Maverick	Maverick	2455	161440	0.10
Unknown	Unknown	6263	943233	0.58
Total		**56901**	**12902029**	**7.99**

Table 3. Satellite DNAs identified in the *D. buzzatii* genome.

Tandem repeat family	Repeat length	GC content (%)	Genome fraction (%)[a]	Consensus Sequence[b]	Distribution
pBuM189	189	29	0.039	GCAAAAGACTCCGTCAATTAGAAAACAAAAAATGTTATAGTTTTGAGGATTAACCGGCAAAAACCGTATTATTTGTTATATGATTTCTGTATGGAATACCGTTTTAGAAGCGTCTTTTATCGTATTACTCAGATATATCTTAAGATTTAGCATAATCTAAGAACTTTTTGAAATATTCACATTTGTCCA	*D. buzzatii* cluster species *D. mojavensis*
DbuTR198	198	34	0.027	AAGGTAGAAAGGTAGTTGGTGAGATAAACCAGAAAAAGAGCTAAAAACGGCTAAAAACGGCTAGAAAATAGCCAGAAAGGTAGATTGAACATTAATGGGCAAATGGATGGATAAATAAGACTGGTCATCATCCAATGAACAGAATCATGATTAAGAGATAGAAATATGATTAGAAAGTAGGATAGAAAGGTTAGAAAG	*D. buzzatii*

[a] Genome fraction was calculated assuming a genome size of 163.547.398 bp (version 1 freeze of all contigs).

[b] Consensus sequence generated after clustering TRF results (see Materials and Methods).

Table 4. Median estimates for dn, ds and dn/ds (ω) between *D. buzzatii* and *D. mojavensis* for chromosome X and five autosomes, for recombining and non-recombining regions, and for inverted and non-inverted regions. Only 9017 1:1 orthologs whose chromosomal location is known in *D. mojavensis* by scaffold anchoring (Schaeffer et al. 2008) and with data available for other variables (see text) were included in the analysis.

Chromosome/region	Number of genes	dn	ds	ω
All chromosomes	9017	0.0343	0.4043	0.0895
X	1352	0.0371	0.4168	0.0943
2	2303	0.0346	0.4077	0.0884
3	1683	0.0354	0.4102	0.0889
4	1806	0.0327	0.3920	0.0868
5	1844	0.0334	0.3932	0.0901
6 (dot)	29	0.0718	0.4943	0.1379
Autosomes (all)	7665	0.0340	0.4016	0.0889
Autosomes (2-5)	7636	0.0339	0.4012	0.0887
Non-recombining regions	603	0.0419	0.4564	0.0928
Recombining regions	8414	0.0339	0.3993	0.0892
Inverted regions	4220	0.0348	0.4048	0.0899
Non-inverted regions	4797	0.0338	0.4033	0.0891

Table 5. Linear regression model for divergence rates using seven regressor variables. The coefficient of determination R^2 as well as the relative contribution (%) of each variable is shown. Significant values ($P < 0.05$) are given in boldface; ns = non significant. [1]RC = Relative contribution.

	dn			ds			ω		
Linear model	Coefficient		P-value	Coefficient		P-value	Coefficient		P-value
Multiple R^2	**11.56**		< 2.2 e-16	**11.44**		< 2.2 e-16	**6.16**		< 2.2 e-16
Variable	RC[1]	Slope	P-value	RC[1]	Slope	P-value	RC[1]	Slope	P-value
Type	**1.47**	6.8 e-3	3.9 e-5	**2.33**	2.1 e-2	8.6 e-8	**0.90**	1.1 e-2	0.0247
Recombination	0.36	5.1 e-3	0.0348	**9.31**	6.3 e-2	< 2 e-16	0.09	4.7 e-3	ns
State	0.02	6.5 e-4	ns	**0.66**	8.6 e-3	0.0032	0.00	-5.0 e-4	ns
Protein length	0.40	2.9 e-6	ns	**22.95**	7.9 e-5	< 2 e-16	**8.08**	-2.0 e-5	7 e-5
Number of exons	**25.15**	-3.3 e-3	< 2 e-16	**46.60**	-1.6 e-2	< 2 e-16	**14.37**	-3.7 e-3	4.5 e-7
Breadth	**72.58**	-1.0 e-2	< 2 e-16	**16.00**	-1.1 e-2	< 2 e-16	**76.49**	-2.3 e-2	< 2 e-16
Max expression level	0.02	-1.2 e-7	ns	**2.15**	-3.2 e-6	2 e-6	0.07	-5.4 e-7	Ns
Total	100			100			100		

Table 6. GO analysis of putative genes under positive selection detected by both site models (SM) and branch-site models (BSM). Only categories showing an enrichment with a p-value < 1.0e-03 are included.

Codon subst. Models	Lineage (branch number)	Number of candidates	GO enrichment					
			Molecular Function		Biological Process		Interpro domain	
			Id	Fold enrichment	Id	Fold enrichment	Id	Fold enrichment
Site Model (SM)	Cactophilic #3	772	Antiporter activity	1.77	Regulation of transcription	4.90	Src Homology-3 domain	1.60
			Transcription factor activity	1.56				
Branch site models (BSM)	*D. buzzatii* #1	350	DNA binding	1.36	Regulation of transcription DNA dependent	1.36	Immunoglobulin-like	1.33
					Phosphate Metabolic Process	0.72		
	D. mojavensis #2	172	Dopamine beta-monooxigenase activity	2.35	Heterocycle catabolic process	2.35	DOMON (DOpamine beta-MOnooxygenase N-terminal domain)	2.35
					Cation transport	0.98		
					Histidine family amino acid catabolic process	2.35		
	Cactophilic #3	458	Zinc ion binding	2.01	Cytoeskeleton organization	1.67	Zinc Finger, PHD-type	1.93
			Transition Metal Ion Binding	2.01	Regulation of transcription DNA dependent	1.06	Proteinase inhibitor I1 kazal	2.20
			DNA binding	1.66				

FIGURES

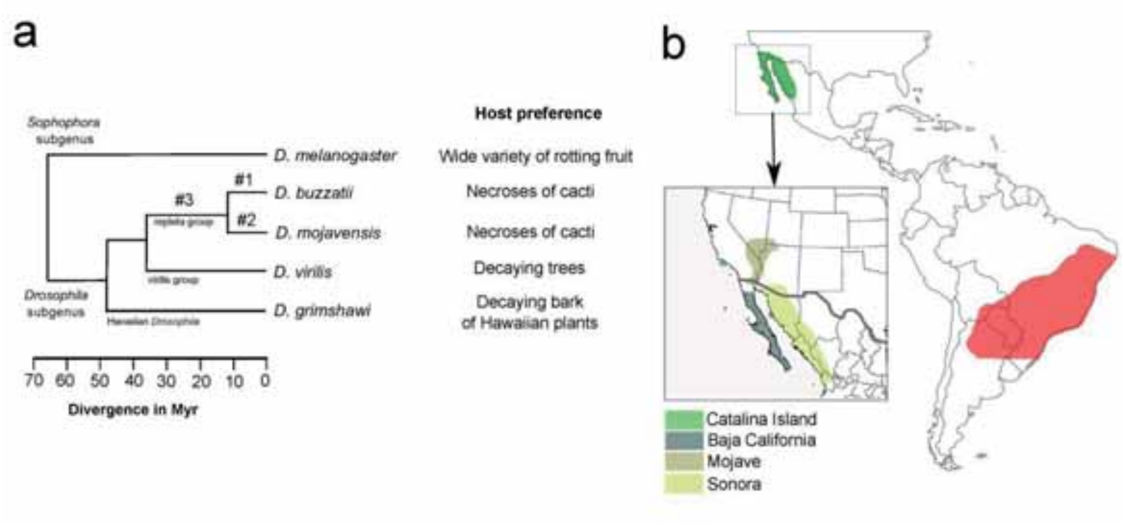


Figure 1. (a) Phylogenetic relationship of fruit fly species considered in our comparative analysis and their host preference. (b) Geographical distribution of cactophilic species *D. buzzatii* (red) and *D. mojavensis* (green) in America.

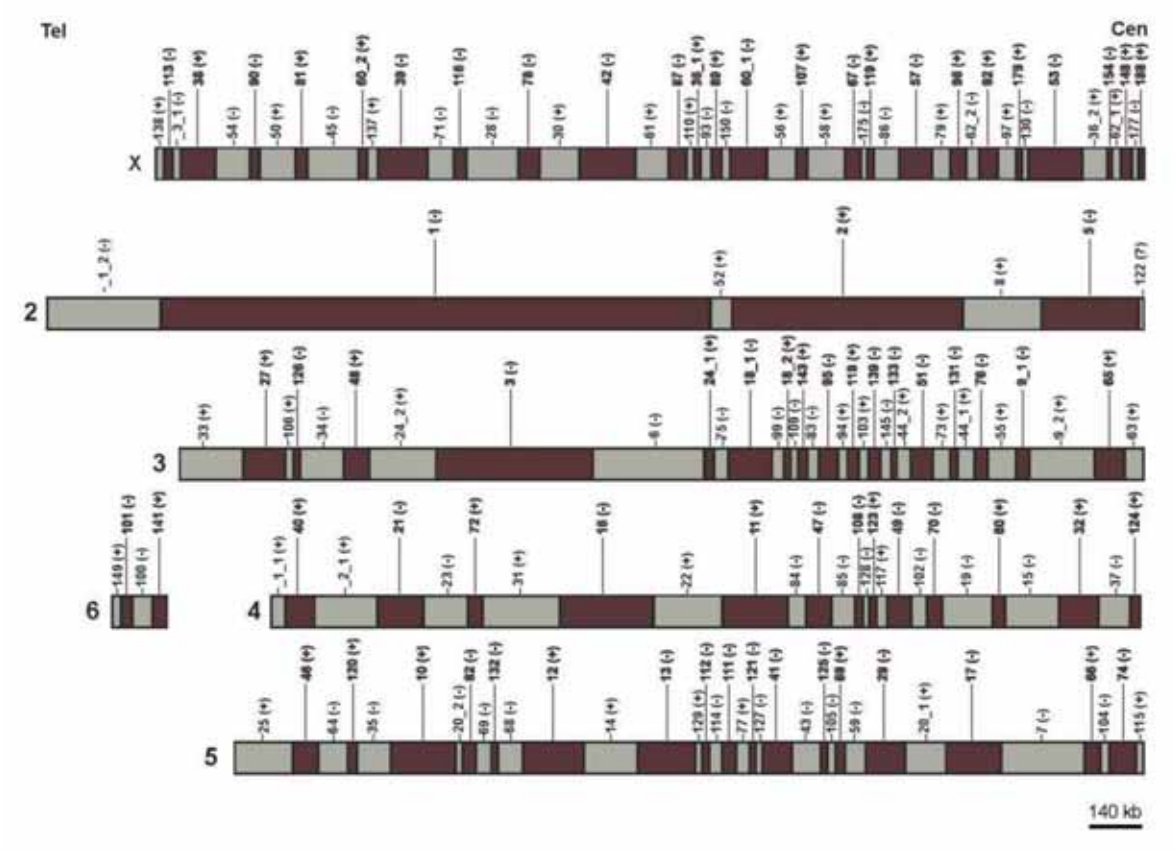


Figure 2. Order and orientation of Freeze 1 scaffolds included in N90 index within *D. buzzatii* chromosomes. Each scaffold is represented as a solid block and its orientation relative to telomere is marked by a positive (+) or negative (-) sign next to its identification number (? if direction is unknown).

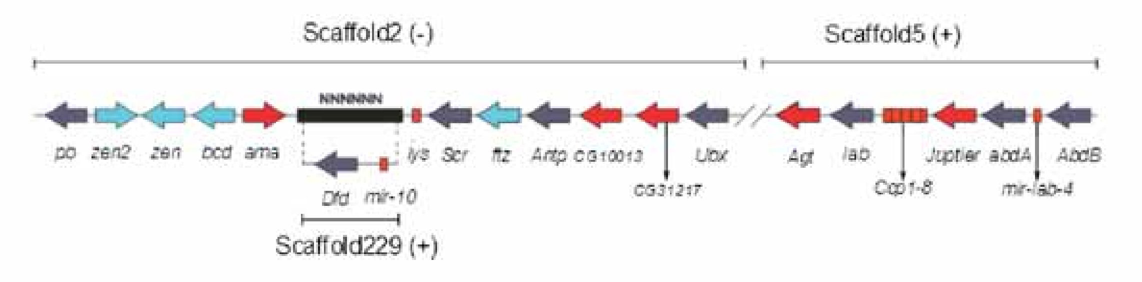


Figure 3. HOM-C structural organization in *D. buzzatii* genome. Hox genes are in dark blue, Hox-derived genes in light blue and non-Hox genes in red. The black rectangle indicates a large gap where scaffold 229 should be located.

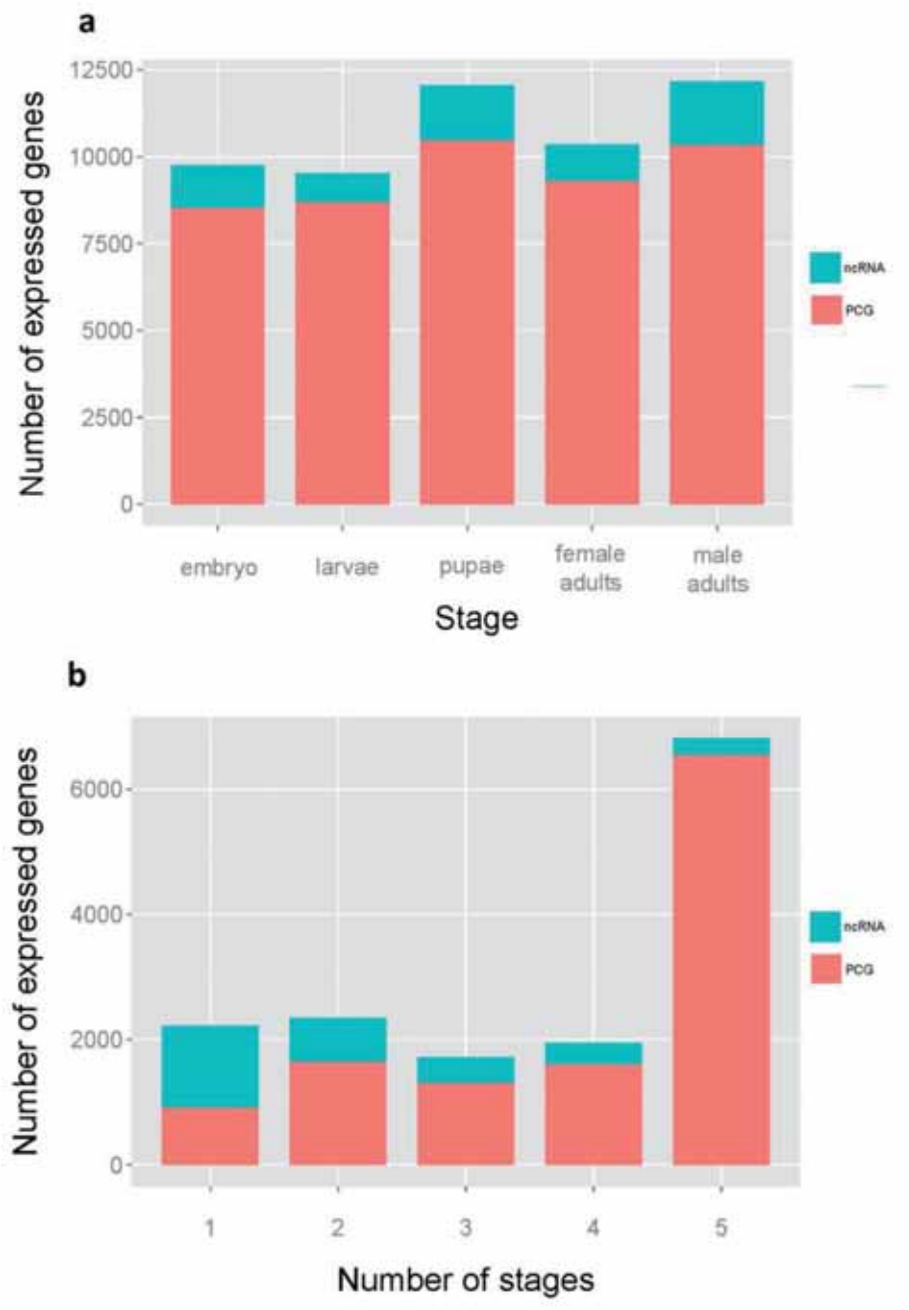


Figure 4. Developmental expression profile of *D. buzzatii* genes. (a) Number of expressed PCG (red) and ncRNA genes (blue) along five developmental stages. (b) Classification of PCG and ncRNA genes according to the number of stages where they are expressed.

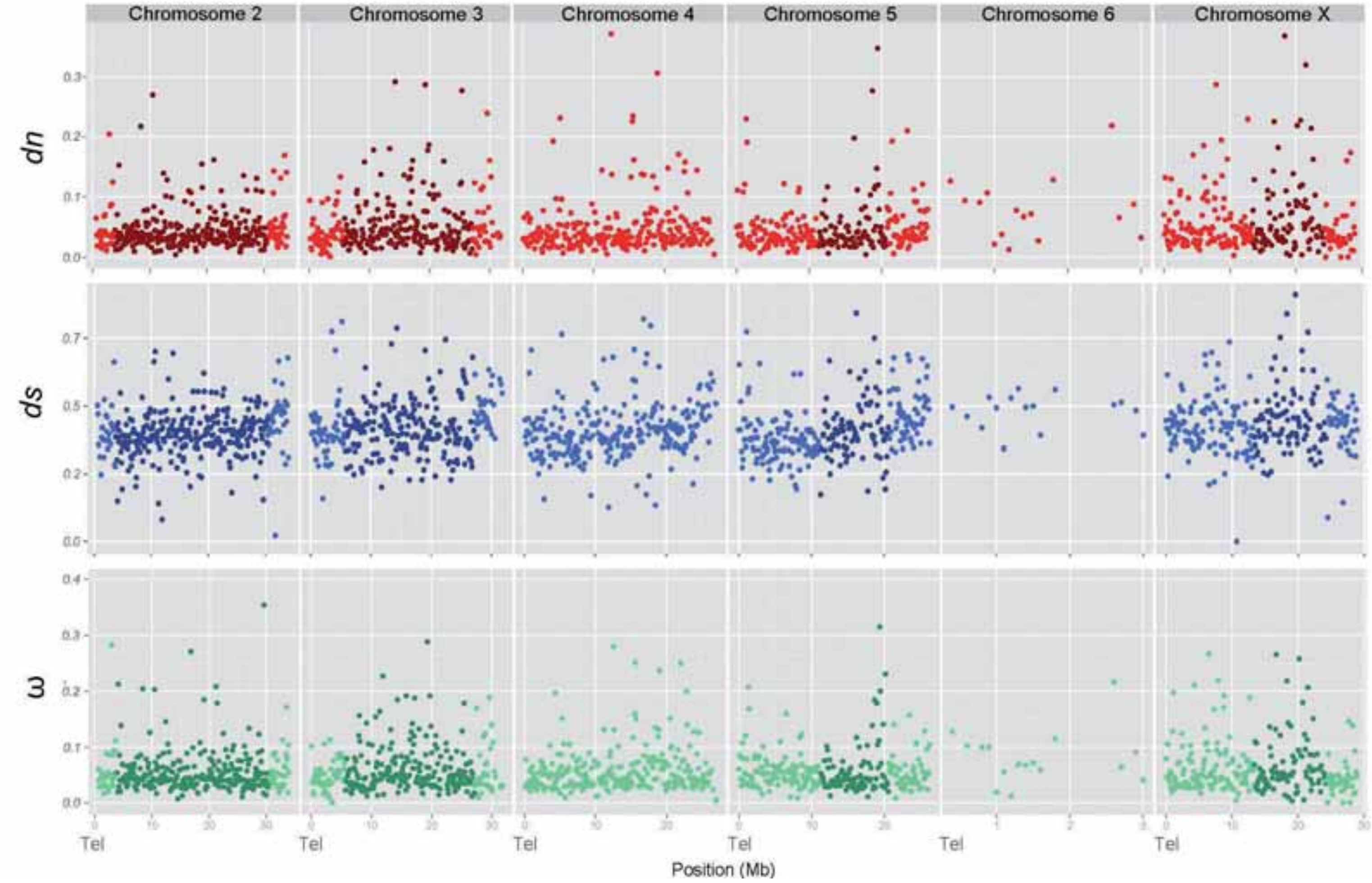

Chromosome 2
Chromosome 3
Chromosome 4
Chromosome 5
Chromosome 6
Chromosome X
dn
0.3
0.2
0.1
0.0
ds
0.7
0.5
0.2
0.0
ω
0.4
0.3
0.2
0.1
0.0
0
10
20
30
Tel
1
2
3
Position (Mb)

Figure 5. Patterns of divergence *D. buzzatii-D. mojavensis* along six *D. mojavensis* chromosomes. To construct the graph parameters were calculated in non-overlapping 100kb-windows. Coordinate 0 of x-axis corresponds to telomere. *D. mojavensis* scaffold 6540 is negatively oriented relative to telomere; thus the scaffold coordinates had to be reverted to represent chromosome 2. Windows included in regions that have been involved in chromosomal inversions are represented in darker colors (dark red for dn, dark blue for ds and dark green for ω).

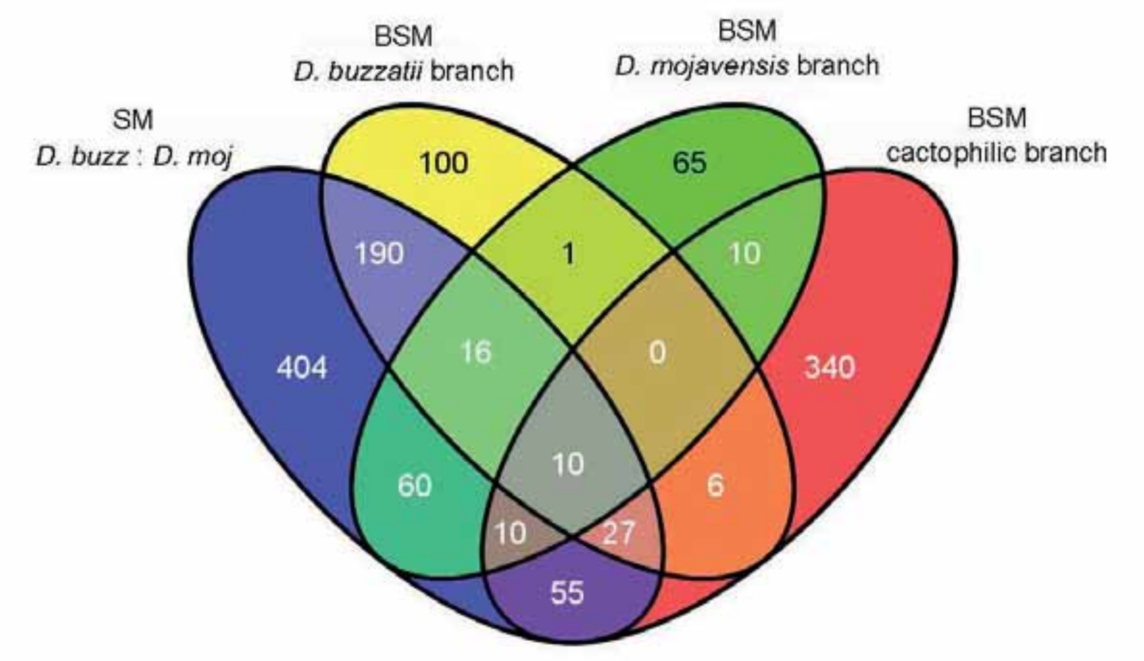


Figure 6. Venn diagram showing the number of genes under positive selection detected by two different methods, site models (SM) and branch-site models (BSM) using three different lineages as foreground branches.

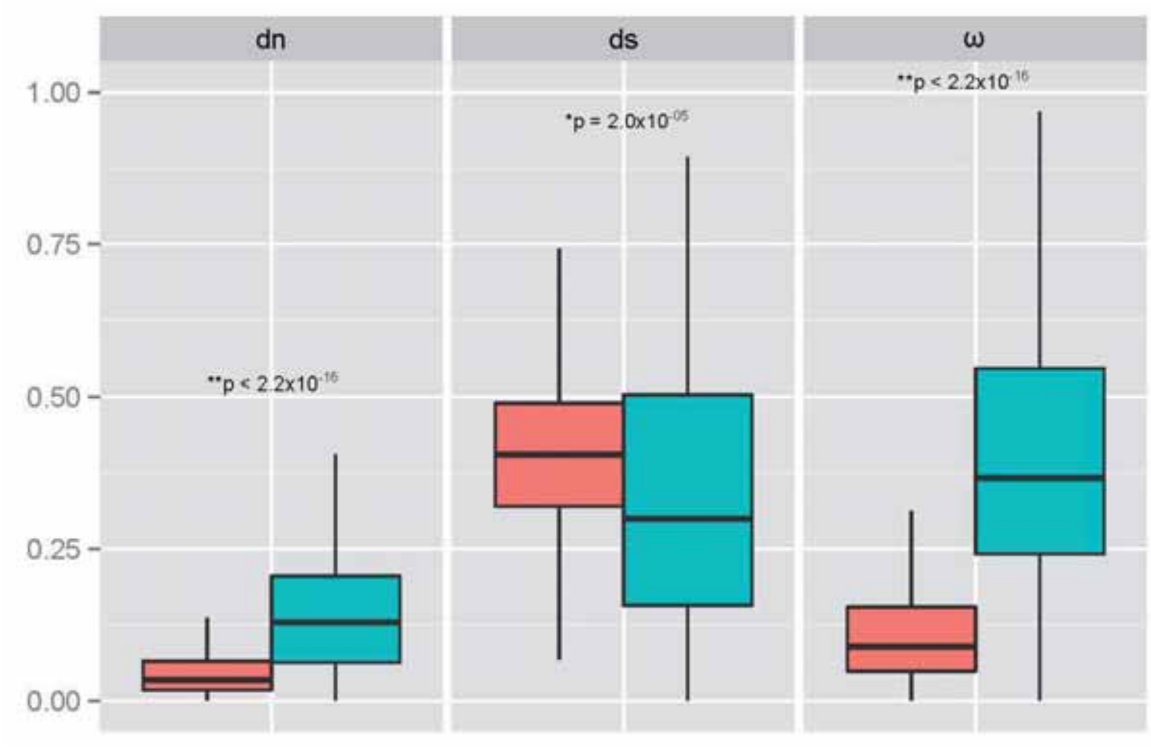


Figure 7. Patterns of divergence in orphan and non-orphan genes. Orphan genes (blue) have significantly higher dn and ω values compared to that of non-orphan genes (red). Non-orphan genes show significantly higher ds.

SUPPLEMENTAL INFORMATION - MATERIALS AND METHODS

Flies

Two strains of *Drosophila buzzatii*, st-1 and j-19, were used. Strain st-1 was isolated from flies collected in Carboneras (Spain) by repeated sib-mating and selection for chromosome arrangement *2st* (Betrán et al. 1998). This strain is isogenic for the major part of chromosome 2 and highly inbred for the rest of the genome. Strain j-19 was isolated from flies collected in Ticucho (Argentina) using the balanced-lethal stock Antp/⍰^{5} (Piccinali et al. 2007). Individuals of j-19 strain are homozygous for chromosome arrangement *2j* (Cáceres et al. 2001).

DNA extraction and sequencing

DNA was extracted from male and female adults of strains st-1 and j-19 using the sodium dodecyl sulfate (SDS) method (Milligan 1998) or the method described by Piñol and colleagues (Piñol et al. 1988) for isolating high molecular weight DNA.

Reads from different sequencing platforms were generated for strain st-1 in order to achieve an accurate assembly of the genome of this strain (Figure S1 and Table S5). Shotgun reads (3 plates, ~8x) and paired-end (PE) reads (2 plates, ~3x) were generated using GS-FLX platform (454-Roche) at the Centre for Research in Agricultural Genomics (CRAG, Barcelona, Spain). PE reads were produced from three different libraries with inserts of 6 kb (one half-plate), 7 kb (one plate) and 8 kb (one half-plate). We removed duplicate reads from 454 sequences using CDHIT 3.1.2 (Li and Godzik 2006). We also generated ~100 bp PE reads (4 lanes, ~76x) from libraries with an insert size of ~500 bp using HiSeq2000 platform (Illumina) at the Centre Nacional d'Anàlisi

Genòmica (CNAG, Barcelona, Spain). An accurate pipeline was designed in order to filter Illumina reads based on their length and quality. We first trimmed the read ends discarding bases with a quality lower than Q20 and then filtered low quality sequences (keeping only those with at least 95% of the bases with quality ≥ Q20). The final step was to discard exact duplicates and reverse complement exact duplicates from the final dataset. A mate pair (MP) library with ~7.5 kb fragments was also obtained and sequenced (one lane, ~12x) with Illumina at Macrogen Inc. (Seoul, Korea). Low quality reads as well as exact duplicates were removed (as before). Finally, we also used information provided by BAC end-sequences (BES) of 1,152 BAC clones covering *D. buzzatii* chromosome 2 (Guillén and Ruiz 2012).

De novo assembly

The assembly of the genome of strain st-1 was performed in three stages (Table S6). In the first stage, Newbler 2.6 was fed with filtered 454 reads (shotgun and PE), Sanger BES and one of the four Illumina PE lane to obtain an initial *de novo* preassembly (Figure S1). Prior to the assembly, false or chimeric 454 PE reads were discarded by mapping all the paired sequences against the *D. mojavensis* masked genome (Drosophila 12 Genomes Consortium et al. 2007) using gsMapper (Newbler 2.6). Those reads coming from the same fragment that aligned to different chromosomes as well as those aligning to multiple locations in the *D. mojavensis* scaffolds were removed. Likewise, all BES were previously filtered by mapping them against the *D. mojavensis* genome in order to remove chimeric mates and artifacts using gsMapper. Out of the initial 2304 BES, 1799 reads were used for the preassembly. We used the "*heterozygotic mode*" option in Newbler 2.6 to allow for residual nucleotide variability in the inbreed st-1 strain. We also run the "*large or complex genome*" option as we were assembling a eukaryotic genome. Thus the assembly algorithm was prepared to deal with the problem of high-copy regions, although the number of output contigs was expected to be high. The preassembly contained 2,306 scaffolds. To estimate the number of chimeric artifacts, the 38 scaffolds contained in the N50 index were mapped to the *D. mojavensis* masked genome using NUCmer (Delcher et al. 2003). Three scaffolds that matched two or more regions located in different *D. mojavensis* chromosomes were considered chimeric and split.

In a second stage, Illumina MP reads were used by SSPACE (Boetzer et al. 2011) to link output >3kb scaffolds from the preassembly and obtain 815 larger scaffolds (Table S6). A minimum number of three mate pairs were required to connect two sequences (k=3). Prior to this operation, all Illumina MP reads were mapped against the *D. buzzatii* contigs obtained from the preassembly stage (Table S6) using *bowtie2* (Langmead and Salzberg 2012). We used only MP reads that obeyed the following criteria: (I) both end sequences from the same fragment mapped to different contigs (at unknown distance); and (II) both ends mapped in the same contig at a distance greater than 4.5 kb (thus excluding inward paired end contamination). SSPACE, the software used for the scaffolding step, excluded mates not mapping at the expected set distance. After this step, a second control for chimerism was performed (as before), detecting another three chimeric scaffolds (4, 26 and 98), which were split resulting in six new scaffolds.

The third stage consisted of filling the gaps (N's) using the three short PE Illumina libraries that were not included in the pre-assembly (Table S6). GapFiller (Nadalin et al. 2012) was used in this stage, running 10 iterations and at least 4 reads needed to call a base during an extension (Figure S1). To further control for chimerism, the 818 scaffolds in the N90 scaffold index resulting from the third assembly step were blasted against the *D. mojavensis* masked genome using MUMMER and the resulting hits were reordered according to the *D. mojavensis* coordinates. This method allowed the detection of inversion breakpoint regions shared by these two species and putative chimeric scaffolds. Under a conservative criterion, eight scaffolds (9, 18, 20, 24, 36, 44, 60, 62) mapping in more than one location in the same chromosome but in regions where no inversion breakpoints or other rearrangements were expected (see Results) were split. The final assembly, named Freeze 1, thus contains 826 scaffolds >3kb and N50 and N90 index are 30 and 158, respectively.

Fold redundancy and base composition

The distribution of read depth in the st-1 genome preassembly (Figure S2) shows a Gaussian distribution with a prominent mode centered at ~22x (Figure S2). Conceivably, the scaffolding and gap filling stages of the assembly did not alter significantly this distribution. However, its variance is much larger than that expected by random (~30 times higher), showing that there is an important bias on the coverage. In particular there is a long right tail that might reflect cases where highly similar repetitive sequences or duplicated genes were merged into the same consensus sequence. One such case of misassembly was observed in the Hsp68 genes. In most Drosophila genomes there are two almost identical Hsp68 gene copies arranged head-to-head (Guillén and Ruiz 2012). In the *D. buzzatii* genome only one copy was found but it was in the vicinity of a gap (filled with N's) about the same size, suggesting that the assembler had merged all Hsp68 reads into a single gene leaving a gap in the place of the second copy.

Base composition of genes, exons and overall for Freeze 1 assembly is summarized in Table S7. CG content is ~35% overall, ~42% in gene regions (including introns) and reaches ~52% in exons. Unidentified nucleotides (N's) represent ~9% overall, ~4% in gene regions and 0.004% in exons. These patterns agree well with the reported higher CG content of genes and exons in many genomes including those of

Drosophila (Adams et al. 2000; Heger and Ponting 2007; Díaz-Castillo and Golic 2007) and humans (Bulmer 1987; Lander et al. 2001).

Sequence quality assessment and nucleotide polymorphism

To assess the quality of the Freeze 1 assembly sequence, we used ~800 kb of Sanger sequences corresponding to five *D. buzzatii* BAC clones: 40C11 (Negre et al. 2005), 5H14 (Negre et al. 2003), 20O19 and 1N19 (Calvete et al. 2012) and 1B03 (Prada et al. 2010). These BAC sequences were aligned against the genome sequence using MUMmer (Delcher et al. 2003). Some BAC regions containing repetitive elements matched multiple scaffold locations and were excluded (Table S8). Using only the unambiguously covered regions (97.6%), the genome sequence resulted 99.95% identical to that of the BAC sequences, giving an error rate of 0.0005 and a PHRED quality score of ~Q33.

In a second sequence quality assessment, we mapped the three Illumina runs (99,124,355 reads) that were used in the GapFiller stage of the assembly (Figure S1) and RNAseq data from adult males (44,840,622 reads, see below) against the Freeze 1 assembly using bowtie2 (Langmead and Salzberg 2012). Mapping of genomic reads allowed us to assess the overall genome error rate, including both expressed and non-expressed regions, whereas mapping of RNAseq reads reported the error rate exclusively for expressed regions. We considered as assembly errors those positions where 80% or more of the reads did not match the genome base and at least 80% of these unmatched positions had the same nucleotide (Figure S3). Under a conservative criterion the overall error rate was estimated to 0.0005 and the average quality ~Q33, as before. A similar value was estimated when aligning the RNAseq reads to the expressed regions of the genome (Table S9).

The strain (st-1) used for generating the *D. buzzatii* reference genome was isogenic for a large portion of chromosome 2 and highly inbreed for the remaining genome (see above). We estimated the amount of residual nucleotide polymorphism in this strain by aligning the Illumina reads against the genome Freeze 1 assembly (Figure S3). An overall proportion of segregating sites of ~0.1% was estimated (Table S10). About 15% of all the SNPs are located in gene sequences and 4% in coding exons. Thus the vast majority of SNPs are located in non-coding regions.

Genome size estimation

The genome size of two *D. buzzatii* strains, st-1 and j-19, was estimated by Feulgen Image Analysis Densitometry. The genome size of *D. mojavensis* 15081-1352.22 strain (193,826,310 bp) was used as reference (Drosophila 12 Genomes Consortium et al. 2007). Testicles from anesthetized males of both species and strains were dissected in saline solution and fixed in acetic-alcohol 3:1. Double preparations of *D. mojavensis* and *D. buzzatii* were obtained by crushing the fixed testicles in 50% acetic acid. Following Ruiz-Ruano et al. (2011), the samples were stained by Feulgen reaction including a 5N HCl incubation for 5 minutes. Images obtained by optical microscopy were analyzed with the pyFIA software (Table S11, Figure S4).

Chromosome organization and evolution

The 158 scaffolds in the N90 index were assigned to chromosomes by aligning their sequences with the *D. mojavensis* genome using blastn from MUMMER (Delcher et al. 2003). Six (out of seven) scaffolds mapping to chromosome 2 were ordered and oriented using BES and the *D. buzzatii* physical map (Gonzalez et al. 2005). The scaffolds included in N90 index mapping to chromosomes X, 4, 5 and 6 were ordered and oriented by conserved linkage (Schaeffer et al. 2008). Briefly, we looked for the position in *D. mojavensis* of genes located at the ends of *D. buzzatii* scaffolds. When two of these genes are closely located in the *D. mojavensis* genome (<200 kb in most cases) we can infer that they are also close in *D. buzzatii*, assuming synteny conservation, and then the respective scaffolds must be adjacent. This method works as far as there are no inversion breakpoints between the two scaffolds and gave consistent results for the four forementioned chromosomes. In contrast, for chromosome 3, it yielded ambiguous or inconsistent results. We had to resort to *in situ* hybridization of PCR generated probes to anchor chromosome 3 scaffolds to *D. buzzatii* polytene chromosomes (Delprat et al. in preparation).

In order to determine the organization of the HOX gene complex (HOM-C), the eight Drosophila HOX genes were searched bioinformatically in the *D. buzzatii* genome and found in three chromosome 2 scaffolds: 2, 5 and 229. Scaffold 2 contained four Hox genes (*pb*, *Scr*, *Antp* and *Ubx*) and scaffold 5 another three (*lab*, *abdA* and *AbdB*) (see Results). The eighth HOX gene, *Dfd*, was found in the small scaffold 229 (49,930 bp). We looked for the genomic position of this scaffold using BAC-end sequences and found that those of three BACs (3A12, 9B20 and 25B04) anchored this scaffold inside scaffold 2, precisely within the HOX gene complex where a 65-kb gap filled with N's was found (Figure 3). We concluded that this was a case of misassembly and the correct order of *D. buzzatii* HOX genes at this chromosomal site must be *pb*, *Dfd*, *Scr*, *Antp* and *Ubx.* All genes (HOX genes, HOX-derived genes and non-HOX genes) within the HOM-C were manually annotated using the available information (Negre et al. 2005), the annotated *D. mojavensis* and *D. melanogaster* genomes, and the RNA-seq data generated for *D. buzzatii* (Table S1).

Repeat identification and masking

A library of transposable elements (TEs) was constructed combining three different collections of repeats. The first collection was compiled blasting FlyBase canonical set of TEs against an early assembly of *D. buzzatii* genome. For each query several significant hits were manually inspected in order to recover the most complete TE copy. The second collection was build with RepeatScout 1.0.5 (Price et al. 2005) and classified by Repclass (Feschotte et al. 2009) and the third is the result of RepeatModeler 1.0.5 (Smit and Hubley 2008), with RepeatScout and RECON (Bao and Eddy 2002), both using the *D. buzzatii* early assembly. Manual analyses to reduce redundancy and remove possible protein-coding genes were performed with RepeatMasker and blast searches resulting in a library with 357 TE sequences. This library was used to mask the repeats from Freeze 1 assembly with RepeatMasker v3.2.9 (Smit et al. 1996) and annotate the protein-coding genes (see below).

A second and more comprehensive TE library (4,808 sequences) was generated adding Repbase (Jurka et al. 2005) repeats from *Insecta* species to the previous library and running again RepeatScout and RepeatModeler with *D. buzzatii* Freeze 1 assembly. Additionally, sequences classified as simple repeats, satellite or low complexity, were removed from the library. Finally, a blast analysis was performed to filter non-TE related sequences. Sequences with significant hits (e-value<1e-25) to *D. mojavensis* coding sequences (cds) and at the same time with no significant similarity to repeats deposited in Repbase were removed. This second TE library was then used to annotate and classify *D. buzzatii* TEs running RepeatMasker with the following options cutoff 250, -nolow and –norna, to prevent masking any low complexity regions and small RNA genes.

In order to identify satDNAs (highly abundant tandemly repeated DNA motifs) from the genome of *D. buzzatii*, we used the Tandem Repeats Finder (TRF) software (version 4.04) (Benson 1999). Tandem repeats searches were performed in the version 1 freeze of all contigs using the command line version of TRF with parameters 1, 1, 2, 80, 5, 200 and 750 for *match*, *mismatch*, *indel*, *probability of match*, *probability of indel*, *min. score* and *max. period*, respectively. Repeats with less than 50 bp were eliminated from the dataset. We developed a series of scripts and pipelines for clustering similar tandem repeats into major families and to eliminate redundancy between families (de Lima et al. in preparation). The outcome produced a table containing the repeat size, consensus sequence and genomic fraction of every tandem repeat family identified. From the final collection of tandem repeats, we selected the most likely satDNA families based on three main parameters: (i) abundance; (ii) no sequence similarity with transposable elements or to other non-satellite genomic elements (inferred by screening the Repbase, Genbank and FlyBase databases) and (iii) the presence of several contigs made exclusively by repeats from the same tandem repeat family.

Developmental transcriptome

Ten to twenty individuals from each of five different life stages (embryo, larvae, pupae, adult males and adult females) were collected and frozen at -80ºC. RNA from frozen samples was processed using TruSeq RNA sample preparation kit provided by Illumina. The protocol included a poly-A selection to enrich for mRNA. Library preparation was carried out at Cornell's Molecular Biology and Genetics Department, whereas RNA sequencing was done at Weill Cornell Medical College. The average insert size of the libraries from the 5 samples was 264 bp. Sequencing at PE 100 bp was performed on a Hi-Seq2000 Illumina Sequencer. A total of 378,647,052 raw reads were generated (38 Gb of sequence) comprising between 60 and 89 million reads from each of the 5 samples. RNAseq reads were trimmed and filtered by quality (at least 95% of the bases had a quality ≥ Q20) (Table S12). Filtered reads were mapped to Freeze 1 masked genome using TopHat version 1.3.3 allowing only for uniquely mapped reads

(Trapnell et al. 2009). The common setting parameters used among different stages were: -g 1 (maximum multihits) -F 0 (suppression of transcripts below this abundance level) and -i 40 (minimum intron length). The rest of parameters were set by default.

We run Cufflinks to reconstruct transcripts models and their expression level for each stage (Trapnell et al. 2010) using Annotation Release 1 as reference (-g option activated). This allowed us to identify new isoforms from expressed protein-coding genes (PCGs) and also non-coding RNA (ncRNA) genes. Transcription levels along the genome sequence and transcripts inferred by Cufflinks for each stage are included in the genome browser of the *D. buzzatii* Genome Project web (http://dbuz.uab.cat).

Protein coding gene annotation

PCGs contained by masked Freeze 1 assembly were annotated by a strategy that combined both *ab initio* and homology-based predictions. We used two HMM-based algorithms, Augustus (Stanke and Waack 2003) and SNAP (Korf 2004), and a dual-genome *de novo* software, N-SCAN (Korf et al. 2001) using as guide the alignment between *D. buzzatii* Freeze 1 assembly and *D. mojavensis* masked genome (release 1.3). Exonerate was run to identify conserved genes aligning both *D. mojavensis* and *D. melanogaster* protein databases to Freeze 1 assembly (Slater and Birney 2005). All these predictions were combined by a weight-based consensus generator, EVidence Modeler (EVM) (Haas et al. 2008) using the following weights: Exonerate *D. mojavensis* (9), Exonerate *D. melanogaster (6)*, NSCAN (6), Augustus (2) and SNAP (2). The EVM gene set contained 12,102 gene models.

There were 1,555 genes annotated by Exonerate but not reported by EVM due to their structural properties. We included these genes in Annotation Release 1 by combining EVM and Exonerate annotations using mergeBed tool from Bedtools package

(Quinlan and Hall 2010). The Annotation Release 1 includes 13,657 annotated genes (12,102 annotated by EVM and 1,555 genes detected only by Exonerate). The 1,555 genes annotated only by exonerate were shorter (Wilcoxon test, W=81226636, p-value<2.2e-16) and had less exons (W=15142546, p-value<2.2e-16). This fact indicates that algorithms that annotate genes by generating a consensus from multiple evidences are not efficient at identifying short and monoexonic genes. Some genes from the Annotation Release 1 contain internal stop codons and/or lack stop or start codons suggesting they might be misannotated PCGs or pseudogenes (Table S3).

We computed the number of wrong assembled positions contained in the total span of the gene models as well as the errors located within exons of Annotation Release 1 (see above). The vast majority of genes and exon sequences showed no assembly error positions, 91.3% and 99.2% respectively. Thus, we concluded that assembly errors are mainly contained in non-exonic regions, and both the detection of positive selection and the divergence pattern analyses carried out subsequently will not be significantly altered by misassembled sequences (Schneider et al. 2009).

Protein coding Gene Evolution

The RSD (Reciprocal Smallest Distance) algorithm (Wall and Deluca 2007) was used to identify 1:1 orthologs between *D. mojavensis* and *D. buzzatii.* The parameters used were -d 0.2 (estimated distance between species), -e 1e-08 (e-value cutoff) and the rest were set by default. Posterior alignments between pairs of orthologous proteins were performed by Clustal W (Thompson et al. 1994). To convert protein alignments to codon alignments we used *pal2nal* software (Suyama et al. 2006). Codon alignments were fed to *codeml* module of PAML 4.4 package (Yang 2007) to estimate dn, ds and ω ratio (dn/ds) of 11,154 pairs of orthologs (setting NSsites=0, single ω fixed across the phylogeny for each alignment). The orthologous pairs that reported ds>1 were considered artifacts and thus removed from the final set of genes. The 2,040 orthologs that showed a length difference higher than 20% were not considered. Our analysis evidenced that these gene pairs biased the posterior results (Figure S5).

Several causes might have generated these length differences between orthologs. Firstly, the most likely explanation is a wrong detection of exon structure of one of the copies. Secondly, RSD can report artifactual relationships, establishing wrong orthology due to the existence of similar widespread protein domains. Finally, the length difference might be a consequence of the inference of “non-ortholog isoforms” from the same pair of orthologs, i.e., the comparison of two different isoforms from the same gene in the two species compared. To investigate this possibility we calculated the correlation of the number of exons per gene between the two copies of an orthologous pair. The results indicate that there is a strong positive correlation between exon/gene ratio from orthologous gene pairs (R=0.8522, p-value<2,2e-16). It implies that the vast majority of the orthologs share the same exon-intron structure. To test whether the length difference between single-copy orthologs was caused by a wrong predicted structure of genes we performed a correlation test between the exon ratio (exon number of the *D. buzzatii* gene / exon number of the *D. mojavensis* gene) and the % protein length ratio (*D. buzzatii* protein length / *D. mojavensis* protein length). The results indicate that there exists a positive correlation between exon and length ratios (W = 125237304, p-value < 2.2e-16) and therefore the length difference between orthologs is likely due to a wrong exon-structure prediction of one of the copies.

Analysis of divergence patterns

The analysis of divergence patterns was carried on a set of 9,017 *D. buzzatii-D. mojavensis* orthologs whose chromosomal location in *D. mojavensis* is known using the statistical programming language R. The package *ggplot2* was used to generate the graphs representing dn, ds and ω medians for genes included in non-overlapping 100-kb windows across *D. mojavensis* chromosomes (Figure 5). The location of orthologous genes in *D. mojavensis* chromosomes was extracted from Schaeffer et al. (2008). Inverted chromosomal regions (dots in darker colors in Figure 5) correspond to regions involved in fixed chromosomal inversions between *D. mojavensis* and *D. buzzatii* (Guillén and Ruiz 2012; this work).

Divergence parameters were compared using the non-parametric Kruskal-Wallis test. Four tests were performed: (i) among all chromosomes; (ii) chromosome X versus autosomes; (iii) chromosome 6 (dot) versus non-dot autosomes (2-5); and (iv) chromosomes 2+3 versus chromosomes 4+5. The degrees of freedom in each case are 5, 1, 1 and 1, respectively.

We used linear models to test the joint effect on divergence of seven variables: type, recombination, state, protein length, number of exons, expression breadth and maximum expression level. Type refers to X-linked (1) or autosomal (0) gene location. Recombination was tested by comparing genes located in the non-recombining chromosome 6 (dot) or in the 3-Mb centromeric regions of the other chromosomes that have a reduced recombination rate (1) with those in the rest of chromosomal regions, presumably with normal levels of recombination (0). State indicates whether genes are located in rearranged regions (1), those involved in at least one inversion fixed between *D. mojavensis* and *D. buzzatii*, or in non-rearranged (collinear) regions (0). Protein length (in aa) and number of exons were taken from the *D. buzzatii* genome (Annotation Release 1). Expression variables (breadth and level) were assessed from the RNA-seq data collected for five life stages in *D. buzzatii* (see above). Expression breadth was measured simply as the number of life stages (0-5) in which each gene is expressed (FPKM > 1). Finally, expression level was assessed as the maximum FPKM value observed across all life stages. Three linear models were tested, one for each divergence rate (dn, ds and ω), as response variable, and the seven variables as main effects (no interaction terms were included). To assess the relative importance of each of the analyzed genomic factors in the linear models we run *pmvd* metric included in R package *relaimpo* (Groemping 2006).

Detection of genes under positive selection

To test for positive selection we made a comparison between different pairs of codon substitution models. We first run two site models on the orthologs set between *D. buzzatii* and *D. mojavensis*: M7(beta), which does not allow for positively selected sites (ω>1), and M8(beta&ω), which includes one extra class of sites to the beta model allowing for sites with ω>1 (Yang 2007). Both models were then compared using a likelihood-ratio test (LRT). We also run two more site models, M1a and M2a, and compared them again using the LRT test. Only genes that were detected as being under positive selection by both model comparisons were analyzed in further detail (see Results).

To perform the branch-site test of positive selection (Test 2) we identified 1:1:1:1 orthologs among the four available Drosophila subgenus species: *D. buzzatii*, *D. mojavensis*, *D. virilis* and *D. grimshawi* using OrthoDB version 6 database (Kriventseva et al. 2008). Branch-site models allow us detecting positive selection that affects particular sites and branches of the phylogeny. We decided to test for positive selection on three different lineages: *D. mojavensis* lineage, *D. buzzatii* lineage, and the lineage that led to the two cactophilic species (*D. buzzatii* and *D. mojavensis*) (Table S4). We run Venny software (Oliveros 2007) to create a Venn diagram showing shared selected genes among the different models. Gene expression information for positively selected genes was extracted from the Cufflinks output (see above).

Detection of orphan genes

We identified genes that are only present in the two cactophilic species, *D. mojavensis* and *D. buzzatii*, by blasting the amino acid sequences from the 1:1 orthologs between *D. mojavensis* and *D. buzzatii* (excluding missannonated genes) against all the proteins from the remaining 11 Drosophila species available in FlyBase protein database (excluding *D. mojavensis*). Proteins that showed no similarity with any Drosophila known gene product were considered putative orphans. We used a cutoff value of 1e-05 to avoid spurious hits. From the initial single-copy orthologs set between *D. mojavensis* and *D. buzzatii*, 117 proteins showed no similarity with any predicted Drosophila polypeptides. We used this set to study genes unique to the cactophilic lineage (Supplemental Table S4) and analyzed their expression pattern with TopHat and Cufflinks (see above).

SUPPLEMENTAL REFERENCES

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, et al. 2000. The genome sequence of Drosophila melanogaster. *Science* **287**: 2185–2195.

Bao Z, Eddy SR. 2002. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res* **12**: 1269–1276.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**: 573–580.

Betrán E, Santos M, 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.

Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. 2011. Scaffolding pre-assembled contigs using SSPACE. *Bioinforma Oxf Engl* **27**: 578–579.

Bulmer M. 1987. A statistical analysis of nucleotide sequences of introns and exons in human genes. *Mol Biol Evol* **4**: 395–405.

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.

Calvete O, González J, Betrán E, Ruiz A. 2012. Segmental duplication, microinversion, and gene loss associated with a complex inversion breakpoint region in Drosophila. *Mol Biol Evol* **29**: 1875–1889.

Delcher AL, Salzberg SL, Phillippy AM. 2003. Using MUMmer to identify similar regions in large sequence sets. *Curr Protoc Bioinforma Ed Board Andreas Baxevanis Al* **Chapter 10**: Unit 10.3.

Díaz-Castillo C, Golic KG. 2007. Evolution of gene sequence in response to chromosomal location. *Genetics* **177**: 359–374.

Drosophila 12 Genomes Consortium, Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA, Kaufman TC, Kellis M, Gelbart W, et al. 2007. Evolution of genes and genomes on the Drosophila phylogeny. *Nature* **450**: 203–218.

Feschotte C, Keswani U, Ranganathan N, Guibotsy ML, Levine D. 2009. Exploring repetitive DNA landscapes using REPCLASS, a tool that automates the classification of transposable elements in eukaryotic genomes. *Genome Biol Evol* **1**: 205–220.

Gonzalez 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**: 885–889.

Groemping U. 2006. Relative Importance for Linear Regression in R: The Package relaimpo. *1* **17**.

Guillén Y, Ruiz A. 2012. Gene alterations at Drosophila inversion breakpoints provide prima facie evidence for natural selection as an explanation for rapid chromosomal evolution. *BMC Genomics* **13**: 53.

Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR. 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol* **9**: R7.

Heger A, Ponting CP. 2007. Evolutionary rate analyses of orthologs and paralogs from 12 Drosophila genomes. *Genome Res* **17**: 1837–1849.

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. 2005. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* **110**: 462–467.

Korf I. 2004. Gene finding in novel genomes. *BMC Bioinformatics* **5**: 59.

Korf I, Flicek P, Duan D, Brent MR. 2001. Integrating genomic homology into gene structure prediction. *Bioinforma Oxf Engl* **17 Suppl 1**: S140–148.

Kriventseva EV, Rahman N, Espinosa O, Zdobnov EM. 2008. OrthoDB: the hierarchical catalog of eukaryotic orthologs. *Nucleic Acids Res* **36**: D271–275.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, 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.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357–359.

Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinforma Oxf Engl* **22**: 1658–1659.

Milligan B. 1998. Total DNA isolation. In *Molecular Genetic Analysis of Population: A practical approach*, pp. 29–64, Oxford University Press, Oxford, NY, Tokyo.

Nadalin F, Vezzi F, Policriti A. 2012. GapFiller: a de novo assembly approach to fill the gap within paired reads. *BMC Bioinformatics* **13 Suppl 14**: S8.

Negre B, Casillas S, Suzanne M, Sánchez-Herrero E, Akam M, Nefedov M, Barbadilla A, de Jong P, Ruiz A. 2005. Conservation of regulatory sequences and gene expression patterns in the disintegrating Drosophila Hox gene complex. *Genome Res* **15**: 692–700.

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.

Oliveros J. 2007. VENNY. An interactive tool for comparing lists with Venn diagrams. *BioinfoGP CNB-CSIC*.

Piccinali R, Mascord L, Barker J, Oakeshott J, Hasson E. 2007. Molecular Population Genetics of the α-Esterase5 Gene Locus in Original and Colonized Populations of Drosophila buzzatii and Its Sibling Drosophila koepferae. *J Mol Evol* **64**: 158–170.

Piñol J, Francino O, Fontdevila A, Cabré O. 1988. Rapid isolation of Drosophila high molecular weight DNA to obtain genomic libraries. *Nucleic Acids Res* **16**: 2736.

Prada CF, Delprat A, Ruiz A. 2011. Testing chromosomal phylogenies and inversion breakpoint reuse in Drosophila. The martensis cluster revisited. *Chromosome Res Int J Mol Supramol Evol Asp Chromosome Biol* **19**: 251–265.

Price AL, Jones NC, Pevzner PA. 2005. De novo identification of repeat families in large genomes. *Bioinforma Oxf Engl* **21 Suppl 1**: i351–358.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinforma Oxf Engl* **26**: 841–842.

Schaeffer SW, Bhutkar A, McAllister BF, Matsuda M, Matzkin LM, O'Grady PM, Rohde C, Valente VLS, Aguade M, Anderson WW, 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.

Schneider A, Souvorov A, Sabath N, Landan G, Gonnet GH, Graur D. 2009. Estimates of positive Darwinian selection are inflated by errors in sequencing, annotation, and alignment. *Genome Biol Evol* **1**: 114–118.

Slater GS, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**: 31.

Smit A, Hubley R. 2008. *RepeatModeler*. http://www.repeatmasker.org.

Smit A, Hubley R, Green P. 1996. *RepeatMasker*. http://www.repeatmasker.org.

Stanke M, Waack S. 2003. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinforma Oxf Engl* **19 Suppl 2**: ii215–225.

Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res* **34**: W609–612.

Thompson JD, Higgins DG, Gibson TJ. 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.

Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinforma Oxf Engl* **25**: 1105–1111.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**: 511–515.

Wall DP, Deluca T. 2007. Ortholog detection using the reciprocal smallest distance algorithm. *Methods Mol Biol Clifton NJ* **396**: 95–110.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**: 1586–1591.

SUPPLEMENTAL TABLES

Table S1. Manual annotation of protein-coding genes in *D. buzzatii* HOMC.

Proboscipedia (pb)

Transcript	Exon	Region	BAC 40C11	Dbuz scaffold2	Size
pb-PA	9	**UTR3'**	75576..75841	1919784..1920049	266
		CDS	75842..76292	1920050..1920500	451
	8	**CDS**	76834..77603	1921042..1921811	770
	7	**CDS**	77673..77848	1921881..1922056	176
	6	**CDS**	77916..78044	1922124..1922252	129
	5	**CDS**	78965..79079	1923173..1923287	115
	4	**CDS**	79424..79581	1923632..1923789	158
	3	**CDS**	96599..96613	1940950..1940964	15
	2	**CDS**	109654..110131	1953998..1954475	478
		UTR5'	110132..110214	1954476..1954558	83
	1	**UTR5'**	111204..112277	1955542..1956615	1074
pb-PB	9	**UTR3'**	75576..75841	1919784..1920049	266
		CDS	75842..76292	1920050..1920500	451
	8	**CDS**	76834..77603	1921042..1921811	770
	7	**CDS**	77673..77848	1921881..1922056	176
	6	**CDS**	77916..78044	1922124..1922252	129
	5	**CDS**	78965..79079	1923173..1923287	115
	4	**CDS**	79424..79566	1923632..1923774	143
	3	**CDS**	96599..96613	1940950..1940964	15
	2	**CDS**	109654..110131	1953998..1954475	478
		UTR5'	110132..110214	1954476..1954558	83
	1	**UTR5'**	111204..112277	1955542..1956615	1074

Transcript	Exon	Region	BAC 40C11	Dbuz scaffold2	Size
pb-PC	8	**UTR3'**	75576..75841	1919784..1920049	266
		CDS	75842..76292	1920050..1920500	451
	7	**CDS**	76834..77603	1921042..1921811	770
	6	**CDS**	77673..77848	1921881..1922056	176
	5	**CDS**	77916..78044	1922124..1922252	129
	4	**CDS**	78965..79079	1923173..1923287	115
	3	**CDS**	79424..79581	1923632..1923789	158
	2	**CDS**	109654..110131	1953998..1954475	478
		UTR5'	110132..110214	1954476..1954558	83
	1	**UTR5'**	111204..112277	1955542..1956615	1074
pb-PD	8	**UTR3'**	75576..75841	1919784..1920049	266
		CDS	75842..76292	1920050..1920500	451
	7	**CDS**	76834..77603	1921042..1921811	770
	6	**CDS**	77673..77848	1921881..1922056	176
	5	**CDS**	77916..78044	1922124..1922252	129
	4	**CDS**	78965..79079	1923173..1923287	115
	3	**CDS**	79424..79566	1923632..1923774	158
	2	**CDS**	109654..110131	1953998..1954475	478
		UTR5'	110132..110214	1954476..1954558	83
	1	**UTR5'**	111204..112277	1955542..1956615	1074

Deformed (Dfd)

Transcript	Exon	Region	*Dmoj scaffold_6540*	Size	*Dbuz scaffold_229*	Size	Identity	Gaps
Dfd-RA	1	5'UTR			19414..19795	382		
		CDS	16520570..16521341	772	19796..20567	772	93%	0
	2	CDS	16522602..16522693	92	21660..21751	92	93%	0
	3	CDS	16522755..16522929	175	21815..21989	175	94%	0
	4	CDS	16531918..16532309	392	30769..31148	380	98%	12
	5	CDS	16533307..16533654	348	32309..32641	333	95%	15
		3'UTR			32642..33030	389		

Sex combs reduced (Scr)

Transcript	Exon	Region	*Dmoj scaffold_6540*	Size	*Dbuz scaffold_2*	Size	Identity	Gaps
Scr-RA	1	**UTR5'**			2092196..2091356	841		
	2	**UTR5'**			2083768..2083738	31		
		CDS	16460577..16461525	949	2083737..2082795	943	96%	22
	3	**CDS**	16482110..16482417	308	2063379..2063072	308	98%	0
		UTR3'			2063071..2060846	2226		
Scr-RB	1	**UTR5'**			2093601..2093085	517		
	2	**UTR5'**			2083768..2083738	31		
		CDS	16460577..16461525	949	2083737..2082795	943	96%	22
	3	**CDS**	16482110..16482417	308	2063379..2063072	308	98%	0
		UTR3'			2063071..2060846	2226		

Antennapedia (Antp)

Transcript	Exon	Region	*Dmoj scaffold_6540*	Size	*Dbuz scaffold_2*	Size	Identity	Gaps
Antp-RA	1	5'UTR			2271808..2270969	840		
	2	5'UTR			2238817..2238740	78		
	3	5'UTR			2166782..2166543	240		
	4	5'UTR			2166486..2166361	126		
		CDS	16377826..16378449	624	2166360..2165746	615	95%	9
	5	CDS	16378611..16378649	39	2165590..2165552	39	95%	0
	6	CDS	16378763..16378985	223	2165454..2165220	235	97%	12
	7	CDS	16390892..16391142	251	2154093..2153843	251	98%	0
		3'UTR			2153842..2151440	2403		
Antp-RB	1	5'UTR			2191767..2191542	226		
	2	5'UTR			2166782..2166543	240		
	3	5'UTR			2166486..2166361	126		
		CDS	16377826..16378449	624	2166360..2165746	615	95%	9
	4	CDS	16378611..16378649	39	2165590..2165552	39	95%	0
	5	CDS	16378763..16378985	223	2165442..2165220	223	97%	0
	6	CDS	16390892..16391142	251	2154093..2153843	251	98%	0
		3'UTR			2153842..2151440	2403		
Antp-RC	1	5'UTR			2191767..2191542	226		
	2	5'UTR			2166782..2166543	240		
	3	5'UTR			2166486..2166361	126		
		CDS	16377826..16378449	624	2166360..2165746	615	95%	9
	4	CDS	16378763..16378985	223	2165442..2165220	223	97%	0
	5	CDS	16390892..16391142	251	2154093..2153843	251	98%	0
		3'UTR			2153842..2151440	2403		

Utrabithorax (Ubx)

Transcript	Exon	Region	*Dmoj scaffold_6540*	Size	*Dbuz scaffold_2*	Size	Identity	Gaps
Ubx-RA	1	5'UTR			2440200..2439170	1031		
		CDS	16091974..16092706	733	2439169..2438437	733	97%	0
	2	CDS	16102527..16102577	51	2429353..2429303	51	100%	0
	3	CDS	16122625..16122675	51	2410980..2410930	51	100%	0
	4	CDS	16190146..16190450	305	2348684..2348380	305	99%	0
		3'UTR			2348379..2345906	2474		
Ubx-RC	1	5'UTR			2440200..2439170	1031		
		CDS	16091974..16092706	733	2439169..2438437	733	97%	0
	2	CDS	16122625..16122675	51	2410980..2410930	51	100%	0
	3	CDS	16190146..16190450	305	2348684..2348380	305	99%	0
		3'UTR			2348379..2345906	2474		
Ubx-RD	1	5'UTR			2440200..2439170	1031		
		CDS	16091974..16092706	733	2439169..2438437	733	97%	0
	2	CDS	16102527..16102577	51	2429353..2429303	51	100%	0
	3	CDS	16122625..16122675	51	2410980..2410930	51	100%	0
	4	CDS	16190146..16190450	305	2348684..2348380	305	99%	0
		3'UTR			2348379..2347576	804		
Ubx-RE	1	5'UTR			2440200..2439170	1031		
		CDS	16091974..16092706	733	2439169..2438437	733	97%	0
	2	CDS	16102527..16102577	51	2429353..2429303	51	100%	0
	3	CDS	16122625..16122675	51	2410980..2410930	51	100%	0
	4	CDS	16190146..16190450	305	2348684..2348380	305	99%	0
		3'UTR			2348379..2347125	1255		

Labial (lab)

Transcript	Exon	Region	*BAC 5H14*	*Dbuz scaffold5*	Size
lab-RA	1	5'UTR	101795..102584	2677351..2678140	790
		CDS	102585..103893	2678141..2679449	1309
	2	CDS	122396..122775	2697698..2698077	380
	3	CDS	123463..123753	2698765..2699055	291
		3'UTR	123754..124024	2699056..2699326	271

Abdominal A (abdA)

Transcript	Exon	Region	*BAC 5H14*	*Dbuz scaffold_5*	Size
abdA-PA	1	UTR5'	1799..3370	2576284..2577855	1572
	2	UTR5'	4454..4576	2578939..2579061	123
	3	UTR5'	4675..4965	2579160..2579450	291
		CDS	4966..5054	2579451..2579539	89
	4	CDS	6414..6664	2580897..2581147	251
	5	CDS	10030..10077	2584994..2585041	48
	6	CDS	24314..24537	2599551..2599774	224
	7	CDS	24635..25018	2599872..2600255	384
		UTR3'	25019..26921	2600256..2600255	1903 / 1899
abdA-PB	1	UTR5'	1799..3370	2576284..2577855	1572
	2	UTR5'	4336..4350	2578821..2578835	15
		CDS	4351..5054	2578836..2579539	704
	3	CDS	6414..6664	2580897..2581147	251
	4	CDS	10030..10077	2584994..2585041	48
	5	CDS	24314..24537	2599551..2599774	224
	6	CDS	24635..25018	2599872..2600255	384
		UTR3'	25019..26921	2600256..2600255	1903 / 1899

Abdominal B (AbdB)

Transcript	Exon	Region	Dmoj scaffold_6540	Size	Dbuz scaffold5	Size	Identity	Gaps
AbdB-RA	1	5'UTR			2415774..2416013	240		
	2	5'UTR			2433706..2433751	46		
	3	5'UTR			2442652..2442800	149		
		CDS*	The translation start is different		2448001..2448195	195	97%	2
	4	CDS	2037953..2037746	208	2448344..2448551	208	96%	0
	5	CDS	2037346..2037132	215	2449020..2449234	215	92%	0
	6	CDS	2037058..2036867	192	2449303..2449494	192	97%	0
		3'UTR			2449495..2451421	1927		
		* *D. mojavensis* has more annotated exons than *D. buzzatii*						
AbdB-RB	1	5'UTR			2444187..2446373	2187		
		CDS	Not corresponding with Dbuz		2446374..2446761	388	97%*	27
	2	CDS	2038308..2038112	197	2447999..2448195	197	97%	0
	3	CDS	2037953..2037746	208	2448344..2448551	208	96%	0
	4	CDS	2037346..2037132	215	2449020..2449234	215	92%	0
	5	CDS	2037058..2036867	192	2449303..2449494	192	97%	0
		3'UTR			2449495..2451421	1927		
		*In *D. mojavensis* CDS1 is annotated otherwise. Identity (97%) of the alignment of the predicted gene with *D. mojavensis*						
AbdB-RC	1	5'UTR			2410168..2410605	438		
	2	5'UTR			2433706..2433751	46		
	3	5'UTR			2442652..2442800	149		
		CDS*	The translation start is different		2448001..2448195	195	97%	2
	4	CDS	2037953..2037746	208	2448344..2448551	208	96%	0
	5	CDS	2037346..2037132	215	2449020..2449234	215	92%	0
	6	CDS	2037058..2036867	192	2449303..2449494	192	97%	0
		3'UTR			2449495..2451421	1927		
		* *D. mojavensis* has more annotated exons than *D. buzzatii*						

AbdB-RD	**1**	**5'UTR**			2432555..2432940	386		
	2	**5'UTR**			2433706..2433751	46		
	3	**5'UTR**			2442652..2442800	149		
		CDS*	The translation start is different		2448001..2448195	195	97%	2
	4	**CDS**	2037953..2037746	208	2448344..2448551	208	96%	0
	5	**CDS**	2037346..2037132	215	2449020..2449234	215	92%	0
	6	**CDS**	2037058..2036867	192	2449303..2449494	192	97%	0
		3'UTR			2449495..2451421	1927		
		* *D. mojavensis* has more annotated exons than D. buzzatii						
Abd-RE	**1**	**5'UTR**			2444187..2444359	173		
		CDS	The translation start is different		2444360..2444414	55	100%	0
	2	**CDS**			2446312..2446761	450	95%	27
	3	**CDS**	2038308..2038112	197	2447999..2448195	197	97%	0
	4	**CDS**	2037953..2037746	208	2448344..2448551	208	96%	0
	5	**CDS**	2037346..2037132	215	2449020..2449234	215	92%	0
	6	**CDS**	2037058..2036867	192	2449303..2449494	192	97%	0
		3'UTR			2449495..2451421	1927		

zen2

Transcript	**Exon**	**Region**	*BAC 40C11*	*Scaffold_2 Dbuz*	**Size**
zen2-RA	1	**5'UTR**	116230..116292	1960568..1960630	63
		CDS	116293..116343	1960631..1960681	51
	2	**CDS**	116411..117253	1960749..1961591	843
		3'UTR	117254..117320	1961592..1961652	67

Zen

Transcript	Exon	Region	*BAC 40C11*	*Scaffold_2 Dbuz*	Size
zen-RA	1	5'UTR	127297..127247	1971634..1971584	51
		CDS	127246..127166	1971583..1971503	81
	2	CDS	127101..126187	1971438..1970524	915
		3'UTR	126186..125954	1970523..1970291	233

Fushi tarazu (ftz)

Transcript	Exon	Region	*Scaffold_6540 Dmoj*	Size	*Scaffold_2 Dbuz*	Size	Identity	Gaps
ftz-Ra	1	5'UTR			2107569..2107514	56		
		CDS	16434077..16434333	257	2107513..2106667	847	93-94%	9
		CDS	16434406..16434932	527				12
	2	CDS	16435039..16435619	581	2106545..2105968	578	94%	3
		3'UTR			2105967..2105535	433		

Bicoid (bcd)

Transcript	Exon	Region	*BAC 40C11*	*Scaffold_2 Dbuz*	Size
bcd_RA	1	5'UTR	<132938..132872	1977275..1977209	>67
		CDS	132871..132713	1977208..1977050	159
	2	CDS	130798..130484	1975135..1974821	315
		3'UTR	130483..129584	1974820..1973921	900
bcd_RD	1	5'UTR	<132938..132872	1977275..1977209	>67
		CDS	132871..132713	1977208..1977050	159
	2	CDS	132651..132576	1976988..1976913	76
	3	CDS	131937..130859	1976274..1975196	1079
	4	CDS	130798..130484	1975135..1974821	315
		3'UTR	130483..129584	1974820..1973921	900
bcd_RF	1	5'UTR	132684..132589	1977021..1976926	96
		CDS	132588..132576	1976925..1976913	13
	2	CDS	131937..130859	1976274..1975196	1079
	3	CDS	130798..130484	1975135..1974821	315
		3'UTR	130483..129584	1974820..1973921	900

Amalgam (ama)

Gene	Exon	Region	*Scaffold_6540 Dmoj*	Size	*Scaffold_2 Dbuz*	Size	Identity	Gaps
ama	1	5'UTR			1980360..1980518	159		
		CDS	16561943..16560960*	984	1980519..1981499	981	90%	3
		3'UTR			1981500..1982029	530		

**D.moj* has two coding exons annotated. *RNAseq* from *modENCODE.org* shows this is a misannotation

mir-10

Gene	*Scaffold_6540 Dmoj*	*Scaffold_229 Dbuz*	Size	Identity	Gap
mir-10	16502912..16502988	2233..2309	77	100%	0

CG10013

Gene	Exon	Region	*Scaffold_6540 Dmoj*	Size	*Scaffold_2 Dbuz*	Size	Id.	Gaps
CG10013	1	5'UTR			2310755..2310787	33		
		CDS	16224900..16226273	1374	2310788..2312128	1341	81%	57
		3'UTR			2312129..2312339	211		

CG31217

Gene	Exon	Region	*Scaffold_6540 Dmoj*	Size	*Scaffold_2 Dbuz*	Size	Identity	Gaps
CG31217	1	5'UTR			2344951..2344825	127		
		CDS	16194047..16194113	67	2344824..2344758	67	82%	0
	2	CDS	16194597..16194906	310	2344262..2343953	310	83%	0
	3	CDS	16194965..16195443	479	2343892..2343414	479	84%	0
	4	CDS	16195503..16196058	556	2343356..2342804	553	84%	3
	5	CDS	16196178..16196697	520	2342607..2342103	505	83%	15
		3'UTR			2342102..2341997	106		

Agt

Gene	Exon	Region	*Scaffold_6540 Dmoj*	Size	*Scaffold5 Dbuz*	Size	Identity	Gaps
Agt	1	**5'UTR**			2701229..2701306	78		
		CDS	1790657..1791223	567	2701307..2701873	567	84%	0
		3'UTR			2701874..2701899	26		

Ccp 1-8 *To locate the cluster, only the first and last gene were annotated*

Gene cluster Ccp	Region	*BAC 5H14*	*Scaffold_5 Dbuz*	Size
Ccp1	Exon1 (CDS)	72472..72461	2648501..2648490	12
	Exon2 (CDS)	72389..71703	2648418..2647732	687
Ccp2				
Ccp3				
Ccp4				
Ccp5				
Ccp6				
Ccp7				
Ccp8	Exon1 (CDS)	88874..88863	2663597..2663586	12
	Exon2 (CDS)	88775..88299	2663498..2663022	477

Jupiter (CDS)

Gene	*Scaffold_6540 Dmoj*	*Scaffold_5 Dbuz*	Size	Identity	Gaps
Jupiter CDS	1857120..1857181	2626740..2626801	62	94%	0
	1852438..1852571	2634097..2634230	134	84%	0
	1851902..1851934	2634735..2634767	33	100%	0
	1851197..1851442	2635246..2635491	246	94%	0
	1851000..1851136	2635556..2635692	137	93%	0

mir-iab-4

Gene	*Scaffold_6540 Dmoj*	*Scaffold_5 Dbuz*	Size	Identity	Gap
mir-iab-4	1943744..1943811	2545649..2545589	68	100%	0

Table S2. Protein-coding gene content of *D. buzzatii* genome compared to that of *D. mojavensis* and *D. melanogaster*.

Species	*D. buzzatii*	*D. mojavensis R1.3*	*D. melanogaster R5.55*
Number of genes	13657	14595	13937
Mean gene size (bp)	3108	4429	6656
Mean protein size (aa)	498	494	690
Longest gene size (bp)	67103	299059	396068
Shortest gene size (bp)	63	105	117
Longest protein size (aa)	14469	8926	22949
Shortest protein size (aa)	21	34	11
Mean number of exons	3.80	3.78	5.50

Table S3. Features of PCG models in Annotation Release 1.

	EVM	Exonerate	Total
Annotated PCGs	12102	1555	13657
Putatively correct ORFs	11213	0	11213
ORFs with internal stop codons	334	330	664
ORFs lacking start codon	163	0	163
ORFs lacking stop codons	308	654	962
ORFs lacking start and stop codons	68	571	639
ORFs no multiple of 3	16	0	16

Table S4. Candidate genes under positively selection found by comparing different site (SM) and branch site models (BSM) using the likelihood ratio test (LRT), and orphans (see next page).

SM *D. buzzatii* : *D. mojavensis*

	LRT Results	LRT Results		LRT Results	LRT Results
Flybase	M1a versus	M7 versus	Flybase	M1a versus	M7 versus
geneid	M2a	M8	geneid	M2a	M8
FBgn0084366	11.89	12.45	**FBgn0139771**	19.17	21.17
FBgn0084467	17.69	18.12	**FBgn0139800**	95.31	95.57
FBgn0085089	10.93	11.56	**FBgn0139825**	12.30	12.74
FBgn0132853	12.15	12.95	**FBgn0139908**	13.95	14.93
FBgn0132907	17.01	17.61	**FBgn0139909**	30.31	30.33
FBgn0132923	13.26	13.56	**FBgn0139941**	11.31	12.94
FBgn0133004	15.02	15.29	**FBgn0139944**	16.74	17.62
FBgn0133119	114.55	122.88	**FBgn0139946**	11.27	12.04
FBgn0133171	32.86	33.55	**FBgn0139948**	12.29	14.87
FBgn0133176	35.46	40.16	**FBgn0139969**	37.39	38.25
FBgn0133179	24.31	25.82	**FBgn0140021**	21.53	23.27
FBgn0133199	12.20	12.56	**FBgn0140023**	25.85	30.20
FBgn0133201	12.93	13.20	**FBgn0140036**	60.05	60.50
FBgn0133211	18.92	28.39	**FBgn0140045**	54.51	58.88
FBgn0133225	29.64	29.64	**FBgn0140094**	15.57	15.71
FBgn0133229	21.59	22.62	**FBgn0140142**	27.01	27.06
FBgn0133266	44.22	44.60	**FBgn0140166**	20.15	20.23
FBgn0133272	259.26	259.91	**FBgn0140167**	13.87	14.82
FBgn0133282	10.84	11.69	**FBgn0140218**	21.11	24.45
FBgn0133302	18.83	18.99	**FBgn0140252**	12.61	13.71
FBgn0133309	61.24	62.36	**FBgn0140297**	23.60	27.21
FBgn0133319	11.77	12.81	**FBgn0140310**	12.94	13.39
FBgn0133324	13.91	14.84	**FBgn0140340**	15.44	15.47
FBgn0133334	21.37	21.81	**FBgn0140354**	13.96	18.07
FBgn0133389	20.31	21.01	**FBgn0140377**	14.32	15.99
FBgn0133409	10.96	11.04	**FBgn0140391**	22.50	22.52
FBgn0133455	100.33	100.29	**FBgn0140397**	20.54	20.53
FBgn0133473	17.43	17.64	**FBgn0140405**	20.48	22.19
FBgn0133565	18.64	18.90	**FBgn0140427**	15.68	17.37
FBgn0133573	18.77	18.26	**FBgn0140440**	14.49	14.65
FBgn0133583	11.41	13.71	**FBgn0140449**	37.23	41.73
FBgn0133587	22.90	26.89	**FBgn0140468**	10.89	11.57
FBgn0133615	11.37	11.47	**FBgn0140474**	11.40	11.43
FBgn0133622	16.93	20.47	**FBgn0140488**	11.13	11.14
FBgn0133637	13.98	15.37	**FBgn0140536**	12.28	12.61
FBgn0133665	30.61	30.84	**FBgn0140558**	11.63	14.01

FBgn0133670	71.31	76.72	**FBgn0140562**	12.67	16.04
FBgn0133674	20.57	20.67	**FBgn0140586**	12.03	13.14
FBgn0133679	11.99	12.11	**FBgn0140587**	11.74	12.12
FBgn0133693	18.48	21.12	**FBgn0140719**	20.46	20.92
FBgn0133697	25.61	30.62	**FBgn0140727**	31.15	31.17
FBgn0133698	20.56	21.53	**FBgn0140736**	11.85	12.55
FBgn0133704	62.94	64.92	**FBgn0140743**	15.44	16.73
FBgn0133733	11.75	11.97	**FBgn0140758**	18.60	23.02
FBgn0133743	22.29	23.05	**FBgn0140759**	16.59	18.33
FBgn0133744	14.00	14.89	**FBgn0140765**	39.74	42.82
FBgn0133745	21.32	21.78	**FBgn0140774**	12.49	12.65
FBgn0133753	14.00	15.41	**FBgn0140778**	14.96	15.30
FBgn0133754	24.91	25.76	**FBgn0140825**	11.33	11.67
FBgn0133776	18.35	18.44	**FBgn0140827**	16.45	18.66
FBgn0133819	13.62	16.20	**FBgn0140840**	11.89	12.61
FBgn0133837	12.42	13.93	**FBgn0140871**	13.43	15.32
FBgn0133848	19.39	25.67	**FBgn0140920**	31.19	31.18
FBgn0133866	14.40	15.08	**FBgn0140923**	13.07	13.45
FBgn0133869	12.59	12.65	**FBgn0140983**	34.08	38.63
FBgn0133889	11.68	12.00	**FBgn0141006**	13.15	21.10
FBgn0133897	14.27	14.95	**FBgn0141099**	13.53	13.53
FBgn0133916	23.31	25.52	**FBgn0141105**	19.01	21.22
FBgn0133918	12.74	14.09	**FBgn0141113**	13.08	15.23
FBgn0133924	18.84	19.17	**FBgn0141119**	13.66	14.27
FBgn0133936	11.36	11.93	**FBgn0141170**	28.96	29.26
FBgn0133967	45.54	46.44	**FBgn0141171**	11.44	12.13
FBgn0133981	13.15	14.13	**FBgn0141174**	15.20	15.66
FBgn0134099	15.38	15.41	**FBgn0141178**	25.52	27.21
FBgn0134159	37.28	38.51	**FBgn0141189**	15.32	18.51
FBgn0134184	71.16	71.25	**FBgn0141193**	199.05	201.61
FBgn0134227	22.04	22.19	**FBgn0141205**	92.77	92.81
FBgn0134228	31.84	32.17	**FBgn0141206**	14.04	14.28
FBgn0134235	209.26	209.75	**FBgn0141232**	24.60	25.05
FBgn0134268	12.82	12.81	**FBgn0141244**	32.35	32.34
FBgn0134274	11.77	12.16	**FBgn0141287**	24.50	27.21
FBgn0134284	11.30	11.69	**FBgn0141295**	84.90	87.70
FBgn0134345	11.33	13.35	**FBgn0141315**	11.97	12.15
FBgn0134351	15.24	15.75	**FBgn0141362**	18.26	18.74
FBgn0134358	26.45	28.32	**FBgn0141371**	12.14	12.41
FBgn0134366	86.15	90.69	**FBgn0141373**	34.66	34.66

FBgn0134372	26.58	26.82	**FBgn0141406**	17.55	18.17
FBgn0134377	17.63	17.79	**FBgn0141448**	39.11	40.46
FBgn0134393	10.95	18.54	**FBgn0141463**	21.48	21.54
FBgn0134410	14.89	15.10	**FBgn0141543**	10.93	11.04
FBgn0134420	11.77	12.43	**FBgn0141613**	22.88	23.37
FBgn0134443	105.51	115.38	**FBgn0141659**	16.17	16.89
FBgn0134444	15.99	16.28	**FBgn0141675**	14.63	14.61
FBgn0134468	19.71	20.59	**FBgn0141677**	30.89	33.93
FBgn0134486	12.82	13.26	**FBgn0141681**	48.07	48.11
FBgn0134535	16.67	16.92	**FBgn0141726**	23.03	23.09
FBgn0134537	71.03	75.47	**FBgn0141742**	12.81	12.51
FBgn0134544	43.38	43.74	**FBgn0141750**	11.19	15.35
FBgn0134552	36.69	40.01	**FBgn0141761**	16.28	17.72
FBgn0134565	31.93	33.62	**FBgn0141766**	65.77	66.13
FBgn0134589	15.40	23.22	**FBgn0141783**	38.35	39.64
FBgn0134605	15.69	19.61	**FBgn0141810**	25.10	27.08
FBgn0134610	15.01	15.70	**FBgn0141859**	22.78	23.08
FBgn0134620	12.40	14.31	**FBgn0141861**	15.09	15.60
FBgn0134651	14.16	14.40	**FBgn0141864**	29.00	28.91
FBgn0134666	15.50	15.75	**FBgn0141879**	103.64	103.99
FBgn0134692	35.92	36.18	**FBgn0141887**	33.35	38.86
FBgn0134700	15.46	15.58	**FBgn0141909**	33.54	35.51
FBgn0134753	11.68	13.40	**FBgn0141920**	34.80	35.09
FBgn0134759	18.39	18.89	**FBgn0141945**	17.76	17.41
FBgn0134797	12.55	14.00	**FBgn0141950**	19.20	19.61
FBgn0134800	12.49	12.68	**FBgn0141995**	11.43	16.57
FBgn0134830	14.71	16.07	**FBgn0142012**	28.69	29.11
FBgn0134854	14.25	14.50	**FBgn0142013**	91.57	91.70
FBgn0134858	82.74	86.27	**FBgn0142017**	24.57	25.64
FBgn0134860	12.40	12.41	**FBgn0142038**	14.68	15.82
FBgn0134886	11.43	11.52	**FBgn0142041**	19.27	19.69
FBgn0134901	32.80	36.81	**FBgn0142061**	19.28	19.22
FBgn0134911	47.13	47.14	**FBgn0142064**	32.47	37.15
FBgn0134920	67.13	75.79	**FBgn0142077**	11.43	15.53
FBgn0134937	15.06	15.49	**FBgn0142078**	11.39	12.90
FBgn0134959	13.75	14.32	**FBgn0142086**	119.14	122.29
FBgn0134970	20.14	22.19	**FBgn0142103**	20.74	21.43
FBgn0135018	14.35	15.99	**FBgn0142104**	24.97	25.22
FBgn0135023	29.22	29.35	**FBgn0142105**	14.32	17.08
FBgn0135027	17.09	17.72	**FBgn0142109**	17.14	18.00

FBgn0135037	14.60	14.74	**FBgn0142120**	14.90	15.14
FBgn0135040	17.46	17.52	**FBgn0142135**	22.17	25.35
FBgn0135041	61.34	64.23	**FBgn0142169**	58.75	58.76
FBgn0135054	18.91	20.38	**FBgn0142185**	13.99	13.94
FBgn0135076	153.76	156.25	**FBgn0142192**	22.01	22.04
FBgn0135080	97.72	100.33	**FBgn0142194**	12.21	13.64
FBgn0135081	16.80	16.87	**FBgn0142195**	18.07	29.00
FBgn0135106	17.55	18.15	**FBgn0142210**	15.20	15.52
FBgn0135126	11.19	13.14	**FBgn0142223**	26.14	26.29
FBgn0135138	29.94	30.54	**FBgn0142275**	26.83	27.13
FBgn0135154	17.67	18.17	**FBgn0142318**	13.17	15.73
FBgn0135156	11.71	13.96	**FBgn0142322**	27.36	27.42
FBgn0135164	15.65	16.40	**FBgn0142336**	24.54	26.24
FBgn0135210	13.07	13.31	**FBgn0142345**	34.06	34.15
FBgn0135227	16.76	17.31	**FBgn0142347**	84.00	88.20
FBgn0135228	16.31	16.39	**FBgn0142366**	22.31	26.96
FBgn0135231	35.65	39.95	**FBgn0142379**	15.65	16.65
FBgn0135290	11.05	11.78	**FBgn0142408**	31.68	31.70
FBgn0135306	26.18	26.61	**FBgn0142420**	25.01	33.78
FBgn0135323	13.84	13.85	**FBgn0142424**	20.96	21.37
FBgn0135325	20.65	65.19	**FBgn0142438**	40.31	41.76
FBgn0135348	11.13	11.37	**FBgn0142461**	11.36	11.50
FBgn0135349	16.77	17.23	**FBgn0142475**	46.19	53.16
FBgn0135350	13.54	13.85	**FBgn0142496**	40.05	40.12
FBgn0135360	12.14	16.11	**FBgn0142497**	17.92	18.62
FBgn0135446	15.90	17.99	**FBgn0142503**	30.42	31.87
FBgn0135450	16.39	17.68	**FBgn0142530**	11.41	11.45
FBgn0135464	24.40	27.60	**FBgn0142551**	31.70	32.64
FBgn0135465	13.17	14.01	**FBgn0142553**	161.62	165.33
FBgn0135478	73.99	75.37	**FBgn0142556**	26.27	26.87
FBgn0135480	12.78	15.50	**FBgn0142568**	14.98	15.15
FBgn0135483	47.09	47.38	**FBgn0142574**	28.92	28.90
FBgn0135502	25.62	25.66	**FBgn0142578**	16.18	16.69
FBgn0135526	15.74	16.43	**FBgn0142608**	126.00	126.68
FBgn0135556	21.48	22.59	**FBgn0142618**	49.70	55.31
FBgn0135577	10.97	11.77	**FBgn0142620**	12.76	19.19
FBgn0135584	63.63	63.59	**FBgn0142630**	11.11	11.71
FBgn0135590	18.81	19.12	**FBgn0142635**	16.20	17.53
FBgn0135625	33.48	35.60	**FBgn0142654**	20.08	21.58
FBgn0135627	30.59	33.22	**FBgn0142655**	54.50	62.15

FBgn0135632	41.16	41.66	**FBgn0142678**	125.11	125.31
FBgn0135679	11.60	11.69	**FBgn0142683**	19.82	23.38
FBgn0135693	29.07	33.81	**FBgn0142695**	70.76	77.38
FBgn0135714	26.21	27.01	**FBgn0142705**	96.00	97.61
FBgn0135746	11.15	12.12	**FBgn0142710**	33.60	38.37
FBgn0135775	11.76	12.14	**FBgn0142720**	32.65	32.93
FBgn0135786	31.65	31.73	**FBgn0142728**	24.31	24.62
FBgn0135789	36.08	35.15	**FBgn0142738**	17.35	17.68
FBgn0135804	21.10	25.88	**FBgn0142780**	13.39	13.59
FBgn0135817	30.39	30.44	**FBgn0142825**	12.25	12.81
FBgn0135849	10.87	11.82	**FBgn0142830**	81.80	82.02
FBgn0135864	15.01	22.74	**FBgn0142833**	10.93	10.95
FBgn0135883	33.24	32.68	**FBgn0142845**	18.76	21.06
FBgn0135887	19.49	21.32	**FBgn0142892**	17.27	17.40
FBgn0135890	88.34	88.31	**FBgn0142909**	11.37	15.36
FBgn0135906	21.20	21.71	**FBgn0142945**	15.27	15.49
FBgn0135920	24.23	24.27	**FBgn0142947**	30.64	30.80
FBgn0135941	74.77	82.01	**FBgn0143003**	21.64	22.49
FBgn0135944	23.77	24.83	**FBgn0143017**	14.00	18.62
FBgn0135952	22.30	23.07	**FBgn0143020**	11.24	11.93
FBgn0135955	24.12	24.22	**FBgn0143050**	13.13	13.11
FBgn0135960	26.42	40.63	**FBgn0143063**	14.51	15.39
FBgn0135964	17.58	18.97	**FBgn0143078**	18.35	18.36
FBgn0135982	15.34	17.99	**FBgn0143099**	16.99	17.00
FBgn0135994	13.77	17.47	**FBgn0143111**	13.32	13.33
FBgn0136002	18.14	18.05	**FBgn0143112**	21.26	22.04
FBgn0136008	20.46	21.76	**FBgn0143128**	24.36	24.96
FBgn0136026	49.01	65.46	**FBgn0143137**	12.58	13.07
FBgn0136037	18.22	18.80	**FBgn0143165**	16.34	16.79
FBgn0136039	21.66	22.04	**FBgn0143184**	13.88	14.56
FBgn0136054	11.64	12.41	**FBgn0143189**	51.34	56.36
FBgn0136061	28.88	29.32	**FBgn0143194**	27.19	28.54
FBgn0136065	11.31	12.54	**FBgn0143211**	24.02	24.28
FBgn0136073	92.83	93.28	**FBgn0143240**	18.62	19.20
FBgn0136098	12.31	12.96	**FBgn0143269**	14.68	14.82
FBgn0136118	12.54	13.11	**FBgn0143279**	27.53	29.99
FBgn0136189	15.70	15.84	**FBgn0143280**	65.24	67.44
FBgn0136218	39.41	41.65	**FBgn0143338**	15.78	17.01
FBgn0136257	12.95	16.58	**FBgn0143342**	14.54	14.58
FBgn0136259	11.54	12.50	**FBgn0143393**	19.78	20.79

FBgn0136267	13.85	13.95	**FBgn0143413**	28.45	29.18
FBgn0136304	16.09	18.17	**FBgn0143416**	88.82	100.29
FBgn0136307	12.14	12.78	**FBgn0143420**	20.20	33.93
FBgn0136313	52.65	52.68	**FBgn0143438**	47.64	55.37
FBgn0136314	11.78	13.47	**FBgn0143467**	18.38	18.51
FBgn0136316	54.41	54.45	**FBgn0143470**	43.04	44.94
FBgn0136349	15.56	16.89	**FBgn0143489**	18.25	19.30
FBgn0136354	26.42	26.49	**FBgn0143490**	39.90	40.06
FBgn0136357	16.49	18.78	**FBgn0143533**	22.78	22.83
FBgn0136372	12.39	12.43	**FBgn0143588**	20.44	23.35
FBgn0136408	17.50	17.58	**FBgn0143645**	14.08	13.94
FBgn0136426	32.28	32.75	**FBgn0143696**	22.24	22.28
FBgn0136434	11.48	12.24	**FBgn0143711**	64.18	65.95
FBgn0136441	12.88	12.77	**FBgn0143727**	19.83	20.41
FBgn0136447	11.06	11.13	**FBgn0143728**	20.18	20.18
FBgn0136470	76.77	79.11	**FBgn0143755**	12.02	17.01
FBgn0136508	15.28	16.45	**FBgn0143767**	16.05	16.05
FBgn0136544	19.38	20.00	**FBgn0143791**	13.68	15.19
FBgn0136547	18.77	24.38	**FBgn0143796**	14.27	14.30
FBgn0136549	14.77	15.71	**FBgn0143802**	11.36	11.72
FBgn0136585	13.57	13.84	**FBgn0143824**	12.31	15.76
FBgn0136590	11.42	12.04	**FBgn0143898**	20.75	20.78
FBgn0136604	18.67	18.68	**FBgn0144011**	13.37	21.14
FBgn0136642	21.83	24.18	**FBgn0144045**	18.63	18.89
FBgn0136647	76.35	81.84	**FBgn0144119**	11.06	11.34
FBgn0136663	32.85	35.71	**FBgn0144171**	86.22	96.64
FBgn0136724	20.70	20.68	**FBgn0144199**	13.68	13.73
FBgn0136802	26.62	28.37	**FBgn0144211**	14.08	14.31
FBgn0136806	12.00	14.00	**FBgn0144215**	117.74	119.80
FBgn0136807	18.73	21.11	**FBgn0144218**	20.62	21.40
FBgn0136845	14.26	23.99	**FBgn0144271**	51.49	51.81
FBgn0136889	18.13	20.52	**FBgn0144317**	10.85	11.36
FBgn0136917	15.84	20.00	**FBgn0144326**	13.93	15.08
FBgn0136954	14.02	14.09	**FBgn0144327**	19.52	20.95
FBgn0136984	52.29	52.29	**FBgn0144353**	11.23	12.23
FBgn0136990	17.75	18.22	**FBgn0144363**	23.86	25.42
FBgn0137015	17.80	17.95	**FBgn0144371**	165.63	168.79
FBgn0137027	20.53	21.34	**FBgn0144385**	19.52	20.98
FBgn0137036	14.00	14.17	**FBgn0144392**	10.97	16.71
FBgn0137067	12.10	13.62	**FBgn0144414**	42.15	41.96

FBgn0137078	29.64	29.65	**FBgn0144444**	26.30	29.60
FBgn0137157	11.85	12.53	**FBgn0144482**	11.94	15.40
FBgn0137159	14.82	16.30	**FBgn0144499**	78.62	79.43
FBgn0137257	20.98	25.86	**FBgn0144501**	11.77	13.45
FBgn0137315	15.49	16.66	**FBgn0144503**	14.76	15.16
FBgn0137320	136.41	137.74	**FBgn0144514**	14.63	16.00
FBgn0137378	46.08	58.75	**FBgn0144520**	17.79	19.96
FBgn0137381	41.40	41.77	**FBgn0144526**	29.32	29.97
FBgn0137398	49.09	49.64	**FBgn0144528**	11.35	12.28
FBgn0137401	31.23	31.37	**FBgn0144607**	12.70	12.85
FBgn0137439	37.27	37.46	**FBgn0144647**	14.84	15.28
FBgn0137464	38.06	40.20	**FBgn0144666**	18.34	19.00
FBgn0137467	50.76	52.49	**FBgn0144681**	13.35	13.71
FBgn0137469	11.58	11.95	**FBgn0144684**	115.37	115.93
FBgn0137484	23.79	24.06	**FBgn0144686**	16.99	18.25
FBgn0137504	47.36	47.41	**FBgn0144687**	16.31	16.41
FBgn0137509	14.92	17.84	**FBgn0144689**	34.75	34.94
FBgn0137548	15.00	17.16	**FBgn0144690**	15.96	17.75
FBgn0137553	16.11	17.14	**FBgn0144691**	29.41	28.54
FBgn0137601	18.07	18.67	**FBgn0144727**	33.64	33.67
FBgn0137611	11.21	12.19	**FBgn0144743**	17.94	28.63
FBgn0137613	49.10	49.31	**FBgn0144753**	17.73	18.21
FBgn0137617	19.05	19.75	**FBgn0144757**	12.03	12.73
FBgn0137629	0.00	0.00	**FBgn0144796**	29.56	31.50
FBgn0137631	11.27	11.80	**FBgn0144838**	31.43	40.97
FBgn0137633	27.19	27.37	**FBgn0144858**	22.17	22.47
FBgn0137634	18.30	19.55	**FBgn0144861**	38.75	41.00
FBgn0137643	24.00	33.41	**FBgn0144884**	15.97	17.18
FBgn0137695	14.19	19.12	**FBgn0144886**	20.95	22.24
FBgn0137702	23.68	26.97	**FBgn0144894**	37.78	38.37
FBgn0137715	15.60	18.70	**FBgn0144929**	14.03	14.40
FBgn0137731	26.09	29.66	**FBgn0144933**	18.35	18.39
FBgn0137749	13.30	14.47	**FBgn0144941**	21.25	21.24
FBgn0137797	17.09	17.38	**FBgn0144957**	11.76	11.94
FBgn0137799	23.89	24.99	**FBgn0144970**	15.70	15.88
FBgn0137810	24.83	25.26	**FBgn0144975**	12.18	12.02
FBgn0137820	27.85	31.75	**FBgn0144984**	13.84	14.82
FBgn0137830	16.88	18.37	**FBgn0145031**	17.80	23.44
FBgn0137837	22.97	23.05	**FBgn0145052**	31.55	31.61
FBgn0137845	18.37	24.34	**FBgn0145071**	13.97	15.85

FBgn0137869	17.40	26.43	**FBgn0145093**	23.38	24.24
FBgn0137883	83.47	85.23	**FBgn0145094**	23.92	24.51
FBgn0137896	45.67	47.35	**FBgn0145115**	20.35	24.56
FBgn0137898	35.79	36.73	**FBgn0145116**	12.02	12.07
FBgn0137903	94.27	95.32	**FBgn0145135**	89.54	99.56
FBgn0137904	16.67	18.67	**FBgn0145156**	31.66	33.67
FBgn0137949	16.14	16.21	**FBgn0145172**	11.09	11.24
FBgn0137953	12.35	13.17	**FBgn0145179**	19.67	20.24
FBgn0137954	14.68	26.56	**FBgn0145248**	17.88	18.15
FBgn0137955	61.93	63.39	**FBgn0145250**	13.52	14.72
FBgn0137960	17.11	19.70	**FBgn0145266**	18.97	19.25
FBgn0137964	50.93	51.64	**FBgn0145274**	36.76	36.79
FBgn0137975	120.81	121.22	**FBgn0145275**	11.13	11.60
FBgn0137993	41.73	43.05	**FBgn0145332**	12.17	13.40
FBgn0138000	19.01	20.37	**FBgn0145369**	17.70	17.75
FBgn0138004	12.28	12.54	**FBgn0145375**	48.75	49.24
FBgn0138007	36.14	37.74	**FBgn0145390**	18.70	17.31
FBgn0138016	14.67	15.37	**FBgn0145432**	20.59	20.96
FBgn0138033	13.60	14.69	**FBgn0145493**	55.97	56.58
FBgn0138056	12.49	12.80	**FBgn0145521**	34.64	43.69
FBgn0138060	31.53	36.60	**FBgn0145527**	15.62	16.00
FBgn0138078	49.53	50.82	**FBgn0145602**	13.40	13.69
FBgn0138080	21.39	28.90	**FBgn0145656**	27.57	27.48
FBgn0138086	34.51	36.52	**FBgn0145681**	15.52	15.68
FBgn0138101	16.25	18.71	**FBgn0145701**	20.09	20.22
FBgn0138120	11.15	12.32	**FBgn0145716**	12.88	12.86
FBgn0138130	12.77	15.21	**FBgn0145748**	29.78	32.01
FBgn0138145	23.69	24.02	**FBgn0145753**	16.32	16.60
FBgn0138162	10.94	11.21	**FBgn0145757**	10.94	10.99
FBgn0138178	44.75	56.65	**FBgn0145799**	20.03	22.45
FBgn0138209	14.15	17.11	**FBgn0145831**	28.62	29.90
FBgn0138223	11.60	12.41	**FBgn0145837**	15.34	19.60
FBgn0138227	12.27	12.70	**FBgn0145839**	38.13	39.33
FBgn0138228	11.52	16.12	**FBgn0145851**	18.53	18.56
FBgn0138246	11.52	11.79	**FBgn0145889**	12.85	16.57
FBgn0138276	12.43	13.01	**FBgn0145902**	32.36	36.04
FBgn0138288	17.59	19.07	**FBgn0145908**	17.62	18.85
FBgn0138314	16.52	18.86	**FBgn0145913**	19.84	20.13
FBgn0138357	13.49	13.50	**FBgn0145945**	32.76	33.05
FBgn0138389	36.33	37.87	**FBgn0145961**	13.18	13.95

FBgn0138415	22.93	24.44	**FBgn0145962**	16.12	16.41
FBgn0138416	13.76	17.36	**FBgn0145969**	14.27	14.93
FBgn0138440	35.97	33.47	**FBgn0145979**	26.54	28.12
FBgn0138446	47.26	48.54	**FBgn0146008**	11.75	12.00
FBgn0138459	12.83	13.92	**FBgn0146022**	17.92	20.03
FBgn0138464	11.91	12.07	**FBgn0146039**	12.97	20.22
FBgn0138466	17.19	20.69	**FBgn0146040**	11.63	13.28
FBgn0138487	13.08	13.08	**FBgn0146061**	29.42	31.32
FBgn0138490	18.61	18.92	**FBgn0146082**	24.17	25.14
FBgn0138492	11.17	11.40	**FBgn0146095**	33.59	37.29
FBgn0138504	12.43	13.17	**FBgn0146107**	25.03	26.53
FBgn0138509	31.76	33.57	**FBgn0146159**	32.46	33.18
FBgn0138512	27.44	34.33	**FBgn0146206**	38.97	39.90
FBgn0138523	47.99	51.73	**FBgn0146216**	43.65	47.76
FBgn0138529	59.17	60.69	**FBgn0146243**	24.67	24.84
FBgn0138537	14.81	14.87	**FBgn0146248**	15.56	17.80
FBgn0138545	32.65	32.98	**FBgn0146355**	14.85	15.12
FBgn0138557	15.88	17.46	**FBgn0146373**	82.46	89.24
FBgn0138574	24.83	25.12	**FBgn0146375**	18.28	18.62
FBgn0138578	12.34	12.68	**FBgn0146386**	12.75	13.39
FBgn0138580	13.80	13.90	**FBgn0146393**	13.16	15.67
FBgn0138582	29.78	31.01	**FBgn0146476**	11.72	11.85
FBgn0138599	19.64	20.74	**FBgn0146491**	15.28	16.15
FBgn0138626	25.53	37.18	**FBgn0146561**	20.37	22.56
FBgn0138654	23.01	23.14	**FBgn0146579**	12.21	12.66
FBgn0138666	30.11	35.07	**FBgn0146696**	24.60	24.85
FBgn0138680	30.91	32.04	**FBgn0146700**	45.04	46.29
FBgn0138710	11.84	12.88	**FBgn0146715**	16.32	16.83
FBgn0138714	15.30	17.91	**FBgn0146719**	18.76	22.15
FBgn0138740	19.88	20.36	**FBgn0146753**	11.80	11.92
FBgn0138752	35.94	41.27	**FBgn0146794**	19.15	19.84
FBgn0138754	51.32	54.26	**FBgn0146800**	18.04	18.11
FBgn0138844	47.37	47.85	**FBgn0146829**	11.05	11.95
FBgn0138916	15.01	16.35	**FBgn0146860**	12.90	13.07
FBgn0138927	128.55	131.49	**FBgn0146861**	20.38	20.50
FBgn0138940	33.61	33.62	**FBgn0146863**	56.07	63.92
FBgn0138976	13.29	14.71	**FBgn0146927**	13.81	14.75
FBgn0139010	13.86	14.09	**FBgn0146951**	11.25	11.34
FBgn0139012	17.76	18.98	**FBgn0146954**	17.88	18.32
FBgn0139033	14.40	14.79	**FBgn0146962**	21.23	23.10

FBgn0139050	14.34	14.57	**FBgn0146972**	11.91	12.05
FBgn0139091	24.46	25.41	**FBgn0146986**	45.53	48.09
FBgn0139110	16.07	18.50	**FBgn0146994**	13.37	13.56
FBgn0139116	16.20	16.19	**FBgn0147011**	11.40	11.66
FBgn0139131	28.04	28.04	**FBgn0147049**	20.64	22.53
FBgn0139167	24.82	24.95	**FBgn0147063**	23.06	23.57
FBgn0139187	87.27	89.39	**FBgn0147080**	40.16	41.10
FBgn0139207	62.65	69.40	**FBgn0147085**	27.60	27.59
FBgn0139222	26.26	27.33	**FBgn0147178**	16.09	17.86
FBgn0139237	12.76	13.09	**FBgn0147191**	14.40	14.64
FBgn0139290	13.39	14.07	**FBgn0147196**	122.05	122.65
FBgn0139362	25.13	25.56	**FBgn0147199**	31.73	30.10
FBgn0139406	12.66	13.52	**FBgn0147225**	10.98	12.39
FBgn0139422	12.65	12.85	**FBgn0147235**	11.47	11.63
FBgn0139443	11.40	11.43	**FBgn0147254**	34.95	37.69
FBgn0139458	28.13	28.06	**FBgn0147289**	48.40	48.47
FBgn0139484	17.07	22.58	**FBgn0147291**	11.03	17.77
FBgn0139523	19.13	20.25	**FBgn0147322**	26.96	27.12
FBgn0139524	23.18	24.09	**FBgn0147362**	39.90	45.16
FBgn0139555	24.83	25.85	**FBgn0147364**	13.64	15.82
FBgn0139563	12.51	13.04	**FBgn0147371**	12.26	12.25
FBgn0139577	12.86	14.55	**FBgn0147404**	11.98	12.66
FBgn0139578	14.04	15.08	**FBgn0147425**	10.93	12.10
FBgn0139591	15.35	22.76	**FBgn0147444**	81.09	82.18
FBgn0139603	14.68	14.91	**FBgn0147454**	63.61	64.55
FBgn0139607	57.51	62.07	**FBgn0147467**	32.62	38.76
FBgn0139632	18.76	18.76	**FBgn0147520**	13.99	14.11
FBgn0139678	26.39	32.86	**FBgn0147533**	59.75	59.80
FBgn0139715	10.97	11.00	**FBgn0147560**	31.54	31.54
FBgn0139736	49.47	49.45	**FBgn0147572**	42.84	43.65

BSM *D. buzzatii* lineage

Flybae gene id	LRT Results	Flybae gene id	LRT Results	Flybae gene id	LRT Results
FBgn0067231	26.55	**FBgn0137814**	18.91	**FBgn0142620**	15.48
FBgn0132833	14.14	**FBgn0137820**	13.47	**FBgn0142655**	100.35
FBgn0132834	15.66	**FBgn0137830**	36.19	**FBgn0142678**	151.36
FBgn0132854	14.54	**FBgn0137905**	12.44	**FBgn0142695**	100.85
FBgn0133004	25.06	**FBgn0137931**	20.92	**FBgn0142729**	11.11
FBgn0133171	47.00	**FBgn0137960**	28.37	**FBgn0142804**	18.02
FBgn0133176	38.30	**FBgn0137975**	157.76	**FBgn0142825**	14.85
FBgn0133225	65.62	**FBgn0138000**	24.65	**FBgn0142830**	117.28
FBgn0133236	16.87	**FBgn0138007**	51.61	**FBgn0142833**	11.49
FBgn0133252	16.49	**FBgn0138033**	31.07	**FBgn0142885**	11.37
FBgn0133266	64.31	**FBgn0138078**	75.37	**FBgn0142921**	12.95
FBgn0133272	347.82	**FBgn0138082**	11.11	**FBgn0142927**	19.05
FBgn0133282	24.74	**FBgn0138095**	30.57	**FBgn0142988**	11.41
FBgn0133302	38.57	**FBgn0138145**	35.03	**FBgn0143003**	54.70
FBgn0133309	39.32	**FBgn0138276**	33.90	**FBgn0143128**	33.71
FBgn0133319	28.82	**FBgn0138389**	62.84	**FBgn0143165**	21.05
FBgn0133515	15.18	**FBgn0138466**	48.87	**FBgn0143183**	15.69
FBgn0133565	25.84	**FBgn0138509**	22.57	**FBgn0143189**	93.48
FBgn0133587	54.95	**FBgn0138523**	80.96	**FBgn0143211**	40.20
FBgn0133615	22.23	**FBgn0138529**	35.02	**FBgn0143240**	18.96
FBgn0133663	12.85	**FBgn0138557**	19.88	**FBgn0143276**	20.65
FBgn0133670	81.83	**FBgn0138654**	56.66	**FBgn0143285**	21.35
FBgn0133733	20.87	**FBgn0138752**	115.48	**FBgn0143393**	16.98
FBgn0133743	35.39	**FBgn0138754**	67.48	**FBgn0143420**	36.94
FBgn0133754	12.75	**FBgn0138844**	19.36	**FBgn0143438**	143.84
FBgn0133765	13.48	**FBgn0138894**	10.96	**FBgn0143467**	26.40
FBgn0133776	37.58	**FBgn0138984**	21.16	**FBgn0143670**	18.79
FBgn0133848	68.97	**FBgn0139177**	21.74	**FBgn0143682**	11.88
FBgn0133863	19.96	**FBgn0139187**	33.60	**FBgn0143696**	32.84
FBgn0133926	17.82	**FBgn0139188**	17.60	**FBgn0143711**	86.66
FBgn0134005	12.59	**FBgn0139189**	11.81	**FBgn0143736**	17.96
FBgn0134159	15.26	**FBgn0139207**	177.75	**FBgn0143755**	21.81
FBgn0134235	266.47	**FBgn0139258**	12.45	**FBgn0143854**	25.58
FBgn0134254	23.82	**FBgn0139443**	15.37	**FBgn0143860**	11.58
FBgn0134268	15.30	**FBgn0139555**	63.95	**FBgn0143898**	56.32
FBgn0134345	24.17	**FBgn0139577**	16.01	**FBgn0144119**	11.47
FBgn0134351	23.06	**FBgn0139578**	19.68	**FBgn0144158**	12.92
FBgn0134358	48.08	**FBgn0139736**	71.28	**FBgn0144171**	120.99

FBgn0134393	31.24	FBgn0139763	21.27	FBgn0144363	14.04
FBgn0134468	19.04	FBgn0139771	14.83	FBgn0144371	164.45
FBgn0134484	11.77	FBgn0139866	13.92	FBgn0144383	11.78
FBgn0134537	37.28	FBgn0139890	19.82	FBgn0144402	18.38
FBgn0134552	45.97	FBgn0139927	12.61	FBgn0144414	32.00
FBgn0134565	45.11	FBgn0140021	47.30	FBgn0144482	37.55
FBgn0134605	17.68	FBgn0140045	27.74	FBgn0144499	106.24
FBgn0134629	11.03	FBgn0140066	11.18	FBgn0144526	13.26
FBgn0134666	14.33	FBgn0140094	35.10	FBgn0144666	37.06
FBgn0134700	15.24	FBgn0140104	14.89	FBgn0144681	23.45
FBgn0134773	34.83	FBgn0140166	18.25	FBgn0144691	53.41
FBgn0134797	27.88	FBgn0140252	33.17	FBgn0144698	10.94
FBgn0134800	37.03	FBgn0140391	37.46	FBgn0144753	19.95
FBgn0134830	25.90	FBgn0140397	64.51	FBgn0144762	39.34
FBgn0134911	79.48	FBgn0140422	13.85	FBgn0144787	16.89
FBgn0134920	60.87	FBgn0140434	11.13	FBgn0144796	12.36
FBgn0134937	12.22	FBgn0140544	17.83	FBgn0144861	57.32
FBgn0135018	31.27	FBgn0140586	15.77	FBgn0144884	57.61
FBgn0135023	14.05	FBgn0140587	12.85	FBgn0144886	49.75
FBgn0135037	31.82	FBgn0140827	11.10	FBgn0144894	43.81
FBgn0135076	28.79	FBgn0140920	44.37	FBgn0144950	15.79
FBgn0135080	159.73	FBgn0140945	20.07	FBgn0144955	15.79
FBgn0135227	20.01	FBgn0140958	30.23	FBgn0144970	19.96
FBgn0135228	13.41	FBgn0141105	65.83	FBgn0144984	24.59
FBgn0135231	77.22	FBgn0141113	20.43	FBgn0145025	28.58
FBgn0135323	11.38	FBgn0141193	224.06	FBgn0145052	48.46
FBgn0135435	14.76	FBgn0141205	150.27	FBgn0145093	39.35
FBgn0135464	34.43	FBgn0141278	11.15	FBgn0145115	17.06
FBgn0135584	27.56	FBgn0141287	14.55	FBgn0145116	27.28
FBgn0135627	43.79	FBgn0141295	110.39	FBgn0145156	51.84
FBgn0135693	49.94	FBgn0141300	14.00	FBgn0145175	13.22
FBgn0135751	13.66	FBgn0141362	17.48	FBgn0145247	20.39
FBgn0135786	24.13	FBgn0141373	54.25	FBgn0145275	17.37
FBgn0135789	92.32	FBgn0141406	36.74	FBgn0145375	108.53
FBgn0135883	15.58	FBgn0141410	11.43	FBgn0145467	30.28
FBgn0135941	33.34	FBgn0141448	69.46	FBgn0145527	31.78
FBgn0136002	32.60	FBgn0141463	19.78	FBgn0145656	31.04
FBgn0136039	32.45	FBgn0141523	16.52	FBgn0145701	28.90
FBgn0136061	47.59	FBgn0141603	14.69	FBgn0145748	17.91
FBgn0136304	11.52	FBgn0141677	66.41	FBgn0145753	11.41

FBgn0136313	124.48	FBgn0141681	42.77	FBgn0145837	163.77
FBgn0136316	83.17	FBgn0141704	13.66	FBgn0145846	16.75
FBgn0136318	12.76	FBgn0141766	102.65	FBgn0145851	24.92
FBgn0136354	37.39	FBgn0141810	30.12	FBgn0145884	17.62
FBgn0136406	14.90	FBgn0141887	60.33	FBgn0145902	56.76
FBgn0136426	21.96	FBgn0141920	42.04	FBgn0145908	18.22
FBgn0136428	11.16	FBgn0141946	11.84	FBgn0145913	12.90
FBgn0136441	14.75	FBgn0141999	17.48	FBgn0145945	85.07
FBgn0136544	26.94	FBgn0142008	15.04	FBgn0145969	22.66
FBgn0136604	21.28	FBgn0142012	10.86	FBgn0146022	31.48
FBgn0136663	54.61	FBgn0142013	113.65	FBgn0146095	61.99
FBgn0136689	16.77	FBgn0142061	19.85	FBgn0146155	12.27
FBgn0136810	27.31	FBgn0142105	24.43	FBgn0146159	33.57
FBgn0136917	18.60	FBgn0142109	12.18	FBgn0146311	23.74
FBgn0136984	58.27	FBgn0142135	28.97	FBgn0146373	101.51
FBgn0136989	15.00	FBgn0142169	101.39	FBgn0146375	43.18
FBgn0136990	39.57	FBgn0142192	52.48	FBgn0146456	18.48
FBgn0137041	13.85	FBgn0142194	14.46	FBgn0146552	18.87
FBgn0137159	12.82	FBgn0142195	30.68	FBgn0146647	18.36
FBgn0137173	17.96	FBgn0142210	53.62	FBgn0146715	20.62
FBgn0137291	18.55	FBgn0142223	28.29	FBgn0146719	39.80
FBgn0137320	117.39	FBgn0142275	35.68	FBgn0146829	12.25
FBgn0137378	47.22	FBgn0142322	44.76	FBgn0146860	21.13
FBgn0137398	21.23	FBgn0142345	41.13	FBgn0146904	29.44
FBgn0137401	11.64	FBgn0142379	23.19	FBgn0146954	25.93
FBgn0137416	11.80	FBgn0142408	50.57	FBgn0146955	13.32
FBgn0137464	77.88	FBgn0142414	13.35	FBgn0146962	23.78
FBgn0137467	42.91	FBgn0142475	137.27	FBgn0146986	92.76
FBgn0137469	35.00	FBgn0142503	12.20	FBgn0147085	48.13
FBgn0137471	13.80	FBgn0142513	13.31	FBgn0147185	15.50
FBgn0137484	26.12	FBgn0142537	11.60	FBgn0147196	204.74
FBgn0137504	67.11	FBgn0142538	11.20	FBgn0147254	73.88
FBgn0137605	12.86	FBgn0142551	54.53	FBgn0147289	30.85
FBgn0137613	66.77	FBgn0142553	200.93	FBgn0147371	48.43
FBgn0137631	27.58	FBgn0142556	28.36	FBgn0147374	11.51
FBgn0137634	34.01	FBgn0142590	14.97	FBgn0147444	45.44
FBgn0137643	46.68	FBgn0142591	13.76	FBgn0147454	42.40
FBgn0137673	31.71	FBgn0142598	13.13	FBgn0147533	68.6859
FBgn0137799	39.11	FBgn0142607	16.08		

BSM *D. mojavensis lineage*

Flybase Gene id	LRT Results	Flybase Gene id	LRT Results	Flybase Gene id	LRT Results
FBgn0084656	28.55	FBgn0138311	24.33	FBgn0143408	18.54
FBgn0132955	11.09	FBgn0138402	13.06	FBgn0143413	49.99
FBgn0132962	15.17	FBgn0138509	13.27	FBgn0143533	18.95
FBgn0133171	11.81	FBgn0138529	17.38	FBgn0143555	11.01
FBgn0133289	11.50	FBgn0138621	12.18	FBgn0143593	15.16
FBgn0133455	142.65	FBgn0138927	80.14	FBgn0143749	11.13
FBgn0133474	12.47	FBgn0139016	14.01	FBgn0143785	21.50
FBgn0133698	19.21	FBgn0139290	23.48	FBgn0144010	16.86
FBgn0133704	20.24	FBgn0139324	14.86	FBgn0144076	11.52
FBgn0133753	30.37	FBgn0139458	28.77	FBgn0144215	38.96
FBgn0133773	15.00	FBgn0139771	17.65	FBgn0144232	10.85
FBgn0133848	22.88	FBgn0139786	10.99	FBgn0144273	12.56
FBgn0133897	34.88	FBgn0139909	33.65	FBgn0144363	12.34
FBgn0133936	15.28	FBgn0140033	13.70	FBgn0144383	16.40
FBgn0134260	34.29	FBgn0140036	68.25	FBgn0144414	82.98
FBgn0134526	12.03	FBgn0140273	14.67	FBgn0144444	11.62
FBgn0134537	19.84	FBgn0140310	11.10	FBgn0144526	18.29
FBgn0134552	14.57	FBgn0140543	14.93	FBgn0144684	163.53
FBgn0134620	15.38	FBgn0140562	21.29	FBgn0144796	40.39
FBgn0134858	138.90	FBgn0140587	20.29	FBgn0144819	10.98
FBgn0134891	72.16	FBgn0140729	19.84	FBgn0144929	49.49
FBgn0135227	14.75	FBgn0140827	26.53	FBgn0144941	36.80
FBgn0135331	15.22	FBgn0140923	15.80	FBgn0144956	12.00
FBgn0135446	26.64	FBgn0140957	11.22	FBgn0144975	11.78
FBgn0135483	66.99	FBgn0140969	19.03	FBgn0145117	14.23
FBgn0135804	43.68	FBgn0140975	21.16	FBgn0145172	22.84
FBgn0135817	18.54	FBgn0141072	14.98	FBgn0145328	17.42
FBgn0135941	41.11	FBgn0141080	55.00	FBgn0145369	17.02
FBgn0135944	18.31	FBgn0141174	20.00	FBgn0145376	21.89
FBgn0136008	14.50	FBgn0141272	14.76	FBgn0145892	32.32
FBgn0136054	33.45	FBgn0141298	21.24	FBgn0145962	12.79
FBgn0136055	27.17	FBgn0141404	20.66	FBgn0146059	19.76
FBgn0136073	138.13	FBgn0141810	59.03	FBgn0146243	36.72
FBgn0136118	19.02	FBgn0141840	19.16	FBgn0146332	15.07
FBgn0136259	13.26	FBgn0141950	14.95	FBgn0146373	55.02
FBgn0136363	15.68	FBgn0141962	12.71	FBgn0146501	10.99
FBgn0136372	10.92	FBgn0142013	55.16	FBgn0146561	19.30
FBgn0136447	19.41	FBgn0142061	44.33	FBgn0146665	11.44

FBgn0136486	27.77	**FBgn0142086**	194.75	**FBgn0146709**	29.81
FBgn0136598	11.30	**FBgn0142102**	36.55	**FBgn0146753**	34.86
FBgn0136603	15.59	**FBgn0142104**	15.15	**FBgn0146800**	31.68
FBgn0136642	31.10	**FBgn0142135**	11.58	**FBgn0146863**	125.42
FBgn0136657	13.60	**FBgn0142236**	13.65	**FBgn0146997**	13.33
FBgn0136845	20.79	**FBgn0142366**	14.71	**FBgn0147063**	21.12
FBgn0136954	13.84	**FBgn0142429**	13.75	**FBgn0147080**	17.96
FBgn0137096	16.10	**FBgn0142436**	17.17	**FBgn0147166**	11.84
FBgn0137320	15.22	**FBgn0142459**	17.96	**FBgn0147204**	14.57
FBgn0137398	18.27	**FBgn0142496**	90.54	**FBgn0147215**	12.08
FBgn0137504	20.36	**FBgn0142618**	17.75	**FBgn0147254**	18.90
FBgn0137526	17.89	**FBgn0142688**	21.20	**FBgn0147281**	20.93
FBgn0137602	13.96	**FBgn0142786**	14.20	**FBgn0147303**	11.68
FBgn0137810	39.01	**FBgn0142892**	12.09	**FBgn0147304**	11.22
FBgn0137898	37.54	**FBgn0142995**	46.27	**FBgn0147322**	39.75
FBgn0137975	27.59	**FBgn0143063**	14.96	**FBgn0147362**	62.63
FBgn0137997	11.83	**FBgn0143137**	23.48	**FBgn0147425**	19.35
FBgn0138080	54.09	**FBgn0143279**	29.64	**FBgn0147444**	97.17
FBgn0138120	17.41	**FBgn0143338**	29.38		
FBgn0138209	28.32	**FBgn0143342**	13.95		

BSM cactophilic lineage

Flybase gene id	**LRT Results**	**Flybase gene id**	**LRT Results**	**Flybase gene id**	**LRT Results**
FBgn0084467	23.01	**FBgn0137909**	49.49	**FBgn0142477**	15.58
FBgn0084651	12.34	**FBgn0137911**	19.38	**FBgn0142503**	11.83
FBgn0085089	26.62	**FBgn0137979**	11.84	**FBgn0142533**	12.29
FBgn0085178	11.99	**FBgn0137993**	13.32	**FBgn0142547**	11.83
FBgn0132853	25.29	**FBgn0138012**	16.45	**FBgn0142551**	17.48
FBgn0132868	11.61	**FBgn0138016**	18.79	**FBgn0142553**	27.78
FBgn0132897	40.23	**FBgn0138030**	13.69	**FBgn0142569**	13.46
FBgn0132940	21.04	**FBgn0138060**	11.15	**FBgn0142598**	29.02
FBgn0132962	16.02	**FBgn0138066**	48.19	**FBgn0142625**	18.60
FBgn0133074	13.33	**FBgn0138099**	12.09	**FBgn0142652**	11.36
FBgn0133199	26.42	**FBgn0138139**	12.73	**FBgn0142654**	12.62
FBgn0133207	15.47	**FBgn0138162**	14.68	**FBgn0142710**	26.36
FBgn0133233	33.13	**FBgn0138484**	13.44	**FBgn0142712**	23.85

FBgn0133289	12.47	FBgn0138509	20.90	FBgn0142713	11.82
FBgn0133296	13.13	FBgn0138522	24.91	FBgn0142721	29.86
FBgn0133409	13.95	FBgn0138557	14.77	FBgn0142754	11.52
FBgn0133467	15.36	FBgn0138559	13.32	FBgn0142780	14.67
FBgn0133476	12.49	FBgn0138593	17.14	FBgn0142785	22.59
FBgn0133530	14.12	FBgn0138630	12.77	FBgn0142834	28.42
FBgn0133576	19.05	FBgn0138631	13.14	FBgn0142845	22.10
FBgn0133622	14.67	FBgn0138654	11.23	FBgn0142890	14.61
FBgn0133717	15.52	FBgn0138655	12.34	FBgn0142893	12.24
FBgn0133727	17.19	FBgn0138666	39.72	FBgn0142932	17.62
FBgn0133728	19.77	FBgn0138720	12.99	FBgn0142976	21.69
FBgn0133744	26.81	FBgn0138739	12.19	FBgn0142985	29.97
FBgn0133753	22.96	FBgn0138755	13.55	FBgn0142987	16.76
FBgn0133776	19.57	FBgn0138774	15.38	FBgn0143003	11.56
FBgn0133789	39.36	FBgn0138838	18.59	FBgn0143020	13.78
FBgn0133803	19.24	FBgn0138844	12.91	FBgn0143033	11.11
FBgn0133809	16.90	FBgn0138873	16.70	FBgn0143112	38.12
FBgn0133813	16.63	FBgn0138982	22.97	FBgn0143127	11.17
FBgn0133818	11.16	FBgn0138986	14.10	FBgn0143170	12.86
FBgn0133835	31.74	FBgn0138994	11.97	FBgn0143189	28.00
FBgn0133848	16.77	FBgn0139007	15.23	FBgn0143306	12.93
FBgn0133866	63.10	FBgn0139012	21.06	FBgn0143314	35.18
FBgn0133917	16.73	FBgn0139020	19.84	FBgn0143320	14.50
FBgn0133963	24.65	FBgn0139033	11.86	FBgn0143402	11.33
FBgn0134033	21.26	FBgn0139056	12.59	FBgn0143489	20.97
FBgn0134056	15.38	FBgn0139063	14.98	FBgn0143490	22.07
FBgn0134069	18.70	FBgn0139067	19.62	FBgn0143524	15.79
FBgn0134077	15.35	FBgn0139069	13.90	FBgn0143593	40.73
FBgn0134099	14.65	FBgn0139114	17.09	FBgn0143670	21.36
FBgn0134167	16.05	FBgn0139174	10.99	FBgn0143709	11.62
FBgn0134299	30.55	FBgn0139187	21.60	FBgn0143766	13.26
FBgn0134355	15.78	FBgn0139206	10.84	FBgn0143873	11.30
FBgn0134418	11.88	FBgn0139207	47.47	FBgn0143934	21.79
FBgn0134484	11.18	FBgn0139210	12.98	FBgn0143996	13.77
FBgn0134505	12.65	FBgn0139216	17.75	FBgn0144035	20.80
FBgn0134537	18.05	FBgn0139279	19.08	FBgn0144119	12.15
FBgn0134572	14.38	FBgn0139286	11.70	FBgn0144160	35.15
FBgn0134603	13.78	FBgn0139294	15.16	FBgn0144177	11.26
FBgn0134605	25.93	FBgn0139314	11.84	FBgn0144211	11.15
FBgn0134620	14.89	FBgn0139338	22.12	FBgn0144232	23.80

FBgn0134649	13.29	FBgn0139346	20.32	FBgn0144245	30.11
FBgn0134691	12.99	FBgn0139355	11.07	FBgn0144275	21.87
FBgn0134707	19.71	FBgn0139379	16.74	FBgn0144310	15.38
FBgn0134776	11.07	FBgn0139458	15.71	FBgn0144324	27.78
FBgn0134804	19.85	FBgn0139487	12.58	FBgn0144363	46.36
FBgn0134828	23.49	FBgn0139519	23.11	FBgn0144386	17.60
FBgn0134865	11.10	FBgn0139547	17.37	FBgn0144407	15.22
FBgn0134883	15.27	FBgn0139581	58.87	FBgn0144465	12.47
FBgn0134890	12.41	FBgn0139588	19.27	FBgn0144495	11.28
FBgn0134920	14.02	FBgn0139590	16.41	FBgn0144505	14.88
FBgn0134933	11.72	FBgn0139641	17.99	FBgn0144506	17.41
FBgn0134942	32.88	FBgn0139721	16.98	FBgn0144522	16.81
FBgn0135042	15.56	FBgn0139737	12.41	FBgn0144647	13.11
FBgn0135043	15.46	FBgn0139848	13.40	FBgn0144659	18.59
FBgn0135083	13.36	FBgn0139880	12.91	FBgn0144701	18.99
FBgn0135097	16.91	FBgn0139909	25.66	FBgn0144708	18.66
FBgn0135187	56.16	FBgn0139912	13.25	FBgn0144753	27.36
FBgn0135210	13.29	FBgn0139929	12.22	FBgn0144757	39.68
FBgn0135227	14.28	FBgn0139930	13.87	FBgn0144770	14.06
FBgn0135228	32.62	FBgn0139931	13.84	FBgn0144805	15.20
FBgn0135231	11.28	FBgn0139935	12.20	FBgn0144850	28.00
FBgn0135255	14.84	FBgn0139947	19.48	FBgn0144950	62.22
FBgn0135298	12.69	FBgn0139948	12.24	FBgn0144975	12.43
FBgn0135305	15.04	FBgn0139981	12.37	FBgn0145072	21.39
FBgn0135319	14.66	FBgn0140001	10.85	FBgn0145093	10.97
FBgn0135324	16.43	FBgn0140036	23.17	FBgn0145133	14.78
FBgn0135327	14.54	FBgn0140048	13.09	FBgn0145176	12.69
FBgn0135329	22.88	FBgn0140063	13.59	FBgn0145239	19.11
FBgn0135334	13.67	FBgn0140073	12.62	FBgn0145250	18.93
FBgn0135391	15.86	FBgn0140074	14.22	FBgn0145262	11.12
FBgn0135435	12.54	FBgn0140136	143.97	FBgn0145266	14.39
FBgn0135440	14.19	FBgn0140159	15.54	FBgn0145280	11.23
FBgn0135448	13.61	FBgn0140167	30.07	FBgn0145305	13.00
FBgn0135462	14.12	FBgn0140235	21.79	FBgn0145332	27.59
FBgn0135465	51.83	FBgn0140237	11.18	FBgn0145453	24.56
FBgn0135555	38.87	FBgn0140318	21.38	FBgn0145637	10.86
FBgn0135574	15.86	FBgn0140332	26.97	FBgn0145667	11.19
FBgn0135584	10.86	FBgn0140439	23.51	FBgn0145688	11.44
FBgn0135590	23.98	FBgn0140514	16.85	FBgn0145700	15.76
FBgn0135622	12.04	FBgn0140519	21.58	FBgn0145796	12.51

FBgn0135624	13.26	**FBgn0140535**	12.31	**FBgn0145835**	12.19
FBgn0135629	28.40	**FBgn0140587**	13.53	**FBgn0145879**	16.34
FBgn0135647	22.22	**FBgn0140588**	14.40	**FBgn0145960**	11.90
FBgn0135656	26.88	**FBgn0140637**	24.61	**FBgn0146008**	15.42
FBgn0135657	22.14	**FBgn0140643**	10.90	**FBgn0146028**	11.86
FBgn0135686	18.74	**FBgn0140662**	13.26	**FBgn0146033**	15.04
FBgn0135714	29.72	**FBgn0140691**	19.60	**FBgn0146036**	14.10
FBgn0135716	12.05	**FBgn0140710**	20.99	**FBgn0146040**	21.40
FBgn0135747	15.82	**FBgn0140713**	21.01	**FBgn0146048**	14.42
FBgn0135764	22.54	**FBgn0140767**	18.12	**FBgn0146061**	15.22
FBgn0135837	41.41	**FBgn0140771**	38.59	**FBgn0146082**	20.38
FBgn0135883	23.60	**FBgn0140857**	12.21	**FBgn0146095**	31.63
FBgn0135941	53.40	**FBgn0140928**	17.93	**FBgn0146104**	19.25
FBgn0136028	15.12	**FBgn0140969**	12.19	**FBgn0146112**	14.53
FBgn0136049	15.98	**FBgn0141009**	12.39	**FBgn0146117**	13.44
FBgn0136158	11.69	**FBgn0141072**	13.72	**FBgn0146140**	19.71
FBgn0136180	10.86	**FBgn0141080**	11.77	**FBgn0146185**	15.46
FBgn0136181	20.57	**FBgn0141096**	25.23	**FBgn0146243**	11.43
FBgn0136252	10.99	**FBgn0141179**	10.94	**FBgn0146248**	26.76
FBgn0136373	16.04	**FBgn0141202**	19.12	**FBgn0146255**	13.82
FBgn0136394	15.99	**FBgn0141304**	17.21	**FBgn0146317**	15.72
FBgn0136434	15.88	**FBgn0141318**	23.88	**FBgn0146327**	17.14
FBgn0136460	17.28	**FBgn0141489**	11.95	**FBgn0146366**	25.74
FBgn0136468	12.36	**FBgn0141510**	11.78	**FBgn0146376**	14.45
FBgn0136512	13.86	**FBgn0141593**	33.07	**FBgn0146420**	11.21
FBgn0136544	20.58	**FBgn0141654**	27.58	**FBgn0146556**	15.46
FBgn0136571	19.72	**FBgn0141689**	12.01	**FBgn0146580**	21.21
FBgn0136691	14.43	**FBgn0141699**	29.69	**FBgn0146593**	12.77
FBgn0136693	12.62	**FBgn0141727**	21.43	**FBgn0146600**	11.58
FBgn0136724	15.50	**FBgn0141734**	14.61	**FBgn0146622**	20.00
FBgn0136785	11.24	**FBgn0141742**	16.10	**FBgn0146665**	14.01
FBgn0136802	40.00	**FBgn0141747**	11.33	**FBgn0146729**	17.00
FBgn0136807	27.69	**FBgn0141766**	20.98	**FBgn0146792**	13.43
FBgn0136852	30.43	**FBgn0141808**	18.79	**FBgn0146814**	12.78
FBgn0136873	31.07	**FBgn0141810**	12.76	**FBgn0146841**	17.47
FBgn0136943	11.39	**FBgn0141927**	16.46	**FBgn0146843**	12.33
FBgn0136954	16.26	**FBgn0141995**	23.17	**FBgn0146894**	17.47
FBgn0137015	28.72	**FBgn0142013**	23.27	**FBgn0146946**	18.55
FBgn0137018	14.30	**FBgn0142086**	63.29	**FBgn0146968**	15.83
FBgn0137134	12.05	**FBgn0142102**	21.24	**FBgn0146972**	16.98

FBgn0137168	14.45	**FBgn0142103**	22.78	**FBgn0146986**	18.91
FBgn0137218	11.80	**FBgn0142112**	10.85	**FBgn0147018**	27.45
FBgn0137242	15.58	**FBgn0142120**	13.61	**FBgn0147027**	103.96
FBgn0137289	13.88	**FBgn0142124**	11.71	**FBgn0147047**	12.00
FBgn0137315	18.94	**FBgn0142156**	18.62	**FBgn0147049**	15.90
FBgn0137418	14.76	**FBgn0142160**	13.37	**FBgn0147082**	19.02
FBgn0137428	18.41	**FBgn0142228**	25.73	**FBgn0147108**	24.82
FBgn0137450	12.11	**FBgn0142264**	15.10	**FBgn0147131**	15.95
FBgn0137471	16.78	**FBgn0142267**	30.66	**FBgn0147203**	13.07
FBgn0137553	20.66	**FBgn0142282**	27.17	**FBgn0147362**	33.39
FBgn0137582	20.13	**FBgn0142312**	11.67	**FBgn0147401**	11.51
FBgn0137602	24.14	**FBgn0142333**	15.27	**FBgn0147412**	14.40
FBgn0137607	28.78	**FBgn0142348**	17.19	**FBgn0147440**	14.04
FBgn0137624	12.99	**FBgn0142394**	23.53	**FBgn0147444**	21.08
FBgn0137728	16.83	**FBgn0142400**	15.78	**FBgn0147514**	11.52
FBgn0137799	12.48	**FBgn0142406**	16.64	**FBgn0147533**	82.56
FBgn0137801	26.52	**FBgn0142408**	19.86	**FBgn0147534**	19.98
FBgn0137821	11.96	**FBgn0142413**	22.91	**FBgn0147543**	11.61
FBgn0137831	11.86	**FBgn0142424**	31.08	**FBgn0147547**	15.11
FBgn0137882	20.89	**FBgn0142433**	14.09		

Orphan genes

Flybase gene id	**dn**	**ds**	**ω**	*D. buz.* **protein length (aa)**	*D. moj.* **protein length (aa)**	*D. moj.* **scaffold**	**# exons** *D. buz.*	**# exons** *D. moj.*
FBgn0084252	0.5025	0.7416	0.6776	62	67	6496	1	1
FBgn0132808	0.0576	0.1068	0.5388	93	97	6540	1	1
FBgn0133043	0.1528	0.6435	0.2374	670	756	6540	4	4
FBgn0133050	0.169	0.4093	0.4128	137	139	6540	1	1
FBgn0133106	0.0787	0.0228	3.4527	57	59	6540	1	1
FBgn0133329	0.0869	0.1815	0.4788	53	64	6540	1	1
FBgn0133460	0.0633	0.2846	0.2225	114	116	6540	1	1
FBgn0133573	0.1291	0.3913	0.3298	74	74	6540	2	2
FBgn0133669	0.0000	0.0000	0.4547	66	76	6308	1	2
FBgn0133712	0.1311	0.1522	0.8614	68	69	6308	1	1
FBgn0133791	0.0003	0.3447	0.0010	69	64	6308	2	1
FBgn0133924	0.2180	0.6094	0.3576	239	240	6308	2	2
FBgn0134143	0.0376	0.3273	0.1148	66	66	6500	1	1
FBgn0134228	0.3442	0.4402	0.7819	80	87	6680	2	2

FBgn0134265	0.2052	0.8390	0.2446	77	77	6680	1	1
FBgn0134411	0.1688	0.0610	2.7652	128	130	6680	1	1
FBgn0134416	0.1271	0.5010	0.2537	183	186	6680	1	1
FBgn0134425	0.1348	0.3000	0.4494	99	101	6680	1	1
FBgn0134449	0.1612	0.5036	0.3201	112	99	6680	2	2
FBgn0134461	0.3056	0.6452	0.4737	102	100	6680	1	1
FBgn0134529	0.2551	0.5560	0.4588	79	79	6680	2	2
FBgn0134546	0.0625	0.4643	0.1347	169	161	6680	2	2
FBgn0134618	0.3178	0.5893	0.5393	138	164	6680	1	1
FBgn0134694	0.0778	0.2400	0.3243	108	112	6680	2	2
FBgn0134745	0.0190	0.3515	0.0542	62	62	6680	2	2
FBgn0135138	0.3875	0.2589	1.4966	56	56	6680	1	1
FBgn0135403	0.0217	0.1648	0.1318	84	79	6680	1	1
FBgn0135405	0.0328	0.1447	0.2266	102	101	6680	2	2
FBgn0135406	0.0139	0.1798	0.0770	75	75	6680	2	2
FBgn0135417	0.0933	0.4815	0.1938	205	200	6680	2	2
FBgn0135424	0.1246	0.3913	0.3184	96	84	6680	1	1
FBgn0135497	0.0272	0.2083	0.1308	91	88	6680	1	2
FBgn0135977	0.0061	0.0365	0.1663	75	81	6680	1	1
FBgn0136040	0.1082	0.1693	0.6393	48	53	6680	1	1
FBgn0136167	0.2655	0.5031	0.5277	71	71	6680	2	2
FBgn0136408	0.4980	0.8945	0.5567	90	90	6680	1	1
FBgn0136630	0.0341	0.0936	0.3647	47	54	6680	1	1
FBgn0136903	0.1676	0.4566	0.3671	67	72	6500	1	1
FBgn0137078	0.4446	0.4591	0.9684	81	97	6482	1	1
FBgn0137510	0.1357	0.1572	0.8630	70	80	6473	1	1
FBgn0137563	0.0416	0.1009	0.4119	93	92	6500	1	1
FBgn0137601	0.1653	0.1329	1.2439	60	71	6473	1	1
FBgn0137769	0.1267	0.4284	0.2958	111	111	6473	1	1
FBgn0137782	0.0838	0.1073	0.7811	91	99	6473	2	2
FBgn0137837	0.4290	0.5889	0.7285	159	188	6473	3	2
FBgn0137880	0.0311	0.107	0.2905	77	74	6473	1	1
FBgn0138207	0.2858	0.6694	0.4269	86	86	6473	1	1
FBgn0138211	0.2808	0.6656	0.4219	56	65	6564	1	1
FBgn0138246	0.2160	0.3220	0.6710	94	99	6473	2	1
FBgn0138354	0.1326	0.2037	0.6513	133	135	6500	1	1
FBgn0138370	0.1533	0.6158	0.2489	62	63	6473	2	2
FBgn0138545	0.2815	0.1844	1.5268	78	80	6473	1	1
FBgn0138653	0.0368	0.2407	0.153	112	109	6473	2	2
FBgn0138709	0.3455	0.5356	0.6451	118	121	6500	1	1
FBgn0138769	0.1766	0.0599	2.9485	46	52	1552	1	1
FBgn0138957	0.1297	0.2017	0.643	66	67	6328	1	1
FBgn0139019	0.4062	0.6552	0.6199	115	109	6328	1	1

FBgn0139140	0.2082	0.6932	0.3004	84	84	6328	2	2
FBgn0139154	0.1110	0.4968	0.2235	75	76	6328	2	2
FBgn0139176	0.0489	0.4993	0.098	164	180	6328	3	2
FBgn0139241	0.2243	0.2236	1.0029	61	69	6500	1	2
FBgn0139272	0.0116	0.2826	0.0412	121	121	6654	2	2
FBgn0139281	0.0687	0.2951	0.2329	108	110	6654	2	2
FBgn0139579	0.0569	0.3274	0.1737	209	207	6654	3	3
FBgn0139711	0.0001	0.1263	0.001	34	34	6496	1	1
FBgn0140039	0.1013	0.4171	0.243	60	60	6500	1	1
FBgn0140234	0.1902	0.4059	0.4686	112	114	6500	2	2
FBgn0140674	0.2859	0.7016	0.4075	77	80	6500	1	1
FBgn0140727	0.1053	0.12	0.878	137	127	6500	2	3
FBgn0140953	0.1324	0.3114	0.4253	86	96	6500	1	1
FBgn0140982	0.0597	0.2182	0.2738	65	65	6500	2	2
FBgn0141168	0.1436	0.7842	0.1832	75	75	6496	1	1
FBgn0141206	0.1583	0.2243	0.7055	68	72	6496	1	1
FBgn0141320	0.1477	0.1814	0.8141	203	219	6496	1	2
FBgn0141330	0.0348	0.0345	1.0089	70	72	6496	1	1
FBgn0141408	0.129	0.2108	0.6121	114	112	6496	1	1
FBgn0141633	0.0676	0.1146	0.5898	54	55	6496	1	1
FBgn0141650	0.0347	0.2034	0.1707	108	105	6496	2	2
FBgn0141774	0.1369	0.4015	0.341	85	84	6496	2	2
FBgn0141919	0.1088	0.3221	0.3378	166	175	6496	1	1
FBgn0142106	0.1174	0.2543	0.4619	58	58	6496	1	1
FBgn0142187	0.1551	0.448	0.3463	153	157	6496	1	1
FBgn0142570	0.2199	0.7697	0.2857	320	337	6496	1	1
FBgn0142574	0.248	0.6124	0.405	304	296	6496	1	1
FBgn0142575	0.3313	0.6032	0.5492	215	217	6496	2	1
FBgn0142632	0.1339	0.4996	0.268	146	166	6496	1	1
FBgn0142635	0.2151	0.6067	0.3545	263	262	6496	1	1
FBgn0142669	0.0529	0.0383	1.3813	56	61	6496	1	1
FBgn0142922	0.0838	0.0605	1.3848	60	61	6496	1	1
FBgn0143049	0.2772	0.5803	0.4777	270	276	6500	2	2
FBgn0143114	0.0211	0.0876	0.2408	55	60	6496	1	2
FBgn0143727	0.1116	0.3557	0.3137	228	214	6496	2	2
FBgn0143728	0.1289	0.2642	0.4879	179	198	6496	2	1
FBgn0143730	0.2504	0.6194	0.4042	77	82	6496	1	1
FBgn0143746	0.0534	0.0969	0.5515	63	69	6496	1	1
FBgn0143776	0.0436	0.0509	0.857	70	74	6496	1	1
FBgn0143834	0.104	0.4012	0.2593	70	82	6496	1	1
FBgn0144124	0.2097	0.61	0.3438	77	76	6500	1	1
FBgn0144621	0.1738	0.2918	0.5955	43	53	6510	1	1
FBgn0144673	0.0186	0.1687	0.1104	95	101	6540	2	3

FBgn0144682	0.0211	0.1355	0.1559	73	68	6540	1	1
FBgn0144907	0.1407	0.3043	0.4624	88	87	6540	2	2
FBgn0145065	0.0782	0.1092	0.716	91	98	6540	1	1
FBgn0145390	0.097	0.0625	1.5517	82	85	6500	1	1
FBgn0146213	0.0382	0.2139	0.1788	48	52	6540	1	1
FBgn0146224	0.1032	0.8046	0.1283	82	93	6540	1	1
FBgn0146316	0.0654	0.0603	1.0857	45	56	6540	1	1
FBgn0146405	0.1429	0.0226	6.3091	71	71	6540	2	2
FBgn0146422	0.163	0.6491	0.2511	159	194	6540	1	2
FBgn0146487	0.1084	0.2445	0.4435	43	52	6500	1	1
FBgn0146771	0.1093	0.1544	0.7083	131	135	6540	1	1
FBgn0146861	0.1308	0.1232	1.0616	129	126	6540	1	1
FBgn0147026	0.1559	0.1923	0.8105	126	131	6540	1	1
FBgn0147508	0.1718	0.5965	0.288	150	162	6540	1	1
FBgn0147510	0.3429	0.657	0.5219	125	132	6540	3	1
FBgn0147520	0.2812	0.6444	0.4363	61	61	6540	1	1
FBgn0147538	0.1104	0.2038	0.5417	88	94	6540	1	1

Table S5. Summary of sequencing data.

Strain	Platform	Library		Mean insert size (kb)	#Raw reads	#Filtered reads	Mean read length (bp)	Expected coverage
		Type	# plates (454) or lanes (Illumina)					
st-1	454	Shotgun	3	-	4219296	3857039	335.23	8x
		PE	2	6-8	2501837	1691215	304.92	3x
	Sanger	BES	-	150	2304	1799	698.2	~0.01x
	Illumina	PE	4	0.5	447062156	114499279	106.3	76x
		MP	1	7.5	41846306	19292893	97.8	12x

Table S6. Three assembly stages of *D. buzzatii* st-1 genome.

Stage	Input	# Scaffold (> 3 kb)	# putative chimerics (split)	N50 scaffold index	Max scaffold size	#N's	#gaps
De novo Pre-assembly (Newbler)	All 454 + BES + 1 library Illumina short PE	2306	3 (inter-chromosomal)	38	14579794	18060254	-
Scaffolding (SSPACE)	Pre-assembled scaffolds + MP libray	815	3 (inter-chromosomal)	29	16289485	18991294	13409
Gapfilling (GapFiller)	Scaffolds + 3 Illumina short PE	818	8 (intra-chromosomal)	30	16306990	14974169	11462

Table S7. Base composition by genome features.

Base composition	Genome	Genes	Exons
AT	55.81 %	54.24 %	48.17 %
GC	34.92 %	42.00 %	51.83 %
N	9.27 %	3.76 %	0.004 %
Total bases	161490851	42433860	20364820
Fraction	100 %	26.28 %	12.61 %

Table S8. Quality control of freeze 1 assembly using sequenced BACs.

BAC	Chromosome	Length (bp)	Unambiguous bp covered (%)	Average identity (%)	Matched scaffolds		
					Number of scaffolds	Freeze 1 scaffold id.	Aligned blocks
1B03	2	258840	97.29	99.96	1	scaffold1	8
1N19	2	138724	98.97	99.92	1	scaffold1	8
20O19	2	143293	98.24	100	1	scaffold1	5
40C11	2	132938	100.00	99.88	1	scaffold2	6
5H14	2	124024	93.31	99.97	1	scaffold5	12

Table S9. Assembly error rate inferred by mapping genomic and RNAseq reads to Freeze 1 sequence. The overall error rate was computed using a coverage threshold of 4 aligned reads per position.

	Genomic reads mapping		**RNAseq male adults reads mapping**	
	# Putative assembly sequence errors	Error rate	# Putative assembly sequence errors	Error rate
No coverage threshold	182598	0.00125	71499	0.00153
Coverage threshold ≥4	68898	0.00047	19042	0.00062

Table S10. Polymorphism rate estimation by mapping Illumina reads to Freeze 1 assembly.

	Gapfiller reads mapping	
	# Polymorphic positions	Polymorphism rate
No coverage threshold	148772	0.00102
Coverage threshold ≥4	141648	0.000972

Table S11. Optical Density (IOD) and genomic size estimation.

	IOD		Genome size (pg)		Genome size (Mb)	
Species	**j19**	**st1**	**j19**	**st1**	**j19**	**st1**
D. buzzatii	96.56	467.03	0.149	0.156	146	153
D. mojavensis	128.27	591.20	0.198[a]	0.198[a]	194[b]	194[b]

[a] Estimated by dividing genome size in Mb by 978 Mb/pg.

[b] Total assembly size (Drosophila 12 Genomes Consortium).

Table S12. RNAseq reads per sample

Sample	Yield (Mb)	Reads (x 10^6)	% bp Q ≥ 30	Mean Quality Score	Paired filtered reads (x 10^6)	Reads used by TopHat (x 10^6)	Reads yielding unique hits (x 10^6)
Embryos	9051	89.6	87.05	34.26	68.5	68.4	50.9
Larvae	6084	60.2	87.51	34.42	46.5	46.4	30.3
Pupae	7070	69.9	86.13	33.94	52.4	52.4	45.8
Female adults	8658	85.7	85.77	33.85	63.6	63.6	55.8
Male adults	7382	73.1	87.03	34.25	55.9	55.8	44.8
Total	38245	378.5	86.70	34.14	286.9	286.6	227.6

Table S13. Matrix of correlation coefficients (below diagonal) and p-values (above diagonal) from pairwise correlation tests between each of the genomic factors included in the three linear models.

	Type	**Recomb**	**State**	**Length**	**Exons**	**Breadth**	**Max. expression**
Type	1	0.3107	2.20e-16**	0.3481	0.0016**	0.5135	0.3459
Recomb	0.0107	1	2.2e-16**	0.6195	0.852	0.1973	0.8744
State	-0.1194	-0.2511	1	0.2392	0.4604	0.0266	0.0368*
Length	-0.0099	0.0052	0.0124	1	2.20e-16**	2.149e-07**	6.20e-14**
Exons	-0.0333	0.0020	0.0078	0.6719	1	2.2e-16**	4.59e-06**
Breadth	0.0069	0.0136	-0.0233	0.0546	0.0872	1	7.50e-08**
Max. expression	-0.0099	0.0017	-0.0220	-0.0789	-0.0482	0.0566	1

** Extremely significant (p-values < 0.01)

* Moderately significant (0.01< p-values <0.05)

SUPPLEMENTAL FIGURES

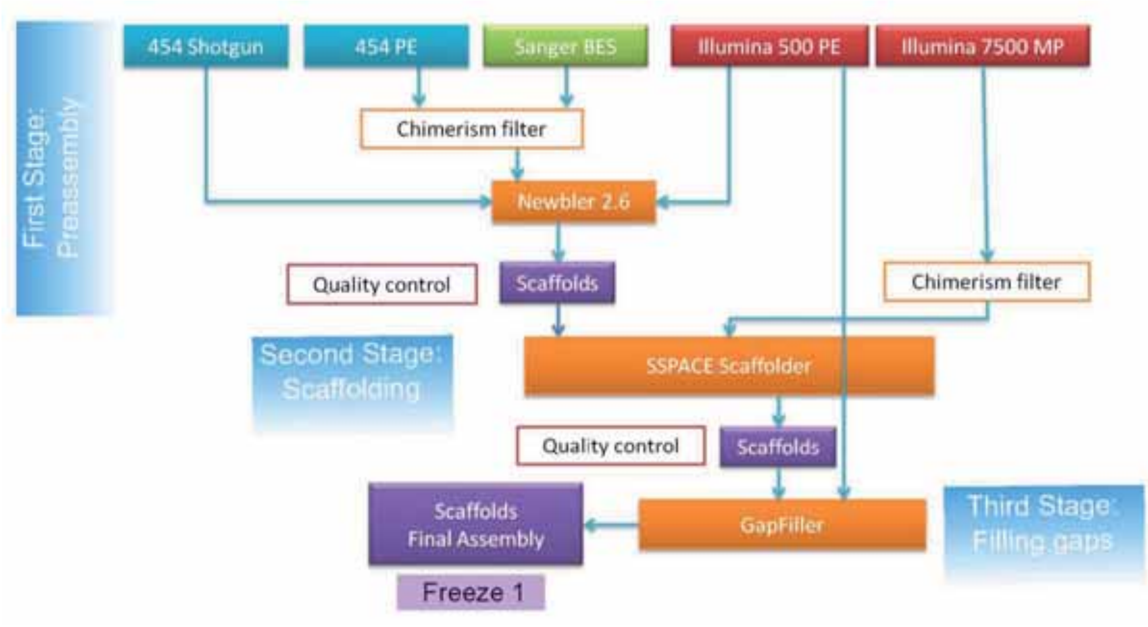


Figure S1. Assembly pipeline followed for st-1 *D. buzzatii* genome.

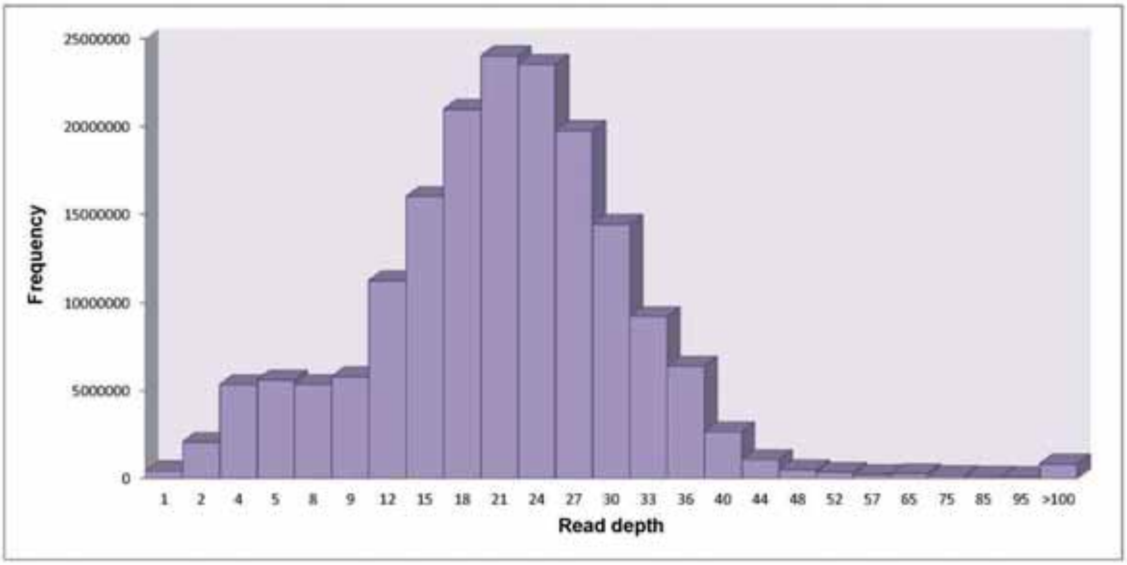


Figure S2. Read depth histogram of *D. buzzatii* preassembly.

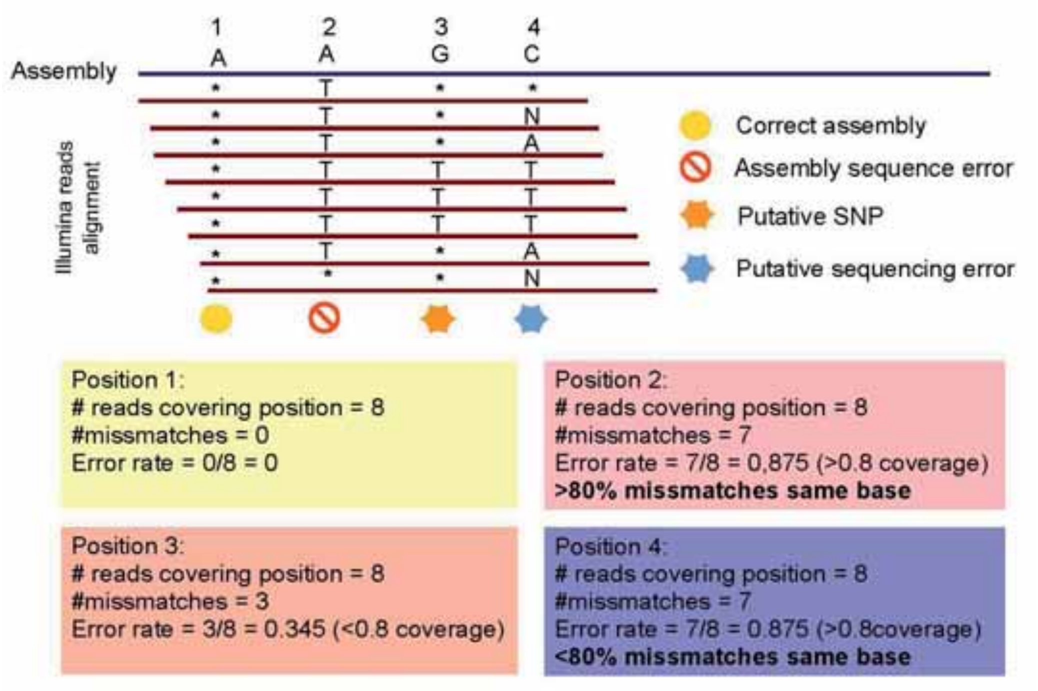

1
2
3
4
A
A
G
C
Assembly
Illumina reads alignment
T
T
T
T
T
T
T
N
A
T
T
T
A
N
Correct assembly
Assembly sequence error
Putative SNP
Putative sequencing error
Position 1:
# reads covering position = 8
#missmatches = 0
Error rate = 0/8 = 0
Position 2:
# reads covering position = 8
#missmatches = 7
Error rate = 7/8 = 0,875 (>0.8 coverage)
>80% missmatches same base
Position 3:
# reads covering position = 8
#missmatches = 3
Error rate = 3/8 = 0.345 (<0.8 coverage)
Position 4:
# reads covering position = 8
#missmatches = 7
Error rate = 7/8 = 0.875 (>0.8coverage)
<80% missmatches same base

Figure S3. Algorithm designed to track putative sequence errors and polymorphic sites in freeze 1 assembly. Four different positions are described according to the results obtained by aligning Illumina reads. Positions with an error rate < 0.8 are considered correct positions (1). Positions in which more than 80% of the aligned reads having the same base do not match the assembly are pinpointing assembly errors (2). Polymorphic positions are detected if less than 80% but more than 20% of the aligned reads do not match the assembly and have the same base (3). Putative sequencing errors are detected when more than 80% of the bases do not match the assembly and they have random bases in the same position. This last category was not further analyzed.

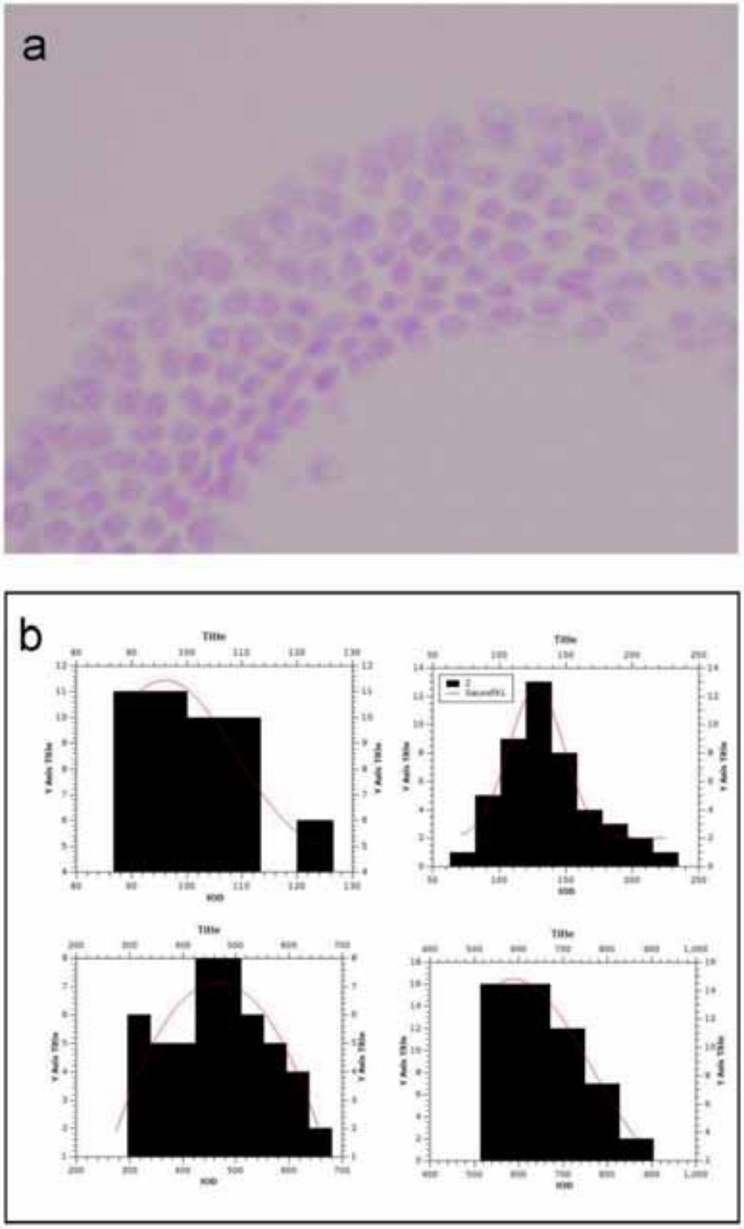


Figure S4. Genome size quantification of *D. buzzatii* st-1 and j-19 strains using IOD. Testicular cells analyzed from *D. buzzatii* st-1 strain (a) and normal distribution profiles that best fit to the IOD histogram representations (b). Fifty cells from each group were analyzed.

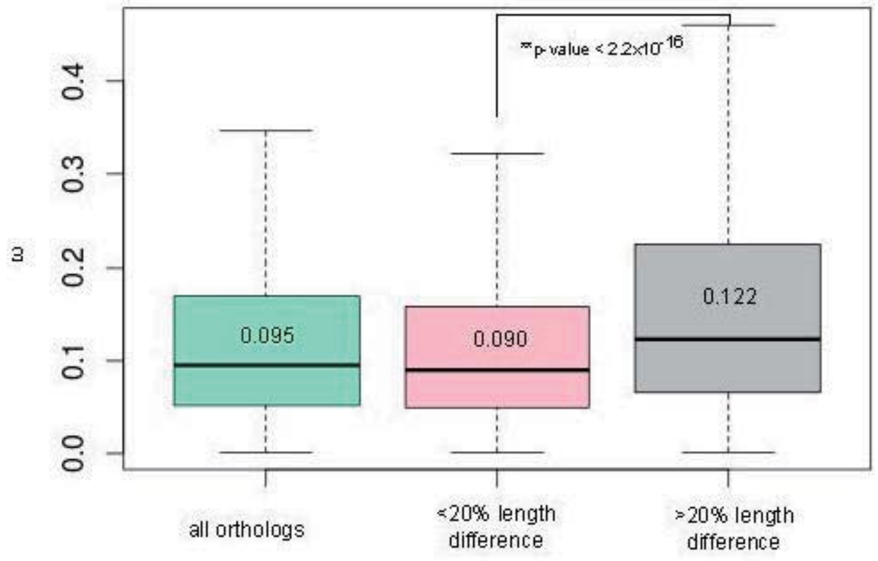


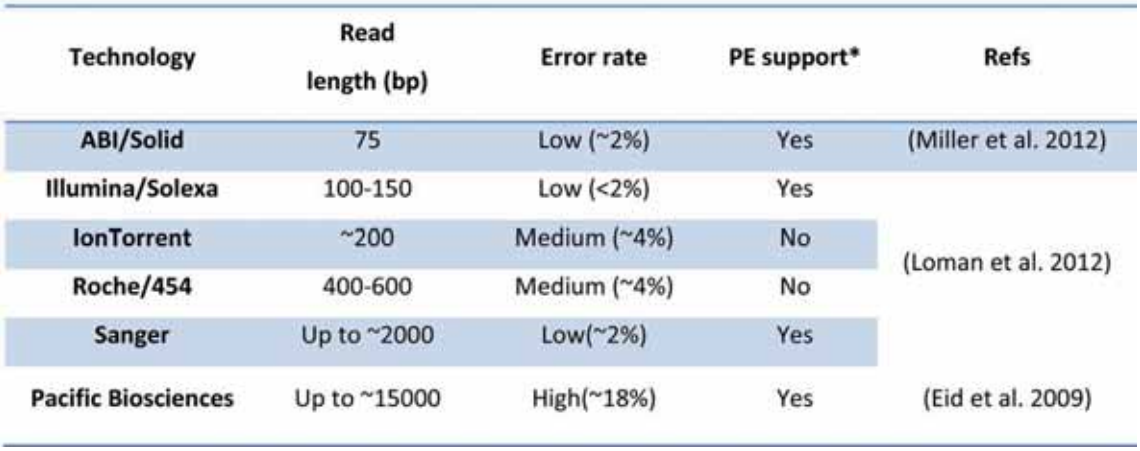
Figure S5. ω distribution of orthologs between *D. buzzatii* and *D. mojavensis*. Orthologous pairs that show a length difference higher than 20% increase the ω median of all gene set.

5. DISCUSSION

5.1 Facing a *de novo* genome assembly

Determining the complete DNA sequence of a genome has become a recurrent task in many laboratories during the last decade. The development of new sequencing technologies makes it more feasible than ever to obtain millions of DNA reads in a relatively short period of time at a reasonable cost (Table 3).

TABLE 3. Comparison of different sequencing platforms.



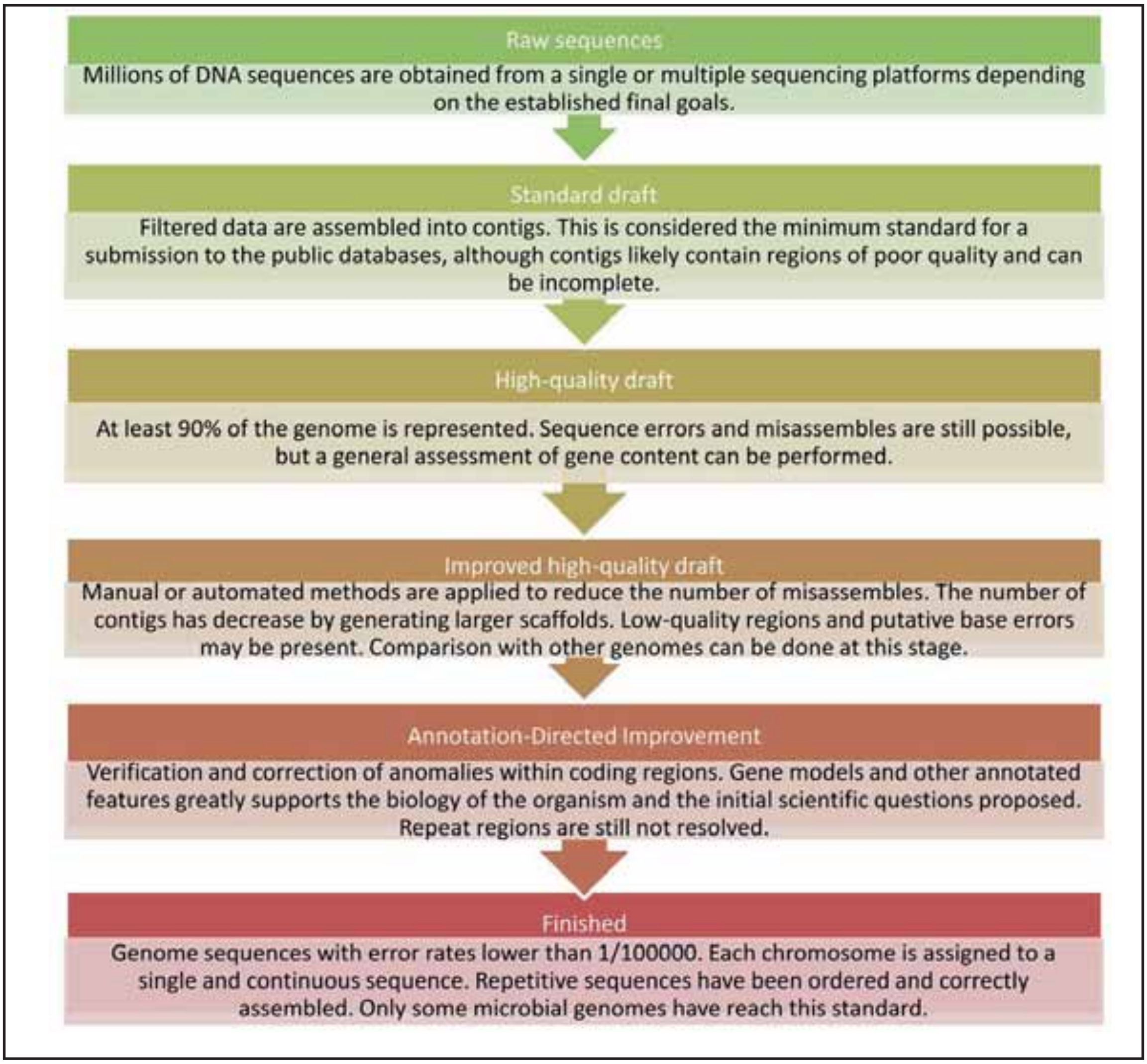
Technology	Read length (bp)	Error rate	PE support*	Refs
ABI/Solid	75	Low (~2%)	Yes	(Miller et al. 2012)
Illumina/Solexa	100-150	Low (<2%)	Yes	(Loman et al. 2012)
IonTorrent	~200	Medium (~4%)	No	
Roche/454	400-600	Medium (~4%)	No	
Sanger	Up to ~2000	Low(~2%)	Yes	
Pacific Biosciences	Up to ~15000	High(~18%)	Yes	(Eid et al. 2009)

**Paired-end support refers to the platform's ability to generate paired-end reads natively. Potentially all sequencing technologies can be used to sequence paired-end libraries obtained by the circularization of long DNA fragments.*

However, to start a new genome project requires facing one of the most complex computational and technical challenges in modern Biology. The abundant levels of repetitive regions in most eukaryotic genomes generate puzzling ambiguities that current short-read assembler software are not able to resolve (Treangen and Salzberg 2012), representing the major obstacle to perform accurate genome analysis. As a result, the increasing number of sequenced genomes has been regrettably accompanied by an overall quality-reduction of genome sequences due to inherent errors in the sequencing technologies, presumably compensated by a decrease in both time and cost-ratios. For this reason, global standards are required for genome sequences to assess the quality of new data sets rapidly generated (Chain et al. 2009) (Figure 13).

All genome assemblers are based on the simple idea that highly similar DNA fragments do overlap. Two different approaches can be used to assemble reads obtained by multiple sequencing platforms: assembly by mapping or assembly *de novo*. If a genome reference sequence is available, DNA reads can be easily mapped against it. This step allows inferring the order and orientation of reads leading to the reconstruction of the genome sequence according to the reference sequence. Assembling by mapping is a technique mainly used to assess structural variants or analyze both inter and intraspecific nucleotide variability. Assembling a genome *de novo* is a more complex and sophisticated procedure which does not require the availability of a reference genome. *De novo* assemblers implement alignment-based algorithms that generate full-length sequences from short DNA fragments. Thus, it allows for the assembly of genomes with no related species sequenced.

Several modern software designed to assembly genomes *de novo* are currently available, and they support different sequencing technologies (Nagarajan and Pop 2013). Choosing among the great variety of assemblers represented one of the most challenging steps in this project. In order to obtain a high quality genome, the strategy that best fits to the sequencing data must be chosen. The available computer resources (mainly computer's memory) are limiting factors in every large-scale project. Hence, a previous knowledge on big data manipulation is required to avoid unexpected failures when running the assembly. Finally multiple alternatives have been proposed to help to improve assemblies. For example, a genome assembly can be assessed by parallel sequencing of the corresponding transcriptome, which facilitates the identification of genes sequence structure. By and large, to sequence a genome is a difficult task that requires coping with several technical barriers but it provides one of the most important sources to thoroughly investigate genomic features. In summary, it is remarkable the big



Raw sequences

Millions of DNA sequences are obtained from a single or multiple sequencing platforms depending on the established final goals.

Standard draft

Filtered data are assembled into contigs. This is considered the minimum standard for a submission to the public databases, although contigs likely contain regions of poor quality and can be incomplete.

High-quality draft

At least 90% of the genome is represented. Sequence errors and misassembles are still possible, but a general assessment of gene content can be performed.

Improved high-quality draft

Manual or automated methods are applied to reduce the number of misassembles. The number of contigs has decrease by generating larger scaffolds. Low-quality regions and putative base errors may be present. Comparison with other genomes can be done at this stage.

Annotation-Directed Improvement

Verification and correction of anomalies within coding regions. Gene models and other annotated features greatly supports the biology of the organism and the initial scientific questions proposed. Repeat regions are still not resolved.

Finished

Genome sequences with error rates lower than 1/100000. Each chromosome is assigned to a single and continuous sequence. Repetitive sequences have been ordered and correctly assembled. Only some microbial genomes have reach this standard.

FIGURE 14. **Community-defined categories of standards that better reflect the quality of genome sequences**.

effort employed herein to obtain a high quality assembly representing the genome of *D. buzzatii.*

5.2 Comparative genomics and evolution

By examining the structural and nucleotide variation between different organisms, comparative genomics offers fundamental and general insight into genome evolution. In this work we have focused on the identification of both macro (chromosomal inversions) and micro (nucleotide substitutions) DNA alterations responsible for environmental adaptation by comparing the genome sequences of species with a well-defined ecology. Two cactophilic fruitflies, *D. buzzatii* and *D. mojavensis*, have been used to carry out our genetic analyses since they exploit a particular range of natural resources providing an excellent model to assess environment-gene interactions (see Introduction).

In the first part of this project we have explored the impact of chromosomal inversions in the evolution of *D. mojavensis* genome. The characterization of the breakpoints associated to the seven inversions fixed in the chromosome 2 of *D. mojavensis* has shed light on the molecular causes and consequences of these rearrangements (see below). There is an increasing interest for the evolutionary dynamics underlying the chromosomal rearrangements, mainly inversions (Kirkpatrick 2010). This is particularly so because the power of DNA sequencing technologies and computer-based algorithms, which are predicted to replace old cytogenetic approaches as reported here, has promoted the identification of chromosomal rearrangements previously overlooked. In the past, the study of structural variation was limited by the restricted amount of available genomic data and by the lack of reliable molecular markers for detecting inversions in Drosophila. The development of bioinformatic tools and the increasing amount of genomic data have facilitated the molecular characterization of breakpoints of many individual genomic rearrangements (Mani and Chinnaiyan 2010). For instance, the availability of the complete genomes of 12 Drosophila species (Drosophila 12 Genomes Consortium et al. 2007) triggered the opportunity to infer genomic distances among more than a dozen species from Drosophila genus. The characterization of all micro and macro inversions provided information about the forces guiding gene-order alterations across Drosophila phylogeny using as reference one of the best known eukaryotic genomes, *D. melanogaster* (Bhutkar et al. 2008) (Figure 8).

Secondly we have examined genetic divergence between *D. mojavensis* and *D. buzzatii* as manifested in the accumulation of nucleotide substitutions in protein-coding genes. In this second step comparative genomics has offered us the opportunity to obtain estimates of selection pressures acting along the genome of the two different cactophilic lineages, as well as to provide an overview of the transcription dynamics along the development of *D. buzzatii*. Furthermore the combination of sequence data from the available species belonging to Drosophila genus has enabled to detect protein-coding genes that show strongest evidence for positive selection, likely indicative of molecular adaptation, and to find taxonomically restricted genes.

Overall, comparative genomics empowered by computed-based methods has provided us the possibility to investigate the genetic basis at both structural and nucleotide levels, of fitness-related traits in cactophilic species.

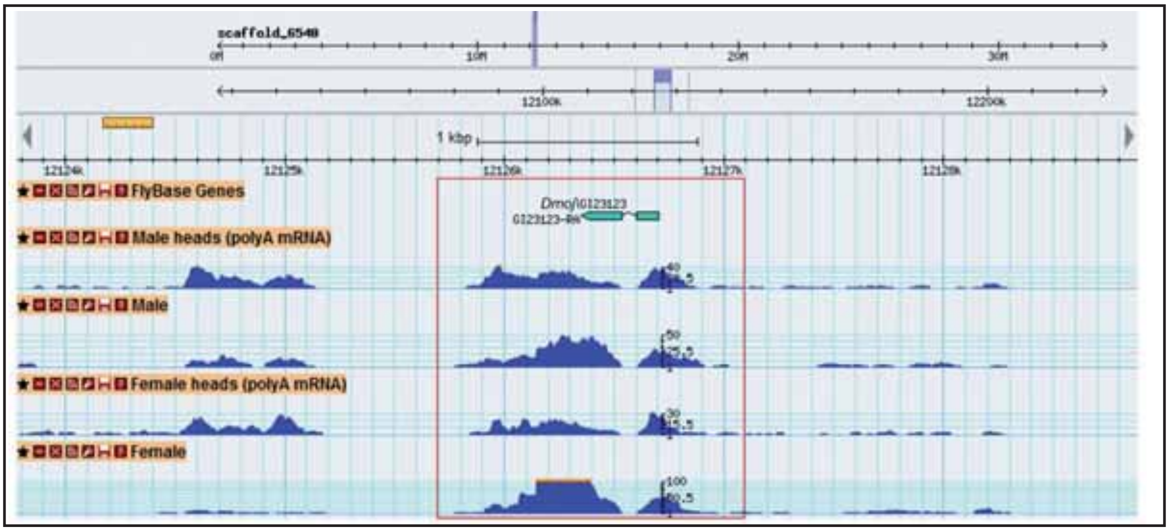
5.3 Chromosomal inversions and their role in adaptation

It has been demonstrated that chromosomal inversions affect the patterns of genomic evolution by reducing recombination, potentially facilitating climatic adaptation (Krimbas and Powell 1992) and inducing reproductive isolation (Rieseberg 2001; Kirkpatrick and Barton 2006). However, in this work (Guillén and Ruiz 2012) we have tested for position effects caused by inversion breakpoints and their consequences on the particular ecology of *D. mojavensis*.

The breakpoint of an inversion can disrupt or modify the expression of a gene that has cascading remarkable effects. Often the consequences of such alteration are expected to be deleterious, likely inducing genetic disorders. But less frequently these alterations can be the source of an adaptive mutation. Thus, the adaptive value of the inversion is given by a mutation at a single gene rather than the prevention of recombination between locally adapted genes (Hoffmann and Rieseberg 2008;

Kirkpatrick 2010). Our results are consistent with the position effect hypothesis since we have found gene alterations associated to inversion breakpoints that may have contributed to the fixation of these rearrangements by natural selection. Within this set of alterations we include the gain of two new genes, the structural change of the sequence coding for a heat shock protein (HSP), the modification of the regulation of another heat shock gene (*hsp*) and the sequence alteration of a gene belonging to *GstD* family as a consequence of its relocation.

It is widely recognized that the generation of new genes is potentially associated to new functions representing an important source to environment adaptation (Kaessmann 2010). Different mechanisms can lead to the generation of novel genes (see Introduction), but we have evidenced for the first time that they can appear as a consequence of an inversion in eukaryotes. Although we did not test for the expression of these two novel genes experimentally, the information provided by the modENCODE project (www.modencode.org) and the conserved domains database (CCD) (Marchler-Bauer and Bryant 2004) suggested that they are potentially functional (Figure 15). Even so it would be necessary to assess the expression pattern of these two genes and to thoroughly explore their functional dynamics in order to corroborate these observations.



<span>FIGURE 15.</span> Expression profile of Dmoj\GI23123 gene in *D. mojavensis*. The data provided by the modENCODE project (www.modencode.org) reveals that the new gene generated by the inversion *2h* is expressed at least in adult males and females.

Heat shock proteins (HSPs) are directly associated to thermotolerance and protection from cellular damage induced by extreme conditions (see Introduction). There is considerable evidence that they are essential for survival at both normal and elevated temperatures (Hoffmann et al. 2003). Recently Calabria et al. (2012) predicted that changes in HSP70 levels associated to a polymorphic inversion in Drosophila were linked to climatic adaptation. Thus, we cannot overlook the alterations that the *hsp* genes suffered as a consequence of the inversions *2s* and *2r* given the extreme thermal conditions surrounding *D. mojavensis*.

Overall whether the genetic differences that distinguish the inverted and ancestral arrangements were responsible for the inversion to be fixated or otherwise they accumulated after it became established for some other reason is an issue that we can not fully resolve. However our results contribute to the expected progress in identifying genes and traits underlying interspecific variation in ecological adaptation and they could represent the first evidence for the adaptive significance of a lineage specific rearrangement.

5.4 TE role in genome evolution

Transposable elements (TEs) affect gene structure and/or expression in several ways suggesting that they greatly contribute to complex evolutionary events (Fedoroff 2012). Here we provide compelling evidence for the implication of the TE *Bu*T5 (Rius et al. 2013) in the generation of the inversion *2s* by ectopic recombination. Moreover the insertion of a *Bu*T5 copy within the promoter associated to *CG10375* gene located in the proximal breakpoint of *2s* inversion indicates that TEs are involved not only in the mechanisms underlying inversions but also in the regulation of gene expression. *Bu*T5 has been classified as a miniature inverted-repeat TE (MITE) associated to the P element (Rius et al. 2013). P-like elements tend to insert into certain regions of the genome, specially sequences associated to *hsp* genes (Bellen et al. 2004; Shilova et al. 2006). It has been shown that heat-shock promoters represent natural "hotspots" for P-like transposable element integration because of the distinctive molecular features of heat shock genes, which seem to facilitate TEs accessibility (Lerman et al. 2003). Furthermore the prevalence of TEs in *Hsp* promoters may be favored by natural selection given the expression changes that undergone *hsp* genes as a consequence of the TE insertion under certain thermal conditions (Michalak et al. 2001; Walser et al. 2006).

It has been previously reported that TEs induce DNA breaks and are associated to chromosomal rearrangements (Finnegan 1989; Cáceres et al. 1999; Gray 2000; Casals et al. 2003). In addition they are important precursors of segmental duplications in Drosophila (Fiston-Lavier et al. 2007). However, the actual implication of TE activity in shaping the structural architecture of host genomes is difficult to assess because of the rapid dynamics of theses sequences. Even there is mounting evidence for the role of TEs in the generation of polymorphic inversions, by the time rearrangements are fixated within a population TEs can be lost or relocated (Bergman et al. 2002). Furthermore, the recurrent observation of TEs at rearrangement breakpoints is not an indicative for their direct implication in their generation as they tend to accumulate in regions with reduced recombination rates (Cáceres et al. 2001; Bartolomé et al. 2002; Casals et al. 2006).

Multiple cases of TEs altering gene expression in different organisms have also been described (Britten 2004; Medstrand et al. 2005; Feschotte 2008). However, as TEs have already become an important part of eukaryotic genomes, it is difficult to ascertain their global impact in gene regulation. In some natural populations of *D. melanogaster* it has been observed that the reduced Hsp70 expression induced by the insertion of a TE in its respective promoter resulted in an adaptation to extreme thermal conditions (Zatsepina et al. 2001). We claim that similar consequences can be expected after analyzing the effects of the *Bu*T5 insertion within the promoter sequence of the constitutive *hsp* gene CG10375 in *D. mojavensis*.

Finally the study of the impact of the polymorphic inversion *2j* in *D. buzzatii* (Puig et al. 2004; Puig 2011) confirmed that TEs are able to regulate the expression pattern of adjacent genes by transcriptional interference (Mazo et al. 2007). The widespread inversion *2j* confers a larger adult body size and a shorter developmental time on carrier individuals than that with the standard arrangement (*2st*). These phenotypic differences are related to the decreasing expression level of the gene CG13167 in *2j* embryos likely due to its silencing by the transcription of an antisense guiding by a *Kepler* copy. Overall our results support the idea that TEs act as potent genomic reorganizers and represent an important source of more complex types of mutation than simple DNA base alterations (Kidwell and Lisch 2000).

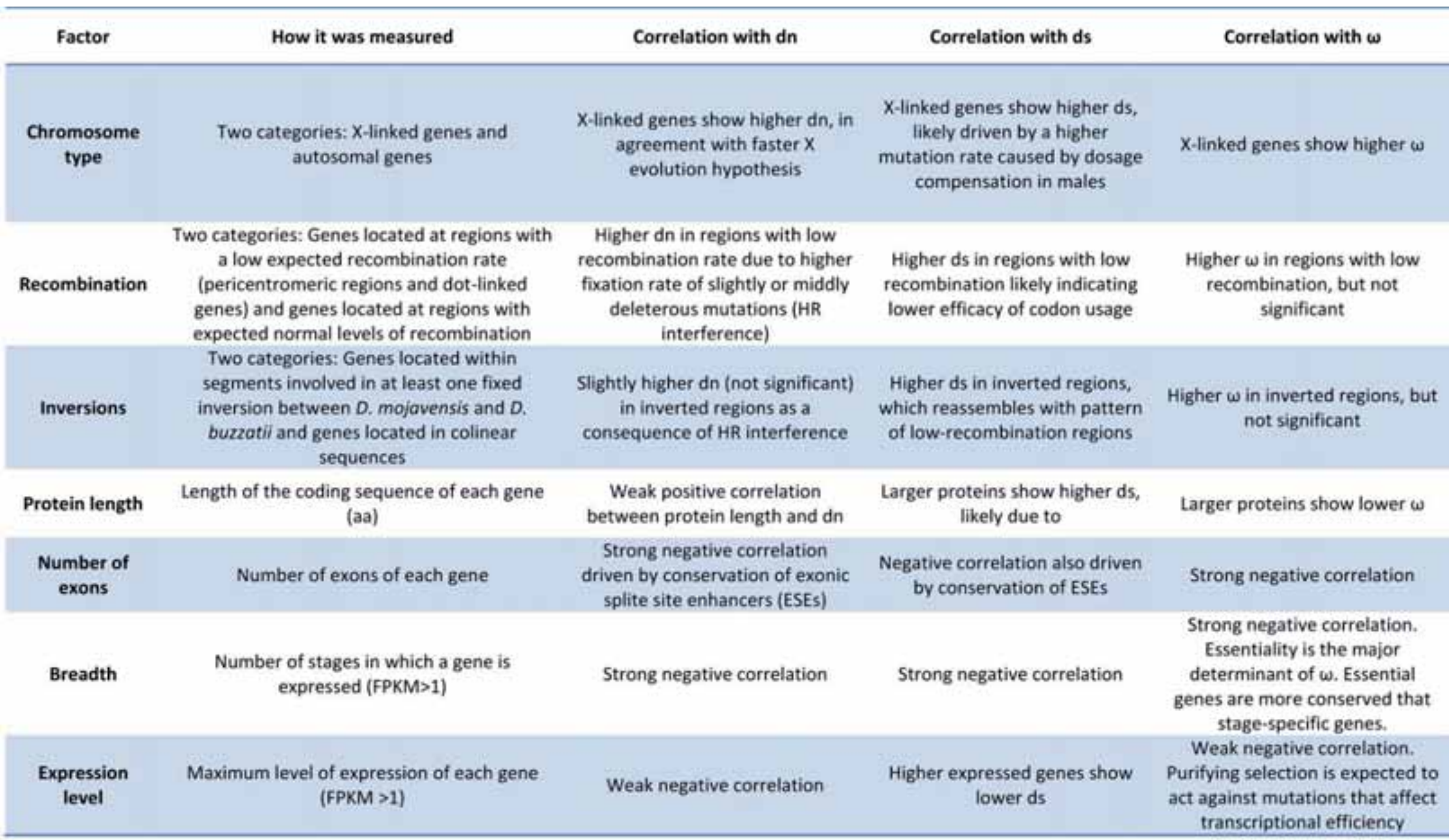
5.5 Divergence patterns and genomic determinants of gene evolution

Protein evolution clearly reflects the footprints of evolutionary adaptation at the molecular level. In order to infer the role of natural selection in functional divergence and to identify traits under positive selection, we have compared the protein-coding sequences of *D. mojavensis* and *D. buzzatii* genomes and we have described their evolutionary pattern. Our results have provided information about the selective determinants that affect the divergence patterns of protein-coding genes between these two species. We have shown that the evolution of protein-coding genes is affected by genomic attributes that interact with each other shaping the patterns of evolutionary variation (Table 4). There have been recent attempts to understand the implication of different factors in evolutionary rate of coding sequences in Drosophila, and similar conclusions have been extracted from all of them (Larracuente et al. 2008; Mackay et al. 2012; Campos et al. 2014).

Gene expression, including both expression bias and level, has been considered the most important determinant of protein evolutionary rates. Our findings are in agreement with previous studies that found that highly expressed genes show a slow rate of evolution (Larracuente et al. 2008). The observed slower rate has been associated to higher codon bias, increased functional importance and/or lower protein complexity of highly expressed genes (Lemos et al. 2005). However, we found that gene expression bias (estimated as the number of stages in which the gene is expressed) seems to have greater effects in shaping evolutionary patterns than expression level (Table 4). Genes that are expressed in more stages evolve slower than genes that are expressed in fewer stages. Larracuente et al. (2008) proposed that narrowly and ubiquitously expressed genes are differentially affected by pleiotropy, which is expected to strength the level of purifying selection on broadly expressed (or more essential) genes. Even that, essentiality does not seem to affect the possibility to experience positive selection. In addition, the effect of protein length, which seems to be independent of gene expression (Duret and Mouchiroud 1999; Lemos et al. 2005) is positive correlated to divergence rates. This indicates that it could be relevant to other aspects of molecular evolution and there is a need of a more detailed examination of this factor.

Patterns of interspecific nucleotide variation also provide a valuable signature of the evolutionary history of fixed inversions. Here we show that the effects of reduced recombination associated to inversions are observable even after they are fixated within the population. Comparing the divergence patterns between the most dynamics chromosomes and the nearly collinear chromosomes between *D. mojavensis* and *D. buzzatii* we have discovered that the divergence pattern in inverted segments resembles that observed in regions with reduced recombination. Thus the maintenance of linkage disequilibrium (LD) by inversions (Hoffmann and Rieseberg 2008) is reflected as an increasing effect of Hill-Robertson (HR) interference. The suppression of the recombination driven by inversions can lead to dramatic effects on individuals fitness (Charlesworth and Charlesworth 2000). One of the most drastic examples of the long-term consequence of suppressed recombination is the mammalian chromosome Y, which is suffering a continuous genetic degeneration (Graves 2006). On the other hand, the suppression of the recombination between alternative chromosomal arrangements can contribute to local adaptation or reproductive isolation. Under this assumption, genes affecting adaptive divergence disproportionally reside within inversions and the effects of the rearrangement contribute to both adaptation and ecological reproductive isolation across habitats (Lai et al. 2005; Hoffmann and Rieseberg 2008; Feder and Nosil 2009). One of the most iconic examples of this theory was described by Lowry and Willis (2010) when they studied the yellow monkeyflower *Mimulus guttatus*. They concluded that a polymorphic inversion that differentiated the two distinct ecotypes of this flower was the responsible for much of the phenotypic variation that distinguished both populations, acting as a supergene.

TABLE 4. Genomic determinants of protein-coding gene evolution in *Drosophila mojavensis* and *Drosophila buzzatii*.



Factor	How it was measured	Correlation with dn	Correlation with ds	Correlation with ω
Chromosome type	Two categories: X-linked genes and autosomal genes	X-linked genes show higher dn, in agreement with faster X evolution hypothesis	X-linked genes show higher ds, likely driven by a higher mutation rate caused by dosage compensation in males	X-linked genes show higher ω
Recombination	Two categories: Genes located at regions with a low expected recombination rate (pericentromeric regions and dot-linked genes) and genes located at regions with expected normal levels of recombination	Higher dn in regions with low recombination rate due to higher fixation rate of slightly or middly deleterous mutations (HR interference)	Higher ds in regions with low recombination likely indicating lower efficacy of codon usage	Higher ω in regions with low recombination, but not significant
Inversions	Two categories: Genes located within segments involved in at least one fixed inversion between *D. mojavensis* and *D. buzzatii* and genes located in colinear sequences	Slightly higher dn (not significant) in inverted regions as a consequence of HR interference	Higher ds in inverted regions, which reassembles with pattern of low-recombination regions	Higher ω in inverted regions, but not significant
Protein length	Length of the coding sequence of each gene (aa)	Weak positive correlation between protein length and dn	Larger proteins show higher ds, likely due to	Larger proteins show lower ω
Number of exons	Number of exons of each gene	Strong negative correlation driven by conservation of exonic splite site enhancers (ESEs)	Negative correlation also driven by conservation of ESEs	Strong negative correlation
Breadth	Number of stages in which a gene is expressed (FPKM>1)	Strong negative correlation	Strong negative correlation	Strong negative correlation. Essentiality is the major determinant of ω. Essential genes are more conserved that stage-specific genes.
Expression level	Maximum level of expression of each gene (FPKM >1)	Weak negative correlation	Higher expressed genes show lower ds	Weak negative correlation. Purifying selection is expected to act against mutations that affect transcriptional efficiency

Assuming that the evolutionary dynamics of a gene partially depends on its mode of inheritance, we expect to observe differences in divergence patterns between the X chromosome and autosomes (Vicoso and Charlesworth 2009). The faster-X effect hypothesis postulates that as X-linked genes are subjected to different levels of selection, mutation, recombination and effective population size, they evolve faster (Charlesworth et al. 1987). The results obtained by comparing the divergence rates of coding genes between autosomes and X chromosome performed herein, are in agreement with this hypothesis. Several studies performed in Drosophila genus have previously supported the faster-X hypothesis by comparing the accumulation of nucleotide substitutions between X-linked and autosomal loci (Figure 16) (Betancourt et al. 2002; Counterman et al. 2004; Begun et al. 2007a; Singh et al. 2008; Vicoso and Charlesworth 2009). When divergence ratios associated to X chromosome are greater than that of autosomes it is said that X chromosome evolve faster. However, by this approach it is not possible to clearly differentiate between adaptive and nonadaptive causes of faster-X evolution and an approach combining both inter and intraspecific nucleotide variation data is recommended (McDonald and Kreitman 1991). Two new tests have provided evidences for a faster-X evolution in addition to classic methods. First the study of the genome of *D. miranda*, which presents a recently formed neo-X chromosome (Zhou and Bachtrog 2012), confirmed that hemizygous neo-X-linked genes evolve faster than effectively diploid genes located in the same chromosome. Second, the analysis of the evolution of X-linked duplicated genes has demonstrated that their divergence rates are higher than autosomal duplicates (Thornton and Long 2002). Finally Bhutkar et al. (2008) observed that X chromosome harbors more inversions than other elements along the Drosophila genus phylogeny. They emphasized that although the higher rate of rearrangement fixation in X could support a higher rate of evolution, this chromosome tends to be the less represented in a genome sequence and as a consequence, it is associated to a higher level of assembly artifacts. Thus, we highlight

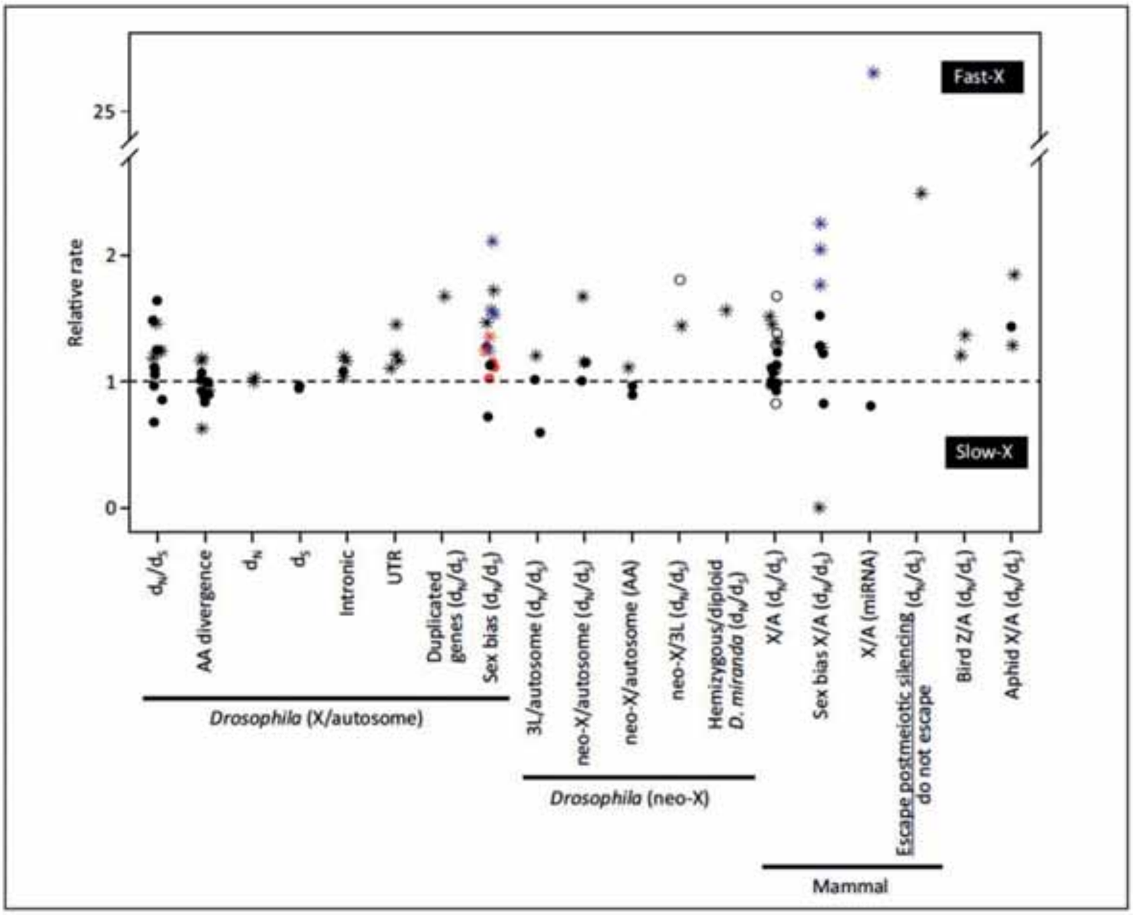


FIGURE 16. **Tests for faster-X divergence in different organisms from Drosophila genus and mammals.** The relative rate of evolution is plotted for different classes of nucleotide site and crhomosome in Drosophila and mammals. The expectation that X-linked and autosomal genes evolve at equal rates is represented by a discontinuous line. Significant deviation from unity in the relative rate is indicated by an asterisk, whereas no significant differences or studies in which significance was absent are indicated by a black or white circle, respectively. In studies where expression was measured (indicated by 'sex bias' in the x-axis label), the color of the point indicates the expression class of the gene (black, non-sex biased; blue, male biased; and red, female biased). Figure extracted from Meisel and Connallon (2013).

the importance of high quality genomes, especially when the results completely depend on heterogeneity in coverage among different genomic regions.

The integration of distinct genomic attributes has allowed us to assess the role of recombination in gene evolution by analyzing genome regions that are differentially exposed to crossing over events. We have highlighted the importance of protein sequence features, expression patterns and gene location among other factors in shaping the evolutionary process of divergence. Although our analyses contribute to disentangle the effect of many biological attributes in gene history, we emphasize that other organismic attributes not incorporated to this study likely influence protein evolution. Thus a use of an extensive range of expression data jointly with the addition of new genomic variables is expected to be incorporated in ongoing projects.

5.6 Inferring positive selection

Positive selection, also known as Darwinian selection, is described as the process by which new advantageous mutations sweep a population. The detection of positive selection has long been considered a challenging task since neutral and deleterious variants predominate over them in frequency. Nowadays the two major recurrent methods to infer positive selection are based on (i) analysis of codon substitutions between multiple species (Yang et al. 2000) and (ii) nucleotide polymorphism within a species compared to interspecific divergence (McDonald and Kreitman 1991; Messer and Petrov 2013).

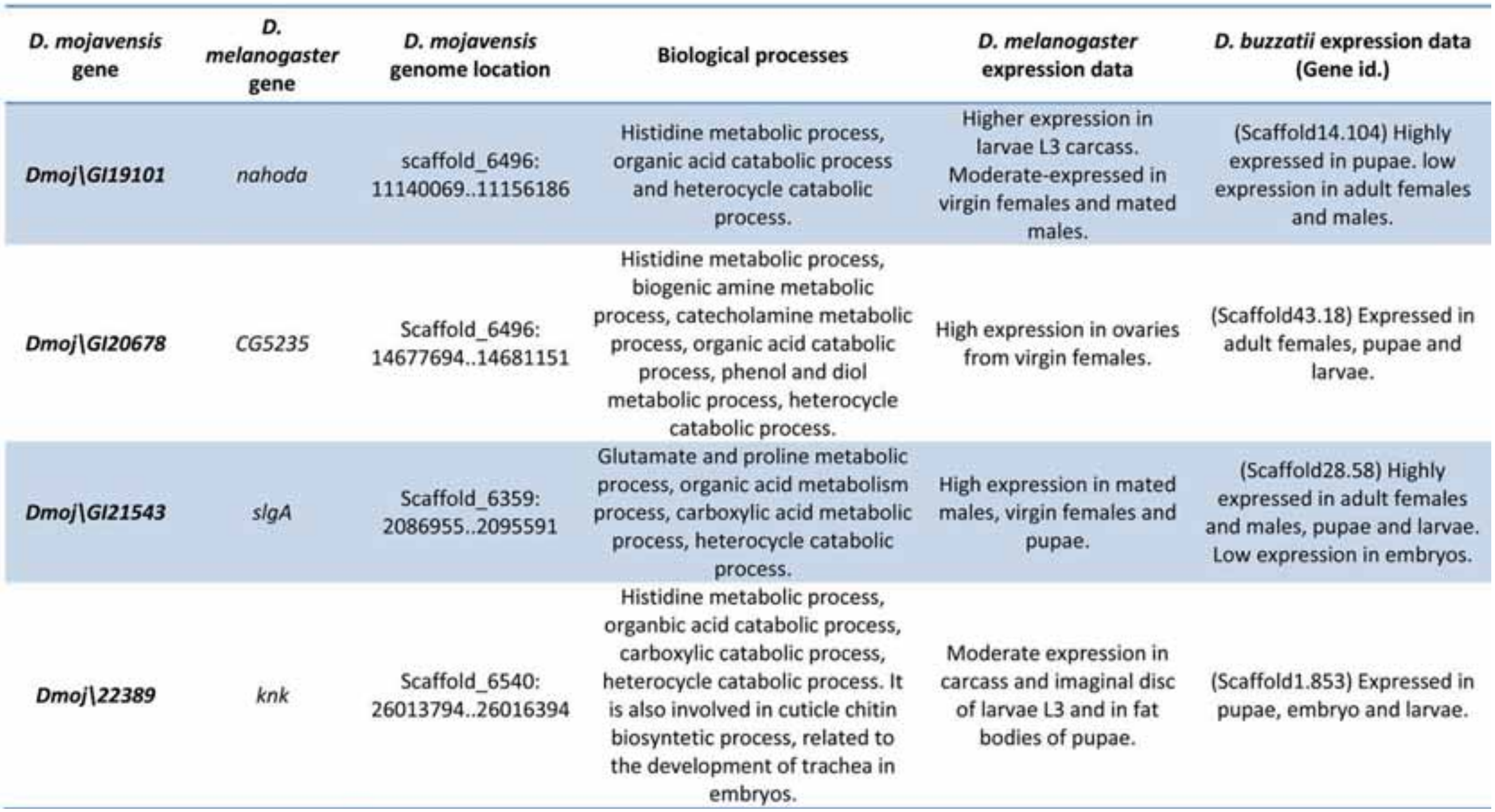
The classical way to infer distinct selective pressures acting on coding genes was based on ka/ks (ω) rate estimation (see Introduction). But ω ratio is a very conservative test of positive selection because many sites might be under strong purifying selection owing to functional constraint, with the ω ratio close to 0 (Figure 6). Indeed, only 15 out of the 9017 (0.16%) orthologs analyzed between *D. mojavensis* and *D. buzzatii* are likely to be under positive selection considering the criteria of ka/ks>1, contrary to the 1214 genes evidencing positive selection using codon substitution models. Thus, nowadays the ω ratio estimation is mainly used as a test for assessing protein-coding regions in genomes assuming that in every gene dn is significantly smaller than ds (Yang 2002).

One of the most robust methods to quantify the rate of adaptive evolution is the McDonald-Kreitman (MK) test. In the MK test the number of segregating variants (polymorphisms) are contrasted to the number of substitutions (divergence) at synonymous and nonsynonymous sites (McDonald and Kreitman 1991). In summary, as beneficial mutations should rapidly spread to fixation, their contribution to polymorphism is expected to be less than their role in divergence, and the proportion of substitutions driven by positive selection can be determined by the α parameter (Eyre-Walker 2006). In this work the identification of genes evolving under positive selection has been performed using only divergence data by testing different codon substitution models (Yang et al. 2000). However, the availability of the genome sequences of two different strains of *D. buzzatii*, *st-1* and *j-19*, allows for the possibility to analyze the adaptive evolution in cactophilic flies combining polymorphism and divergence data in ongoing projects.

As several broad-scale analyses focused on determinate which genes are driven by positive selection are carried out, two principal categories of rapidly evolving genes are being confirmed (Drosophila 12 Genomes Consortium et al. 2007; Heger and Ponting 2007). These two categories are immune defense and reproduction. The constant interaction between hosts and pathogens results in a co-evolutionary process between genes from the two organisms. In addition, sexual selection entails a potent force on genes involved in post mating sperm competition for fertilization (Ellegren 2008).

We found a significant number of genes under positive selection involved in functions related to cell-cell recognition and immune system. However the most represented category in our set of positive selected genes was transcription factor activity. Transcription factors (TFs) are one of the major contributors to complexity in differentiation in animal and plant cells (Phillips and Hoopes 2008). It is known that TFs control many important parts of development and some of them are only activate at a selected few promoters. Thus it is difficult to ascertain the implication of the TFs in the particular ecology of cactophilic flies unless a further analysis is performed. Finally, the enrichment of positively selected genes involved in heterocycle catabolic processes in *D. mojavensis* lineage is a valuable finding given the chemical characteristics of the main host of this species (see Introduction). This enrichment is exemplified by four genes: *Dmoj\GI19101*, *Dmoj\GI20678*, *Dmoj\GI21543* and *Dmoj\GI22389* (Table 5). All of these genes are also involved in processes related to the metabolism of different amino acids and organic compounds. They do not seem to be clustered in a particular region of the genome, and according to the expression data extracted from both *D. melanogaster* and *D. buzzatii* genomes, they cannot be considered constitutive genes. Finally we expect to disentangle the role of these candidate genes in future studies with the help of expression data extracted from several developmental stages of *D. mojavensis*.

TABLE 5. Genes evolving under positive selection in *D. mojavensis* lineage involved in heterocycle catabolic processes.



D. mojavensis gene	*D. melanogaster* gene	*D. mojavensis* genome location	Biological processes	*D. melanogaster* expression data	*D. buzzatii* expression data (Gene id.)
Dmoj\GI19101	*nahoda*	scaffold_6496: 11140069..11156186	Histidine metabolic process, organic acid catabolic process and heterocycle catabolic process.	Higher expression in larvae L3 carcass. Moderate-expressed in virgin females and mated males.	(Scaffold14.104) Highly expressed in pupae. low expression in adult females and males.
Dmoj\GI20678	*CG5235*	Scaffold_6496: 14677694..14681151	Histidine metabolic process, biogenic amine metabolic process, catecholamine metabolic process, organic acid catabolic process, phenol and diol metabolic process, heterocycle catabolic process.	High expression in ovaries from virgin females.	(Scaffold43.18) Expressed in adult females, pupae and larvae.
Dmoj\GI21543	*slgA*	Scaffold_6359: 2086955..2095591	Glutamate and proline metabolic process, organic acid metabolism process, carboxylic acid metabolic process, heterocycle catabolic process.	High expression in mated males, virgin females and pupae.	(Scaffold28.58) Highly expressed in adult females and males, pupae and larvae. Low expression in embryos.
Dmoj\22389	*knk*	Scaffold_6540: 26013794..26016394	Histidine metabolic process, organbic acid catabolic process, carboxylic catabolic process, heterocycle catabolic process. It is also involved in cuticle chitin biosyntetic process, related to the development of trachea in embryos.	Moderate expression in carcass and imaginal disc of larvae L3 and in fat bodies of pupae.	(Scaffold1.853) Expressed in pupae, embryo and larvae.

5.7 From Genomics to Transcriptomics

Next-generation RNA sequencing (RNA-seq) is a powerful tool to study the dynamics of transcriptomes at exceptional resolution (Hoeijmakers et al. 2013). Perhaps the most salient benefit of RNA-seq is that the nucleotide sequence of the target genome is not needed making it possible to analyze poorly characterized organisms. The increasing number of studies focused on transcription dynamics (Mortazavi et al. 2008; Nagalakshmi et al. 2008; Graveley et al. 2011), which extend from single-molecule techniques (Reed et al. 2007) to genome-wide measurements (Trapnell et al. 2010), is unveiling the extraordinary complexity of eukaryotic genomes.

Nowadays, one of the best characterized transcriptomes is that of *D. melanogaster* as a result of the collective effort invested in the modEncode (model organism Encyclopedia of DNA elements) Project (Celniker et al. 2009). The modEncode Project was launched in order to generate an unprecedented detailed catalogue of the functional elements in the *C. elegans* and *D. melanogaster* genomes. In the first stage of the project more than 1900 new transcribed regions in *D. melanogaster* were identified, and other new transcribed elements including highly conserved small non-coding RNAs and microRNAs were discovered. In addition they analyzed the factors underlying alternative splicing events along the development, providing major understanding about the expression dynamics throughout the Drosophila life cycle. It is remarkable that the study of the developmental transcriptome based on deep RNA-seq experiments, as reported here in *D. buzzatii*, has been carried out only in *D. melanogaster* according to the modencode database (www.modencode.org). One of the most outstanding features of Drosophila genome revealed by these studies is the high level of compactness. The pervasive transcription of previously uncharacterized ncRNAs suggests that they can be important determinants in regulating gene expression (Mercer et al. 2009; Hainer and Martens 2011). However, the debate concerning the functional significance of ncRNAs still remains open.

Recent studies performed through improved methods including perturbation experiments have revealed even higher transcriptional complexity in Drosophila (Brown et al. 2014). Most transcriptional complexity is found in genes involved in nervous system, which seems to be entailed by an enrichment of RNA editing events and UTR sequences extensions (Figure 3). Surprisingly sense and antisense transcripts are found in the same cells at the same times, suggesting that transcriptional interference is a conserved and recurrent mechanism to control gene expression. In addition the catalogue describing ncRNAs encoding mostly for putative short amino acids (Ladoukakis et al. 2011) has been expanded. In summary, organismic complexity is demonstrated to be dramatically influenced by the high variability of regulation mechanisms.

Finally, a clear sex biased gene expression has been reported when analyzing the developmental transcriptome of *D. melanogaster* (Graveley et al. 2011; Brown et al. 2014) and *D. buzzatii*. In *D. buzzatii* adult males express up to 1800 more genes than adult females. By and large the presence of sexual dimorphism constitutes the most extreme phenotypic variation within species, so genetic variation between males and females are somehow expected to be reported. Genome-wide studies focused on gene expression patterns have revealed an extensive variety between females and males not only on gene content but also on gene expression (Graveley et al. 2011; Parsch and Ellegren 2013). Indeed some important progresses have been made regarding to sex-biased expression. For example, it has been found that 8% of the genes in *D. melanogaster* show segregating expression variation with opposite fitness effects in females and males, i.e. they are sexually antagonistic (Innocenti et al. 2010). However, the causes underlying gene expression differences between males and females need to be thoroughly analyzed.

5.8 GBrowse and web resource

The dramatic accumulation of genomic data has led to the development of several tools that facilitate the integration of biological information into computerized databases. One of the most recurrent bioinformatics tools are genome browsers. Genome browsers are web-based user interfaces that offer a practical solution to analyze and visualize large quantities of highly interrelated genomic data (Schattner 2008). In order to promote the easy-accessibility of the information provided by the *Drosophila buzzatii* Genome project, we have constructed a database incorporating some of the most important results, as well as a customized browser of the genome of *D. buzzatii.* This browser was launched using the Generic Genome Browse (GBrowse) application (Stein et al. 2002), which has been successfully used to integrate a wide variety of genomic data, from model organisms to humans (Stein 2013). In summary the *D. buzzatii* Genome Project webpage (www.dbuz.uab.cat) is a compilation of the most relevant information regarding to this work, including (i) a description of the project and the partners that have participated (ii) direct links to external databases (iii) a blast-based alignment tool (iv) a genome browser and (v) an interactive section to share information about the *D. buzzatii* genome Project (Figure 17).

The customized GBrowse of the *D. buzzatii* genome incorporates multiple tracks including all the gene and TE annotations produced by different algorithms, orthology relationships with other Drosophila species and the information extracted from the RNAseq-based experiments. Annotations obtained from RNAseq using Cufflinks include coding and non-coding regions (ncRNAs and UTRs) of the genome that are expressed in the five developmental stages that were analyzed (Figure 18). Definitely, the Gbrowser tool offers an intuitive way to explore the *D. buzzatii* genomic features analyzed in this work. In the near future we intend to incorporate all the *D. buzzatii* genome information represented herein into the leading website of Drosophila genomes, the FlyBase webpage (The FlyBase Consortium 2002).

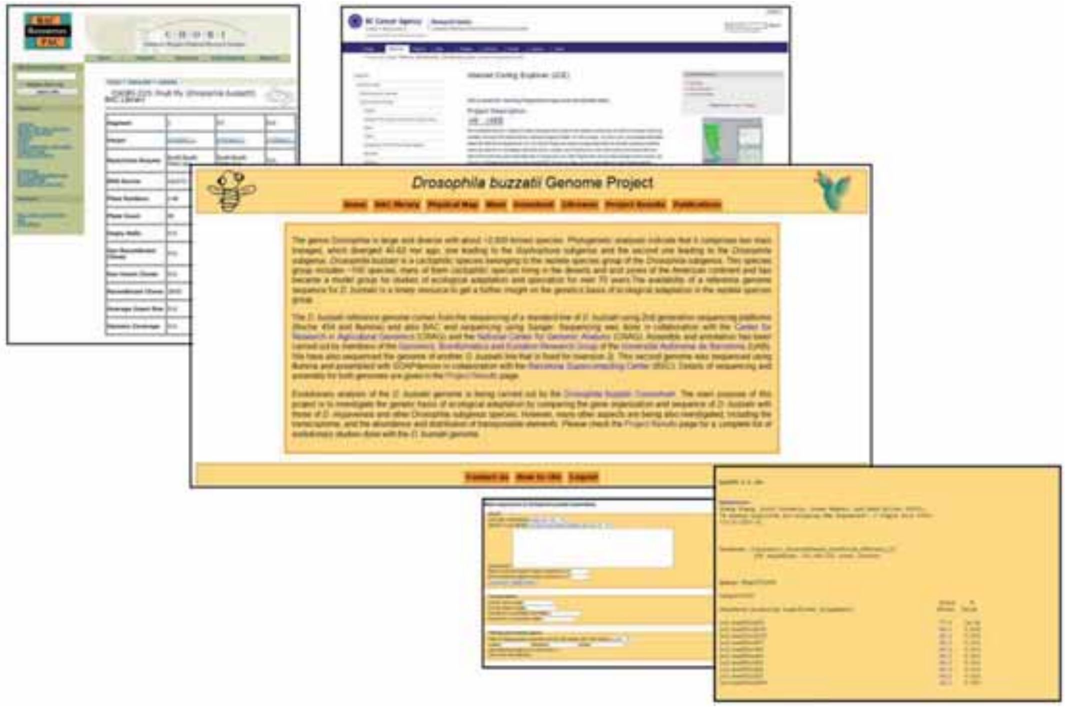


FIGURE 17. Overview of some of the applications implemented in the Drosophila buzzatii Genome Project webpage (www.dbuz.uab.cat). Direct links to both, the BAC library and the physical map of *D. buzzatii* previously constructed, are provided. A blast-based application allows searching nucleotide and protein sequences in the contigs and scaffolds of the genome of *D. buzzatii*.

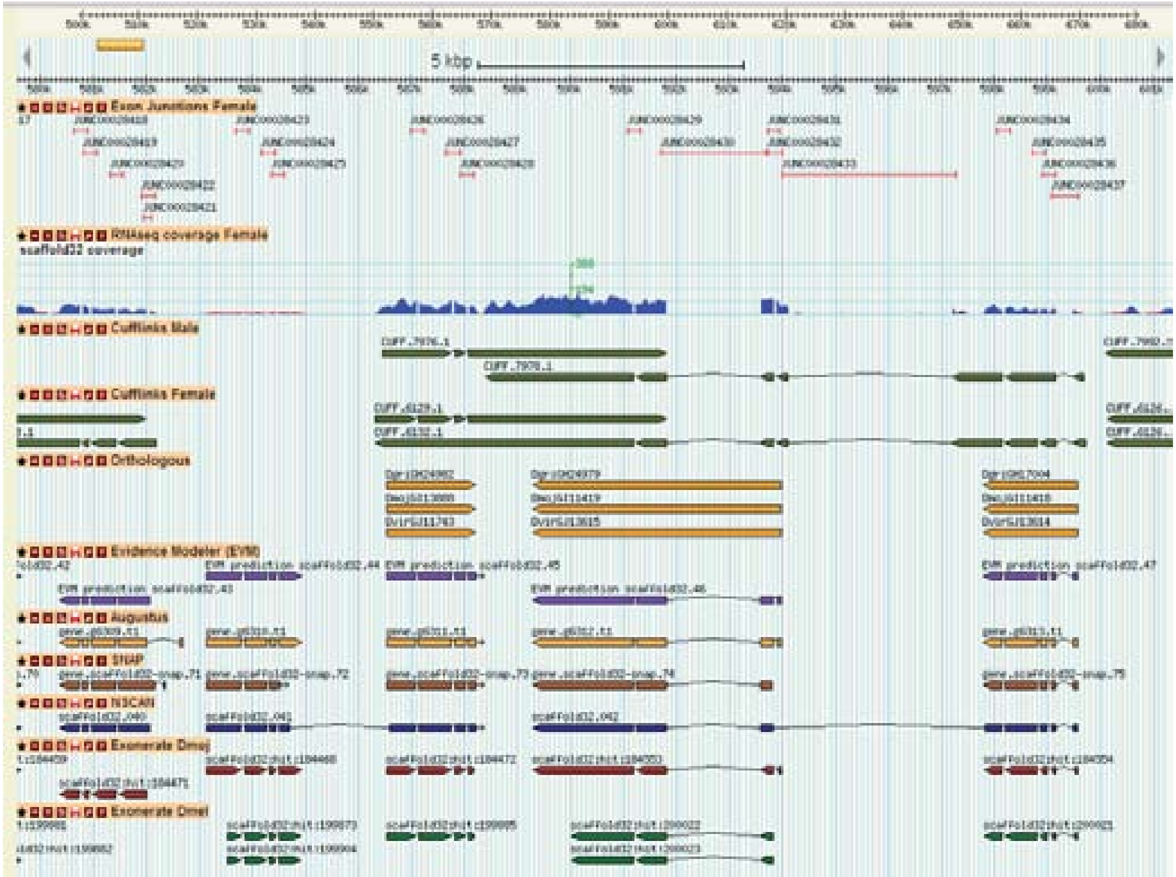


FIGURE 18. **Overview of the genomic features represented in the Gbrowse implemented in the** ***D. buzzatii*** **Genome Project web.**

6. CONCLUSIONS

1. A total of seven inversions (*2s*, *2r*, *2q*, *2h*, *2f*, *2g* and *2c*) have been fixed in the chromosome 2 of *D. mojavensis* since the divergence between *D. mojavensis* and *D. buzzatii.* These results agree with those obtained by previous cytological-based studies.
2. We have provided information about the molecular causes that generated at least three fixed inversions by characterizing all corresponding breakpoints. One of the inversions (*2s*) showed unequivocal evidence for its generation by ectopic recombination between two copies of *Bu*T5, thus demonstrating for the first time the implication of a TE in the generation of a fixed inversion in Dipterans. Two other inversions (*2h* and *2q*) have been likely generated by staggered single-strand breaks and repair by NHEJ, resulting in the duplication of the non-repetitive DNA sequences involved in both single-strand breakages.
3. We have found an excess of breakpoints (four out of 14) that fall between duplicated genes tandemly arranged in the parental genome (*D. virilis*). We argue that either duplicated genes likely undergone structural instability leading to an increasing rate of DNA breakage or they represent breakage permissive regions. We also remark the possibility of beneficial position effects produced by the relocation of duplicated copies entailed by changes in their background genomic landscape.
4. An association between inversion breakpoints and gene transposition events has been reported in this work. We suggest that this association is the result of the intrinsic fragility of sequences undergone breakpoints.
5. Two novel genes (*Dmoj\GI23123* and *Dmoj\22075*) have been originated by *2h* and *2q* inversions respectively, due to the mechanism that generated both inversions. The gene *Dmoj/GI23123* seems to be expressed according to available expression

data from *D. mojavensis* genome. The gene *Dmoj\22075* conserves a MFS domain from the parental copy, suggesting that it could encode a functional protein.

6. Three inversions produced putative structural and/or expression changes in genes adjacent to breakpoints. The relocation of *GstD1* by *2c* inversion could have significant adaptive consequences in species harboring this rearrangement given the demonstrated biological importance of this gene. The inversion *2r* resulted in a size reduction or pseudogeneization of one of the *hsp68* gene copies (*hsp68a*) found in the parental genome. The relocation of the other copy (*hsp68b*) driven by the inversion, made it to acquire a new *cis*-regulatory element likely altering its gene expression pattern. Finally the changes induced by inversion *2s* and *Bu*T5 insertion in the promoter of CG10375, a gene belonging to Hsp40 family, could conferred an adaptive advantage to *D. mojavensis* thermotolerance.

7. The genome of *D. buzzatii* has been sequenced and assembled *de novo* using reads obtained from different platforms (454, Illumina and Sanger). The 158 scaffolds contained in the N90 index have been anchored to chromosomes allowing for the analysis of the structural variation between *D. mojavensis* and *D. buzzatii*.

8. Using a combination of both *ab initio* and homology-based methods, 13657 protein-coding genes have been annotated (Annotation Release 1).

9. The information extracted from RNAseq of five life-stages from *D. buzzatii* revealed that a total of 15573 genes are expressed in at least one developmental stage; from these, 81% are coding genes whereas 19% are ncRNA genes. The expression pattern of ncRNA and coding genes greatly varies along development. A clear sex-biased expression in adults has been observed.

10. Unique orthologous genes between *D. buzzatii* and *D. mojavensis* have been retained from Annotation Release 1 (9017) in order to analyze patterns of divergence. Chromosome type (autosomes vs. X), recombination and inversions have been demonstrated to influence divergence rates at both synonymous and non-synonymous sites (ds and dn, respectively). Other genomic factors including exon number, protein length and expression pattern have significant effect on divergence rate at synonymous sites (ds).

11. We have detected 1294 genes that show evidences for positive selection, representing up to 14% of the total set of 1:1 orthologs between *D. mojavensis* and *D. buzzatii*. X chromosome harbors a significantly higher number of genes evolving under positive selection compared to autosomes. Putative positive selected genes in *D. mojavensis* lineage are enriched in functions related to the characteristic adaptation of *D. mojavensis* to its main host cactus.

12. We found in *D. mojavensis* and *D. buzzatii* genomes 117 coding genes with no similarity to any previously predicted Drosophila protein. RNAseq data revealed that 87% of these orphan genes are expressed in at least one developmental stage. The number of orphan genes that show evidences of positive selection is higher than that expected by random and both divergence and expression patterns clearly differ from that of older genes, evidencing that orphans evolve faster.

APPENDIX

Genomics of ecological adaptation in cactophilic Drosophila: hundreds of genes under positive selection in the D. buzzatii and D. mojavensis lineages

Supplemental information

Table A1. Number of protein-coding genes (PCG) and non-coding genes (ncRNA) expressed along *D. buzzatii* development.

Stage	PCG	ncRNA	Total
Embryo	8552	1208	9760
Larvae	8709	810	9519
Pupae	10485	1574	12059
Female adult	9310	1037	10347
Male adult	10347	1824	12171
Total	**47403**	**6453**	53856

Table A2. Number of PCG and ncRNA expressed in one or more stages.

Stages	PCG	ncRNA	Total
1	925	1292	2217
2	1655	689	2344
3	1322	393	1715
4	1618	326	1944
5	6546	260	6806
Total	12066	2960	15026

Table A3. Distribution of putative positive selected genes expressed along *D. buzzatii* development.

Stage	Positive selected	Non-positive selected	Total
Embryo	881	7671	8552
Larvae	812	7897	8709
Pupae	1069	9416	10485
Female adult	932	8378	9310
Male adult	1000	9347	10347
Total	4694	42709	47403

Table A4. Expression breadth distribution of positive selected genes in *D. buzzatii*.

Stages	Positive selected	Non-positive selected	Total
1	106	819	925
2	166	1489	1655
3	119	1203	1322
4	211	1407	1618
5	611	5935	6546
Total	1213	10853	12066

Table A5. Distribution of orphan genes expression in *D. buzzatii* life cycle.

Stage	Orphans	Non-orphans	Total
embryo	21	8531	8552
larvae	49	8660	8709
pupae	51	10434	10485
female	35	9275	9310
male	54	10293	10347
Total	210	47193	47403

Table A6. Number of orphans and non-orphans expressed in one or more stages of *D. buzzatii* life cycle.

Stage	Orphans	Non-orphans	Total
1	29	896	925
2	18	1637	1655
3	11	1311	1322
4	8	1610	1618
5	16	6530	6546
Total	82	11984	12066

Table A7. Chromosome location of putative positive selected genes detected by site models (SM). The location of one of the 772 gene candidates was unknown.

Chromosome	Positive selected (SM)	Non-positive selected	Total
X	168	1259	1427
2	154	2151	2305
3	129	1557	1686
4	155	1653	1808
5	161	1686	1847
6	4	25	29
Total	771	8331	9102

Table A8. Chromosome location of putative positive selected genes detected by all models (SM and BSM). The chromosome location of two of the 1294 gene candidates was unknown.

Chromosome	Positive	Non-positive selected	Total
X	260	1167	1427
2	264	2041	2305
3	238	1448	1686
4	245	1563	1808
5	277	1570	1847
6	8	21	29
Total	1292	7810	9102

BIBLIOGRAPHY

Acuña R, Padilla BE, Flórez-Ramos CP, Rubio JD, Herrera JC, Benavides P, Lee S-J, Yeats TH, Egan AN, Doyle JJ, et al. 2012. Adaptive horizontal transfer of a bacterial gene to an invasive insect pest of coffee. *Proc Natl Acad Sci* **109**: 4197–4202.

Adams J. 2008. Transcriptome: Connecting the Genome to Gene Function. *Nat Educ* **1**: 195.

Adams KL, Wendel JF. 2005. Polyploidy and genome evolution in plants. *Curr Opin Plant Biol* **8**: 135–141.

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, et al. 2000. The genome sequence of Drosophila melanogaster. *Science* **287**: 2185–2195.

Amemiya CT, Alföldi J, Lee AP, Fan S, Philippe H, MacCallum I, Braasch I, Manousaki T, Schneider I, Rohner N, et al. 2013. The African coelacanth genome provides insights into tetrapod evolution. *Nature* **496**: 311–316.

Aminetzach YT, Macpherson JM, Petrov DA. 2005. Pesticide resistance via transposition-mediated adaptive gene truncation in Drosophila. *Science* **309**: 764–767.

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.

Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408**: 796–815.

Ashburner M, Bergman CM. 2005. Drosophila melanogaster: A case study of a model genomic sequence and its consequences. *Genome Res* **15**: 1661–1667.

Ayala D, Fontaine MC, Cohuet A, Fontenille D, Vitalis R, Simard F. 2011. Chromosomal inversions, natural selection and adaptation in the malaria vector Anopheles funestus. *Mol Biol Evol* **28**: 745–758.

Bailey JA, Eichler EE. 2006. Primate segmental duplications: crucibles of evolution, diversity and disease. *Nat Rev Genet* **7**: 552–564.

Bailey SM, Meyne J, Cornforth MN, McConnell TS, Goodwin EH. 1996. A new method for detecting pericentric inversions using COD-FISH. *Cytogenet Cell Genet* **75**: 248–253.

Barker JSF, Starmer WT. 1982. *The Cactus-Yeast-Drosophila Model System*. Academic Press, Sidney, Australia.

Bartolomé C, Maside X, Charlesworth B. 2002. On the abundance and distribution of transposable elements in the genome of Drosophila melanogaster. *Mol Biol Evol* **19**: 926–937.

Begun DJ. 1997. Origin and Evolution of a New Gene Descended From alcohol dehydrogenase in Drosophila. *Genetics* **145**: 375–382.

Begun DJ, Holloway AK, Stevens K, Hillier LW, Poh Y-P, Hahn MW, Nista PM, Jones CD, Kern AD, Dewey CN, et al. 2007a. Population Genomics: Whole-Genome Analysis of Polymorphism and Divergence in Drosophila simulans. *PLoS Biol* **5**: e310.

Begun DJ, Lindfors HA, Kern AD, Jones CD. 2007b. Evidence for de novo evolution of testis-expressed genes in the Drosophila yakuba/Drosophila erecta clade. *Genetics* **176**: 1131–1137.

Bellen HJ, Levis RW, Liao G, He Y, Carlson JW, Tsang G, Evans-Holm M, Hiesinger PR, Schulze KL, Rubin GM, et al. 2004. The BDGP gene disruption project: single transposon insertions associated with 40% of Drosophila genes. *Genetics* **167**: 761–781.

Bergman CM, Pfeiffer BD, Rincón-Limas DE, Hoskins RA, Gnirke A, Mungall CJ, Wang AM, Kronmiller B, Pacleb J, Park S, et al. 2002. Assessing the impact of comparative genomic sequence data on the functional annotation of the Drosophila genome. *Genome Biol* **3**: research0086.

Bergman CM, Quesneville H, Anxolabéhère D, Ashburner M. 2006. Recurrent insertion and duplication generate networks of transposable element sequences in the Drosophila melanogaster genome. *Genome Biol* **7**: R112.

Betancourt AJ, Presgraves DC, Swanson WJ. 2002. A test for faster X evolution in Drosophila. *Mol Biol Evol* **19**: 1816–1819.

Bhutkar A, Schaeffer SW, Russo SM, Xu M, Smith TF, Gelbart WM. 2008. Chromosomal Rearrangement Inferred From Comparisons of 12

Drosophila Genomes. *Genetics* **179**: 1657–1680.

Blattner FR, Plunkett G 3rd, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, et al. 1997. The complete genome sequence of Escherichia coli K-12. *Science* **277**: 1453–1462.

Bridges CB. 1935. Salivary chromosome maps with a key to the banding of the chromosomes of Drosophila melanogaster. *J Hered* **26**: 60–64.

Britten RJ. 2004. Coding sequences of functioning human genes derived entirely from mobile element sequences. *Proc Natl Acad Sci U S A* **101**: 16825–16830.

Brown JB, Boley N, Eisman R, May GE, Stoiber MH, Duff MO, Booth BW, Wen J, Park S, Suzuki AM, et al. 2014. Diversity and dynamics of the Drosophila transcriptome. *Nature*.

C. elegans Sequencing Consortium. 1998. Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* **282**: 2012–2018.

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.

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.

Calabria G, Dolgova O, Rego C, Castañeda LE, Rezende EL, Balanyà J, Pascual M, Sørensen JG, Loeschcke V, Santos M. 2012. Hsp70 protein levels and thermotolerance in Drosophila subobscura: a reassessment of the thermal co-adaptation hypothesis. *J Evol Biol* **25**: 691–700.

Calvete O, González J, Betrán E, Ruiz A. 2012. Segmental duplication, microinversion, and gene loss associated with a complex inversion breakpoint region in Drosophila. *Mol Biol Evol* **29**: 1875–1889.

Campos JL, Halligan DL, Haddrill PR, Charlesworth B. 2014. The Relation between Recombination Rate and Patterns of Molecular Evolution and Variation in Drosophila melanogaster. *Mol Biol Evol* **31**: 1010–1028.

Casacuberta E, González J. 2013. The impact of transposable elements in environmental adaptation. *Mol Ecol* **22**: 1503–1517.

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.

Casals F, González J, Ruiz A. 2006. Abundance and chromosomal distribution of six Drosophila buzzatii transposons: BuT1, BuT2, BuT3, BuT4, BuT5, and BuT6. *Chromosoma* **115**: 403–412.

Casals F, Navarro A. 2007. Chromosomal evolution: Inversions: the chicken or the egg? *Heredity* **99**: 479–480.

Casola C, Hucks D, Feschotte C. 2008. Convergent domestication of pogo-like transposases into centromere-binding proteins in fission yeast and mammals. *Mol Biol Evol* **25**: 29–41.

Casola C, Lawing AM, Betrán E, Feschotte C. 2007. PIF-like transposons are common in drosophila and have been repeatedly domesticated to generate new host genes. *Mol Biol Evol* **24**: 1872–1888.

Celniker SE, Dillon LAL, Gerstein MB, Gunsalus KC, Henikoff S, Karpen GH, Kellis M, Lai EC, Lieb JD, MacAlpine DM, et al. 2009. Unlocking the secrets of the genome. *Nature* **459**: 927–930.

Chain PSG, Grafham DV, Fulton RS, FitzGerald MG, Hostetler J, Muzny D, Ali J, Birren B, Bruce DC, Buhay C, et al. 2009. Genome Project Standards in a New Era of Sequencing. *Science* **326**: 236–237.

Charlesworth B, Charlesworth D. 2000. The degeneration of Y chromosomes. *Philos Trans R Soc Lond B Biol Sci* **355**: 1563–1572.

Charlesworth B, Coyne JA, Barton NH. 1987. The Relative Rates of Evolution of Sex Chromosomes and Autosomes. *Am Nat* **130**: 113–46.

Chen B, Walser JC, Rodgers TH, Sobota RS, Burke MK, Rose MR, Feder ME. 2007. Abundant, diverse, and consequential P elements segregate in promoters of small heat-shock genes in Drosophila populations. *J Evol Biol* **20**: 2056–2066.

Cirera S, Martin-Campos JM, Segarra C, Aguade M. 1995. Molecular Characterization of the Breakpoints of an Inversion fixed between D. melanogaster and D. suboscura. *Genetics* **139**: 321-326.

Coghlan A, Eichler EE, Oliver SG, Paterson AH, Stein L. 2005. Chromosome evolution in eukaryotes: a multi-kingdom perspective. *Trends Genet* **21**: 673–682.

Coghlan A, Wolfe HK. 2002. Fourfold Faster Rate of Genome Rearrangement in Nematodes Than in Drosophila. *Genome Res* **12**:857-867.

Conrad B, Antonarakis SE. 2007. Gene duplication: a drive for phenotypic diversity and cause of human disease. *Annu Rev Genomics Hum Genet* **8**: 17–35.

Cordaux R, Batzer MA. 2009. The impact of retrotransposons on human genome evolution. *Nat Rev Genet* **10**: 691–703.

Cordaux R, Udit S, Batzer MA, Feschotte C. 2006. Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. *Proc Natl Acad Sci U S A* **103**: 8101–8106.

Corradi N, Pombert J-F, Farinelli L, Didier ES, Keeling PJ. 2010. The complete sequence of the smallest known nuclear genome from the microsporidian Encephalitozoon intestinalis. *Nat Commun* **1**: 77.

Coulibaly MB, Lobo NF, Fitzpatrick MC, Kern M, Grushko O, Thaner DV, Traoré SF, Collins FH, Besansky NJ. 2007. Segmental duplication implicated in the genesis of inversion 2Rj of Anopheles gambiae. *PloS One* **2**: e849.

Counterman BA, Ortíz-Barrientos D, Noor MAF. 2004. Using comparative genomic data to test for fast-X evolution. *Evol Int J Org Evol* **58**: 656–660.

Cui L, Wall PK, Leebens-Mack JH, Lindsay BG, Soltis DE, Doyle JJ, Soltis PS, Carlson JE, Arumuganathan K, Barakat A, et al. 2006. Widespread genome duplications throughout the history of flowering plants. *Genome Res* **16**: 738–749.

Delprat A, Negre B, Puig M, Ruiz A. 2009. The transposon Galileo generates natural chromosomal inversions in Drosophila by ectopic recombination. *PloS One* **4**: e7883.

Dobzhansky T. 1970. *Genetics of the Evolutionary Process*. Columbia University Press.

Dobzhansky T, Sturtevant A. 1938. Inversions in the Chromosomes of Drosophila Pseudoobscura. *Genetics* **23**: 28–64.

Drosophila 12 Genomes Consortium, Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA, Kaufman TC, Kellis M, Gelbart W, et al. 2007. Evolution of genes and genomes on the Drosophila phylogeny. *Nature* **450**: 203–218.

Dunning Hotopp JC, Clark ME, Oliveira DCSG, Foster JM, Fischer P, Muñoz Torres MC, Giebel JD, Kumar N, Ishmael N, Wang S, et al. 2007. Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. *Science* **317**: 1753–1756.

Duret L, Mouchiroud D. 1999. Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. *Proc Natl Acad Sci U S A* **96**: 4482–4487.

Eddy SR. 2001. Non–coding RNA genes and the modern RNA world. *Nat Rev Genet* **2**: 919–929.

Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, et al. 2009. Real-Time DNA Sequencing from Single Polymerase Molecules. *Science* **323**: 133–138.

Ellegren H. 2008. Comparative genomics and the study of evolution by natural selection. *Mol Ecol* **17**: 4586–4596.

ENCODE Project Consortium, Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57–74.

Etges WJ, Johnson WR, Duncan GA, Huckins G, Heed WB. 1999. Ecological Genetics of Cactophilic Drosophila. In *Ecology of Sonoran Desert plants and plant communities*, pp. 164–214, University of Arizona Press.

Eyre-Walker A. 2006. The genomic rate of adaptive evolution. *Trends Ecol Evol* **21**: 569–575.

Fangue NA, Hofmeister M, Schulte PM. 2006. Intraspecific variation in thermal tolerance and heat shock protein gene expression in common killifish, Fundulus heteroclitus. *J Exp Biol* **209**: 2859–2872.

Fares MA, Moya A, Escarmís C, Baranowski E, Domingo E, Barrio E. 2001. Evidence for positive selection in the capsid protein-coding region of the foot-and-mouth disease virus (FMDV) subjected to experimental passage regimens. *Mol Biol Evol* **18**: 10–21.

Farfán M, Miñana-Galbis D, Fusté MC, Lorén JG. 2009. Divergent evolution and purifying selection of the flaA gene sequences in Aeromonas. *Biol Direct* **4**: 23.

Feder JL, Nosil P. 2009. Chromosomal inversions and species differences: when are genes affecting adaptive divergence and reproductive isolation expected to reside within inversions? *Evolution* **63**: 3061–3075.

Feder JL, Roethele JB, Filchak K, Niedbalski J, Romero-Severson J. 2003. Evidence for inversion polymorphism related to sympatric host race formation in the apple maggot fly, Rhagoletis pomonella. *Genetics* **163**: 939–953.

Fedoroff NV. 2012. Transposable Elements, Epigenetics, and Genome Evolution. *Science* **338**: 758–767.

Fellows DP, Heed WB. 1972. Factors Affecting Host Plant Selection in Desert-Adapted Cactiphilic Drosophila. *Ecology* **53**: 850–858.

Feschotte C. 2008. Transposable elements and the evolution of regulatory networks. *Nat Rev Genet* **9**: 397–405.

Feuk L. 2010. Inversion variants in the human genome: role in disease and genome architecture. *Genome Med* **2**: 11.

Feuk L, Carson AR, Scherer SW. 2006. Structural variation in the human genome. *Nat Rev Genet* **7**: 85–97.

Finnegan DJ. 1989. Eukaryotic transposable elements and genome evolution. *Trends Genet TIG* **5**: 103–107.

Fiston-Lavier A-S, Anxolabehere D, Quesneville H. 2007. A model of segmental duplication formation in Drosophila melanogaster. *Genome Res* **17**: 1458–1470.

Fitch WM. 1970. Distinguishing Homologous from Analogous Proteins. *Syst Biol* **19**: 99–113.

Fogleman JC, Armstrong L. 1989. Ecological aspects of cactus triterpene glycosides I. Their effect on fitness components ofDrosophila mojavensis. *J Chem Ecol* **15**: 663–676.

Fogleman JC, Danielson PB. 2001. Chemical Interactions in the Cactus-Microorganism-Drosophila Model System of the Sonoran Desert1. *Am Zool* **41**: 877–889.

Fogleman JC, Kircher HW. 1986. Differential effects of fatty acid chain length on the viability of two species of cactophilic Drosophila. *Comp Biochem Physiol A Physiol* **83**: 761–764.

Fontdevila A, Ruiz A, Alonso G, Ocana J. 1981. Evolutionary History of Drosophila buzzatii. I. Natural Chromosomal Polymorphism in Colonized Populations of the Old World. *Evolution* **35**: 148.

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**: 1531–1545.

Fox J, Kling J. 2010. Chinese institute makes bold sequencing play. *Nat Biotechnol* **28**: 189–191.

Frischer LE, Hagen FS, Garber RL. 1986. An inversion that disrupts the Antennapedia gene causes abnormal structure and localization of RNAs. *Cell* **47**: 1017–1023.

Furuta Y, Kawai M, Yahara K, Takahashi N, Handa N, Tsuru T, Oshima K, Yoshida M, Azuma T, Hattori M, et al. 2011. Birth and death of genes linked to chromosomal inversion. *Proc Natl Acad Sci U S A* **108**: 1501–1506.

Goff SA, Ricke D, Lan T-H, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, et al. 2002. A Draft Sequence of the Rice Genome (Oryza sativa L. ssp. japonica). *Science* **296**: 92–100.

Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, et al. 1996. Life with 6000 Genes. *Science* **274**: 546–567.

Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol* **11**: 725–736.

González J, Casals F, Ruiz A. 2007. Testing chromosomal phylogenies and inversion breakpoint reuse in Drosophila. *Genetics* **175**: 167–177.

Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin JM, Yang L, Artieri CG, van Baren MJ, Boley N, Booth BW, et al. 2011. The developmental transcriptome of Drosophila melanogaster. *Nature* **471**: 473–479.

Graves JAM. 2006. Sex chromosome specialization and degeneration in mammals. *Cell* **124**: 901–914.

Gray YH. 2000. It takes two transposons to tango: transposable-element-mediated chromosomal rearrangements. *Trends Genet TIG* **16**: 461–468.

Gregory TR. 2014. *Animal Genome Size Database*. http://www.genomesize.com.

Gregory TR. 2005a. Synergy between sequence and size in large-scale genomics. *Nat Rev Genet* **6**: 699–708.

Gregory TR. 2005b. The C-value enigma in plants and animals: a review of parallels and an appeal for partnership. *Ann Bot* **95**: 133–146.

Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A. 2005. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res* **33**: D121–D124.

Guerzoni D, McLysaght A. 2011. De novo origins of human genes. *PLoS Genet* **7**: e1002381.

Guillén Y, Ruiz A. 2012. Gene alterations at Drosophila inversion breakpoints provide prima facie evidence for natural selection as an explanation for rapid chromosomal evolution. *BMC Genomics* **13**: 53.

Hainer SJ, Martens JA. 2011. Transcription of ncDNA. *Transcription* **2**: 120–123.

Hartl DL. 2000. Molecular melodies in high and low C. *Nat Rev Genet* **1**: 145–149.

Hartl DL, Clark AG. 1997. *Principle of Popupaltion Genetics*. Sinauer Associates, Inc., Sunderland, Massachusetts.

Hartwell L. 2011. *Genetics: from genes to genomes*. McGraw-Hill, New York.

Hasson E, Naveira H, Fontdevila A. 1992. The breeding sites of Argentinian cactophilic species of the Drosophila mulleri complex (subgenus Drosophila-repleta group). *Rev Chilena de Hist Nat* **65**: 319–326.

Heed WB. 1978. Ecology and Genetics of Sonoran Desert Drosophila. In *Ecological Genetics: The Interface* (ed. P.F. Brussard), *Proceedings in Life Sciences*, pp. 109–126, Springer New York.

Heed WB, Mangan RL. 1986. Community ecology of the Sonoran Desert Drosophila. In *The genetics and biology of Drosophila*, Vol. 3e of, Academic Press, London.

Heger A, Ponting CP. 2007. Evolutionary rate analyses of orthologs and paralogs from 12 Drosophila genomes. *Genome Res* **17**: 1837–1849.

Henikoff S. 1990. Position-effect variegation after 60 years. *Trends Genet TIG* **6**: 422–426.

Hillier LW, Miller W, Birney E, Warren W, Hardison RC, Ponting CP, Bork P, Burt DW, Groenen MAM, Delany ME, et al. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**: 695–716.

Hiraizumi Y. 1971. Spontaneous recombination in Drosophila melanogaster males. *Proc Natl Acad Sci U S A* **68**: 268–270.

Hoeijmakers WAM, Bártfai R, Stunnenberg HG. 2013. Transcriptome analysis using

RNA-Seq. *Methods Mol Biol* **923**: 221–239.

Hoffmann AA, Rieseberg LH. 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.

Hoffmann AA, Sgrò CM, Weeks AR. 2004. Chromosomal inversion polymorphisms and adaptation. *Trends Ecol Evol* **19**: 482–488.

Hoffmann AA, Sørensen JG, Loeschcke V. 2003. Adaptation of Drosophila to temperature extremes: bringing together quantitative and molecular approaches. *J Therm Biol* **28**: 175–216.

Hoffmann AA, Willi Y. 2008. Detecting genetic responses to environmental change. *Nat Rev Genet* **9**: 421–432.

Horton IH. 1938. A comparison of the salivary gland chromosomes of Drosophila melanogaster and D. simulans. *Genetics* **24**: 234–243.

Huang L-H, Kang L. 2007. Cloning and interspecific altered expression of heat shock protein genes in two leafminer species in response to thermal stress. *Insect Mol Biol* **16**: 491–500.

Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C. 2004. Detection of large-scale variation in the human genome. *Nat Genet* **36**: 949–951.

Innocenti P, Morrow EH, Hurst LD. 2010. The Sexually Antagonistic Genes of Drosophila melanogaster. *PLoS Biol* **8**: e1000335.

Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, Swofford R, Pirun M, Zody MC, White S, et al. 2012. The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* **484**: 55–61.

Joron M, Frezal L, Jones RT, Chamberlain NL, Lee SF, Haag CR, Whibley A, Becuwe M, Baxter SW, Ferguson L, et al. 2011. Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature* **477**: 203–206.

Kaessmann H. 2010. Origins, evolution, and phenotypic impact of new genes. *Genome Res* **20**: 1313–1326.

Kale PG. 1969. The meiotic origin of spontaneous crossovers in Drosophila ananassae males. *Genetics* **62**: 123–133.

Kaminker JS, Bergman CM, Kronmiller B, Carlson J, Svirskas R, Patel S, Frise E, Wheeler DA, Lewis SE, Rubin GM, et al. 2002. The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective. *Genome Biol* **3**: research0084.

Kaufman TC, Lewis R, Wakimoto B. 1980. Cytogenetic analysis of chromosome 3 in Drosophila melanogaster: The homoeotic gene complex in polytene chromosome interval. *Genetics* **94**: 115–133.

Keeling PJ, Palmer JD. 2008. Horizontal gene transfer in eukaryotic evolution. *Nat Rev Genet* **9**: 605–618.

Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TCG. 2009. More than just orphans: are taxonomically-restricted genes important in evolution? *Trends Genet* **25**: 404–413.

Kidwell, Lisch. 2000. Transposable elements and host genome evolution. *Trends Ecol Evol* **15**: 95–99.

Kidwell MG. 2002. Transposable elements and the evolution of genome size in eukaryotes. *Genetica* **115**: 49–63.

Kimura M. 1968. Evolutionary Rate at the Molecular Level. *Nature* **217**: 624–626.

Kimura M. 1977. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature* **267**: 275–276.

Kimura M. 1983. *The Neutral Theory of Molecular Evolution*. Cambridge University Press.

Kircher HW. 1982. Chemical composition of cacti and its relationship to Sonoran Desert Drosophila. In *Ecological Genetics and Evolution: The Cactus-Yeast-Drosophila Model System*, pp. 143–158, Academic Press, Sydney, Australia.

Kircher HW, Heed WB, Russell JS, Grove J. 1967. Senita cactus alkaloids: their significance to Sonoran Desert ecology. *J Insect Physiol* **13**: 1869–1874.

Kirkpatrick M. 2010. How and Why Chromosome Inversions Evolve. *PLoS Biol* **8**: e1000501.

Kirkpatrick M, Barton N. 2006. Chromosome Inversions, Local Adaptation and Speciation. *Genetics* **173**: 419–434.

Kmita M, Duboule D. 2003. Organizing axes in time and space; 25 years of colinear tinkering. *Science* **301**: 331–333.

Knight CA, Vogel H, Kroymann J, Shumate A, Witsenber H, Mitchell-Olds T. 2006. Expression profiling and local adaptation of Boechera holboellii populations for water use efficiency across a naturally occurring water stress gradient *Mol Ecol* **15**: 1229-1237.

De Koning APJ, Gu W, Castoe TA, Batzer MA, Pollock DD. 2011. Repetitive Elements May Comprise Over Two-Thirds of the Human Genome. *PLoS Genet* **7**: e1002384.

Konopka R, Benzer S. 1971. Clock mutants of Drosophila melanogaster. *Proc Natl Acad Sci USA* **68**: 2112–6.

Koonin EV. 2005. Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet* **39**: 309–338.

Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, et al. 2007a. Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**: 420–426.

Korbel JO, Urban AE, Grubert F, Du J, Royce TE, Starr P, Zhong G, Emanuel BS, Weissman SM, Snyder M, et al. 2007b. Systematic prediction and validation of breakpoints associated with copy-number variants in the human genome. *Proc Natl Acad Sci U S A* **104**: 10110–10115.

Krassovsky K, Henikoff S. 2014. Distinct chromatin features characterize different classes of repeat sequences in Drosophila melanogaster. *BMC Genomics* **15**: 105.

Krimbas CB, Powell JR. 1992. *Drosophila Inversion Polymorphism*. CRC Press.

Kristensen DM, Wolf YI, Mushegian AR, Koonin EV. 2011. Computational methods for Gene Orthology inference. *Brief Bioinform* **12**: 379–391.

Ladoukakis E, Pereira V, Magny EG, Eyre-Walker A, Couso JP. 2011. Hundreds of putatively functional small open reading frames in Drosophila. *Genome Biol* **12**: R118.

Lai Z, Nakazato T, Salmaso M, Burke JM, Tang S, Knapp SJ, Rieseberg LH. 2005. Extensive chromosomal repatterning and the evolution of sterility barriers in hybrid sunflower species. *Genetics* **171**: 291–303.

Lakich D, Kazazian HH Jr, Antonarakis SE, Gitschier J. 1993. Inversions disrupting the factor VIII gene are a common cause of severe haemophilia A. *Nat Genet* **5**: 236–241.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, 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.

Lang M, Murat S, Clark AG, Gouppil G, Blais C, Matzkin LM, Guittard E, Yoshiyama-Yanagawa T, Kataoka H, Niwa R, et al. 2012. Mutations in the neverland gene turned Drosophila pachea into an obligate specialist species. *Science* **337**: 1658–1661.

Larracuente AM, Sackton TB, Greenberg AJ, Wong A, Singh ND, Sturgill D, Zhang Y, Oliver B, Clark AG. 2008. Evolution of protein-coding genes in Drosophila. *Trends Genet* **24**: 114–123.

Larsen PF, Nielsen EE, Williams T, Hemmer J, Chipman JK, Kruhoffer M, Gronkjaer P, George SG, Dryskjot L, Loeschcke V. 2007. Adaptive differences in gene expression in European flounders (Platichthys flesus) *Mol Ecol* **16**: 4674-4683.

Lemos B, Bettencourt BR, Meiklejohn CD, Hartl DL. 2005. Evolution of proteins and gene expression levels are coupled in Drosophila and are independently associated with mRNA abundance, protein length, and number of protein-protein interactions. *Mol Biol Evol* **22**: 1345–1354.

Lerman DN, Feder ME. 2005. Naturally occurring transposable elements disrupt hsp70 promoter function in Drosophila melanogaster. *Mol Biol Evol* **22**: 776–783.

Lerman DN, Michalak P, Helin AB, Bettencourt BR, Feder ME. 2003. Modification of heat-shock gene expression in Drosophila melanogaster populations via transposable elements. *Mol Biol Evol* **20**: 135–144.

Lewis EB. 1978. A gene complex controlling segmentation in Drosophila. *Nature* **276**: 565–570.

Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, Kheradpour P, Ernst J, Jordan G, Mauceli E, et al. 2011. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**: 476–482.

Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, Kamal M, Clamp M, Chang JL, Kulbokas EJ, Zody MC, et al. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**: 803–819.

Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, Pallen MJ. 2012. Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotechnol* **30**: 434–439.

Long M. 2000. A New Function Evolved from Gene Fusion. *Genome Res* **10**: 1655–1657.

Long M, Betrán E, Thornton K, Wang W. 2003. The origin of new genes: glimpses from the young and old. *Nat Rev Genet* **4**: 865–875.

Long M, Langley CH. 1993. Natural selection and the origin of jingwei, a chimeric processed functional gene in Drosophila. *Science* **260**: 91–95.

Long M, VanKuren NW, Chen S, Vibranovski MD. 2013. New Gene Evolution: Little Did We Know. *Annu Rev Genet* **47**: 307–333.

Lowry DB, Willis JH. 2010. A widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation. *PLoS Biol* **8**.

Lynch M. 2007. *The origins of genome architecture*. Sinauer Associates.

Lynch M, Walsh B. 1998. *Genetics and analysis of quantitative traits*. Sinauer, Sunderland, Mass.

Mackay TFC, Richards S, Stone EA, Barbadilla A, Ayroles JF, Zhu D, Casillas S, Han Y, Magwire MM, Cridland JM, et al. 2012. The Drosophila melanogaster Genetic Reference Panel. *Nature* **482**: 173–178.

Mani R-S, Chinnaiyan AM. 2010. Triggers for genomic rearrangements: insights into genomic, cellular and environmental influences. *Nat Rev Genet* **11**: 819–829.

Marchler-Bauer A, Bryant SH. 2004. CD-Search: protein domain annotations on the fly. *Nucleic Acids Res* **32**: W327–331.

Mardis ER. 2008. The impact of next-generation sequencing technology on genetics. Trends Genet.24: 133-141

Markow TA, O'Grady PM. 2007. Drosophila biology in the genomic age. *Genetics* **177**: 1269–1276.

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.

Mattick JS. 2004. RNA regulation: a new genetics? *Nat Rev Genet* **5**: 316–323.

Mattick JS, Makunin IV. 2006. Non-coding RNA. *Hum Mol Genet* **15**: R17–R29.

Matzkin LM. 2012. Population transcriptomics of cactus host shifts in Drosophila mojavensis. *Mol Ecol* **21**: 2428–2439.

Matzkin LM, Markow TA. 2013. Transcriptional differentiation across the four subspecies of drosopihla mojavensis. In *Speciation: Natural Processes, Genetics and Biodiversity*, Nova Scientific Publishers, New York.

Matzkin LM, Merritt TJS, Zhu C-T, 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.

Mazo A, Hodgson JW, Petruk S, Sedkov Y, Brock HW. 2007. Transcriptional interference: an unexpected layer of complexity in gene regulation. *J Cell Sci* **120**: 2755–2761.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in Drosophila. *Nature* **351**: 652–654.

McGinnis N, Kuziora MA, McGinnis W. 1990. Human Hox-4.2 and Drosophila deformed encode similar regulatory specificities in Drosophila embryos and larvae. *Cell* **63**: 969–976.

McGinnis W. 1994. A century of homeosis, a decade of homeoboxes. *Genetics* **137**: 607–611.

Medstrand P, van de Lagemaat LN, Dunn CA, Landry J-R, Svenback D, Mager DL. 2005. Impact of transposable elements on the evolution of mammalian gene regulation. *Cytogenet Genome Res* **110**: 342–352.

Medvedev P, Stanciu M, Brudno M. 2009. Computational methods for discovering structural variation with next-generation sequencing. *Nat Methods* **6**: S13–20.

Mercer TR, Dinger ME, Mattick JS. 2009. Long non-coding RNAs: insights into functions. *Nat Rev Genet* **10**: 155–159.

Messer PW, Petrov DA. 2013. Frequent adaptation and the McDonald–Kreitman test. *Proc Natl Acad Sci* **110**: 8615–20.

Michalak P, Minkov I, Helin A, Lerman DN, Bettencourt BR, Feder ME, Korol AB, Nevo E. 2001. Genetic evidence for adaptation-driven incipient speciation of Drosophila melanogaster along a microclimatic contrast in "Evolution Canyon," Israel. *Proc Natl Acad Sci* **98**: 13195–13200.

Miller JM, Malenfant RM, Moore SS, Coltman DW. 2012. Short reads, circular genome: skimming solid sequence to construct the bighorn sheep mitochondrial genome. *J Hered* **103**: 140–146.

Mitelman F, Johansson B, Mertens F. 2007. The impact of translocations and gene fusions on cancer causation. *Nat Rev Cancer* **7**: 233–245.

Miyata T, Yasunaga T. 1980. Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. *J Mol Evol* **16**: 23–36.

modENCODE Consortium, Roy S, Ernst J, Kharchenko PV, Kheradpour P, Negre N, Eaton ML, Landolin JM, Bristow CA, Ma L, et al. 2010. Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science* **330**: 1787–1797.

Morgan TH. 1914. No Crossing over in the Male of Drosophila of Genes in the Second and Third Pairs of Chromosomes. *Biol Bull* **26**: 195–204.

Morgan TH. 1910. Sex limited inheritance in Drosophila. *Science* **32**: 120–122.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**: 621–628.

Muller HJ, Painter TS. 1932. The differentiation of sex chromosomes of Drosophila into genetically active and inert regions. *Z.iAV* **62**: 316–365.

Muotri AR, Marchetto MCN, Coufal NG, Gage FH. 2007. The necessary junk: new functions for transposable elements. *Hum Mol Genet* **16**: R159–R167.

Muse SV, Gaut BS. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol Biol Evol* **11**: 715–724.

Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. 2008. The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. *Science* **320**: 1344–1349.

Nagarajan N, Pop M. 2013. Sequence assembly demystified. *Nat Rev Genet* **14**: 157–167.

Negre B, Casillas S, Suzanne M, Sánchez-Herrero E, Akam M, Nefedov M, Barbadilla A, de Jong P, Ruiz A. 2005. Conservation of regulatory sequences and gene expression patterns in the disintegrating

Drosophila Hox gene complex. *Genome Res* **15**: 692–700.

Negre B, Ruiz A. 2007. HOM-C evolution in Drosophila: is there a need for Hox gene clustering? *Trends Genet* **23**: 55–59.

Neves G, Zucker J, Daly M, Chess A. 2004. Stochastic yet biased expression of multiple Dscam splice variants by individual cells. *Nat Genet* **36**: 240–246.

Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, Hubisz MJ, Fledel-Alon A, Tanenbaum DM, Civello D, White TJ, et al. 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol* **3**: e170.

Nielsen R, Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148**: 929–936.

Nurminsky DI, Nurminskaya MV, Aguiar DD, Hartl DL. 1998. Selective sweep of a newly evolved sperm-specific gene in Drosophila. *Nature* **396**: 572–575.

Ohno S. 1970. *Evolution by gene duplication.* Allen & Unwin; Springer-Verlag, London; New York.

Ohta T. 1973. Slightly Deleterious Mutant Substitutions in Evolution. *Nature* **246**: 96–98.

Oliveira DCSG, Almeida FC, O'Grady PM, Armella MA, DeSalle R, Etges WJ. 2012. Monophyly, divergence times, and evolution of host plant use inferred from a revised phylogeny of the Drosophila repleta species group. *Mol Phylogenet Evol* **64**: 533–544.

Ometto L, Cestaro A, Ramasamy S, Grassi A, Revadi S, Siozios S, Moretto M, Fontana P, Varotto C, Pisani D, et al. 2013. Linking genomics and ecology to investigate the complex evolution of an invasive Drosophila pest. *Genome Biol Evol* **5**: 745–757.

Pagani I, Liolios K, Jansson J, Chen I-MA, Smirnova T, Nosrat B, Markowitz VM, Kyrpides NC. 2012. The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res* **40**: D571–579.

Papaceit M, Segarra C, Aguadé M. Structure and population genetics of the breakpoints of a polymorphic inversion in Drosophila subobscura. *Evolution* **67**: 66-79

Parsch J, Ellegren H. 2013. The evolutionary causes and consequences of sex-biased gene expression. *Nat Rev Genet* **14**: 83–87.

Patterson JT, Stone WS. 1953. *Evolution in the Genus Drosophila*. MacMillan Co., New York.

Pellicer J, Fay MF, Leitch IJ. 2010. The largest eukaryotic genome of them all? *Bot J Linn Soc* **164**: 10–15.

Petersen L, Bollback JP, Dimmic M, Hubisz M, Nielsen R. 2007. Genes under positive selection in Escherichia coli. *Genome Res* **17**: 1336–1343.

Philip U. 1944. Crossing overs in the males of D. subobscura. *Nature* **153**: 233.

Phillips T, Hoopes L. 2008. Transcription factors and transcriptional control in eukaryotic cells. *Nat Educ* **1**: 119.

Plotkin JB, Kudla G. 2011. Synonymous but not the same: the causes and consequences of codon bias. *Nat Rev Genet* **12**: 32–42.

Powell JR. 1997. *Progress and prospects in evolutionary biology the Drosophila model*. Oxford University Press, New York.

Prazeres da Costa O, González J, Ruiz A. 2009. Cloning and sequencing of the breakpoint regions of inversion 5g fixed in Drosophila buzzatii. *Chromosoma* **118**: 349–360.

Presgraves DC. 2010. The molecular evolutionary basis of species formation. *Nat Rev Genet* **11**: 175–180.

Prud'homme B, Gompel N, Carroll SB. 2007. Emerging principles of regulatory evolution. *Proc Natl Acad Sci U S A* **104 Suppl 1**: 8605–8612.

Puig M. 2011. Functional analysis of position effects of inversion 2j inDrosophila buzzatii gene CG13617 silencing and its adaptative significance. Universitat Autònoma de Barcelona, Bellaterra.

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.

Ranz JM, Casals F, Ruiz A. 2001. How Malleable is the Eukaryotic Genome? Extreme Rate of Chromosomal Rearrangement in the Genus Drosophila. *Genome Res* **11**: 230–239.

Ranz JM, Maurin D, Chan YS, von Grotthuss M, Hillier LW, Roote J, Ashburner M, Bergman CM. 2007. Principles of genome evolution in the Drosophila melanogaster species group. *PLoS Biol* **5**: e152.

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, et al. 2006. Global variation in copy number in the human genome. *Nature* **444**: 444–454.

Reed J, Mishra B, Pittenger B, Magonov S, Troke J, Teitell MA, Gimzewski JK. 2007. Single molecule transcription profiling with AFM. *Nanotechnology* **18**: 44032.

Richards S, Liu Y, Bettencourt BR, Hradecky P, Letovsky S, Nielsen R, Thornton K, Hubisz MJ, Chen R, Meisel RP, et al. 2005. Comparative genome sequencing of Drosophila pseudoobscura: Chromosomal, gene, and cis-element evolution. *Genome Res* **15**: 1–18.

Riehle MM, Bennett AF, Long AD. 2005. Changes in gene expression following high-temperature adaptation in experimentally evolved populations of E. coli. *Physiol Biochem Zool* **78**: 299–315.

Rieseberg LH. 2001. Chromosomal rearrangements and speciation. *Trends Ecol Evol* **16**: 351–358.

Rius N, Delprat A, Ruiz A. 2013. A Divergent P Element and Its Associated MITE, BuT5, Generate Chromosomal Inversions and Are Widespread within the Drosophila repleta Species Group. *Genome Biol Evol* **5**: 1127–1141.

Roger AJ. 1999. Reconstructing Early Events in Eukaryotic Evolution. *Am Nat* **154**: S146–S163.

Rogers RL, Hartl DL. 2012. Chimeric Genes as a Source of Rapid Evolution in Drosophila melanogaster. *Mol Biol Evol* **29**: 517–529.

Romanish MT, Lock WM, van de Lagemaat LN, Dunn CA, Mager DL. 2007. Repeated recruitment of LTR retrotransposons as promoters by the anti-apoptotic locus NAIP during mammalian evolution. *PLoS Genet* **3**: e10.

Rubin GM, Lewis EB. 2000. A Brief History of Drosophila's Contributions to Genome Research. *Science* **287**: 2216–2218.

Ruiz A, Cansian AM, Kuhn GC, Alves MA, Sene FM. 2000. The Drosophila serido speciation puzzle: putting new pieces together. *Genetica* **108**: 217–227.

Ruiz A, Heed WB. 1988. Host-Plant Specificity in the Cactophilic Drosophila mulleri Species Complex. *J Anim Ecol* **57**: 237–249.

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.

Ruiz A, Wasserman M. 1993. Evolutionary cytogenetics of the Drosophila buzzatii species complex. *Heredity* **70**: 582–596.

Runcie DE, Noor MAF. 2009. Sequence signatures of a recent chromosomal rearrangement in Drosophila mojavensis. *Genetica* **136**: 5–11.

Russo CAM, Mello B, Frazão A, Voloch CM. 2013. Phylogenetic analysis and a time tree for a large drosophilid data set (Diptera: Drosophilidae). *Zool J Linn Soc* **169**: 765–775.

Saiki RK, Gelfand DH, Stoffel S, Scharf SJ, Higuchi R, Horn GT, Mullis KB, Erlich HA. 1988. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* **239**: 487–491.

Sawaya MR, Wojtowicz WM, Andre I, Qian B, Wu W, Baker D, Eisenberg D, Zipursky SL. 2008. A Double S Shape Provides the Structural Basis for the Extraordinary Binding Specificity of Dscam Isoforms. *Cell* **134**: 1007–1018.

Schattner P. 2008. *Genomes, Browsers and Databases: Data-Mining Tools for Integrated Genomic Databases*. 1 edition. Cambridge University Press, Cambridge UK ; New York.

Schneider D. 2000. Using Drosophila as a model insect. *Nat Rev Genet* **1**: 218–226.

Sharakhov IV, White BJ, Sharakhova MV, Kayondo J, Lobo NF, Santolamazza F, Della Torre A, Simard F, Collins FH, Besansky NJ. 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.

Sharp PM, Averof M, Lloyd AT, Matassi G, Peden JF. 1995. DNA sequence evolution: the sounds of silence. *Philos Trans R Soc Lond B Biol Sci* **349**: 241–247.

Shilova VY, Garbuz DG, Myasyankina EN, Chen B, Evgen’ev MB, Feder ME, Zatsepina OG. 2006. Remarkable Site Specificity of Local Transposition Into the Hsp70 Promoter of Drosophila melanogaster. *Genetics* **173**: 809–820.

Singh ND, Larracuente AM, Clark AG. 2008. Contrasting the efficacy of selection on the X and autosomes in Drosophila. *Mol Biol Evol* **25**: 454–467.

Singh ND, Larracuente AM, Sackton TB, Clark AG. 2009. Comparative Genomics on the Drosophila Phylogenetic Tree. *Annu Rev Ecol Evol Syst* **40**: 459–480.

Smith G, Fany Y, Liu X, Kenny J, Cossins AR, de Oliveira C, Etges WJ, Ritchie MG. 2013. Transcriptome-wide expression variation associated with environmental plasticity and mating success in cactophilic Drosophila mojavensis. *Evolution* **67**: 1950-1963.

Sonoda E, Hochegger 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* **5**: 1021–1029.

Sperlich D, Pfreim P. 1986. Chromosomal polymorphism in natural and experimental poopulations. In *The genetics and biology of Drosophila* (eds. M. Ashburner, H. Carson, and J. Thompson), pp. 257–309, M , H.L. Carson, J.N. Thompson Jr., London.

Stapley J, Reger J, Feulner PGD, Smadja C, Galindo J, Ekblom R, Bennison C, Ball AD, Beckerman AP, Slate J. 2010. Adaptation genomics: the next generation. *Trends Ecol Evol* **25**: 705–712.

Stefansson H, Helgason A, Thorleifsson G, Steinthorsdottir V, Masson G, Barnard J, Baker A, Jonasdottir A, Ingason A, Gudnadottir VG, et al. 2005. A common inversion under selection in Europeans. *Nat Genet* **37**: 129–137.

Stein LD. 2013. Using GBrowse 2.0 to visualize and share next-generation sequence data. *Brief Bioinform* **14**: 162–171.

Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, et al. 2002. The Generic Genome Browser: A Building Block for a Model Organism System Database. *Genome Res* **12**: 1599–1610.

Straalen NM van, Roelofs, Dick. 2012. *An introduction to ecological genomics*. Oxford University Press, New York.

Su Z, Wang J, Yu J, Huang X, Gu X. 2006. Evolution of alternative splicing after gene duplication. *Genome Res* **16**: 182–189.

Swanson WJ, Yang Z, Wolfner MF, Aquadro CF. 2001. Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals. *Proc Natl Acad Sci U S A* **98**: 2509–2514.

Swift H. 1950. The Constancy of Desoxyribose Nucleic Acid in

Plant Nuclei. *Proc Natl Acad Sci U S A* **36**: 643–654.

Tamura K, Subramanian S, Kumar S. 2004. Temporal patterns of fruit fly (Drosophila) evolution revealed by mutation clocks. *Mol Biol Evol* **21**: 36–44.

The FlyBase Consortium. 2002. The FlyBase database of the Drosophila genome projects and community literature. *Nucleic Acids Res* **30**: 106–108.

Thomas CA. 1971. The Genetic Organization of Chromosomes. *Annu Rev Genet* **5**: 237–256.

Thornton K, Long M. 2002. Rapid divergence of gene duplicates on the Drosophila melanogaster X chromosome. *Mol Biol Evol* **19**: 918–925.

Toll-Riera M, Castelo R, Bellora N, Albà MM. 2009. Evolution of primate orphan proteins. *Biochem Soc Trans* **37**: 778-782.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**: 511–515.

Treangen TJ, Salzberg SL. 2012. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet* **13**: 36–46.

Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, Haugen E, Hayden H, Albertson D, Pinkel D, et al. 2005. Fine-scale structural variation of the human genome. *Nat Genet* **37**: 727–732.

Vicoso B, Charlesworth B. 2009. Effective Population Size and the Faster-X Effect: An Extended Model. *Evolution* **63**: 2413–2426.

Vilela CR. 1983. A revision of the Drosophila repleta species group (Diptera, Drosophilidae). *Revta Bras Ent* **27**: 1–114.

Villanueva-Cañas JL, Laurie S, Albà MM. 2013. Improving genome-wide scans of positive selection by using protein isoforms of similar length **5**:457-467.

Walser J-C, Chen B, Feder ME. 2006. Heat-shock promoters: targets for evolution by P transposable elements in Drosophila. *PLoS Genet* **2**: e165.

Wang W, Yu H, Long M. 2004. Duplication-degeneration as a mechanism of gene fission and the origin of new genes in Drosophila species. *Nat Genet* **36**: 523–527.

Wang Z, Gerstein M, Snyder M. 2009. RNA-seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**: 57-63.

Wasserman M. 1992. Cytological evolution of the Drosophila repleta species group. In *Drosophila inversion polymorphism*, pp. 455–552, CRC Press, Boca Raton, FL.

Wasserman M. 1982. Evolution of the repleta group. In *The genetics and biology of Drosophila*, Vol. 3b of, pp. 61–139, Academic Press, London.

Waterston RH, Cook LL, Delehaunty KD, Fewell GA, Fulton LA, Fulton RS, Graves TA, Hillier LW, Mardis ER, McPherson JD, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**: 520–562.

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* **91**: 3132–3136.

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, 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.

Wilkins AS. 1998. Evolutionary developmental biology: where is it going? *BioEssays* **20**: 783–784.

Wilson GA, Bertrand N, Patel Y, Hughes JB, Feil EJ, Field D. 2005. Orphans as taxonomically restricted and ecologically important genes. *Microbiol Read Engl* **151**: 2499–2501.

Yang, Bielawski. 2000. Statistical methods for detecting molecular adaptation. *Trends Ecol Evol* **15**: 496–503.

Yang Z. 2002. Inference of selection from multiple species alignments. *Curr Opin Genet Dev* **12**: 688–694.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**: 1586–1591.

Yang Z, Nielsen R, Goldman N, Pedersen AM. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155**: 431–449.

Zanotto PM de A, Kallas EG, Souza RF de, Holmes EC. 1999. Genealogical Evidence for Positive Selection in the nef Gene of HIV-1. *Genetics* **153**: 1077–1089.

Zatsepina OG, Velikodvorskaia VV, Molodtsov VB, Garbuz D, Lerman DN, Bettencourt BR, Feder ME, Evgenev MB. 2001. A

DROSOPHILA MELANOGASTER Strain From Sub-Equatorial Africa Has Exceptional Thermotolerance But Decreased Hsp70 Expression. *J Exp Biol* **204**: 1869–1881.

Zhao L, Saelao P, Jones CD, Begun DJ. 2014. Origin and Spread of de Novo Genes in Drosophila melanogaster Populations. *Science* **343**: 769–772.

Zhou Q, Bachtrog D. 2012. Sex-specific adaptation drives early sex chromosome evolution in Drosophila. *Science* **337**: 341–345.

Zhou Q, Zhang G, Zhang Y, Xu S, Zhao R, Zhan Z, Li X, Ding Y, Yang S, Wang W. 2008. On the origin of new genes in Drosophila. *Genome Res* **18**: 1446–1455.

Index of tables

TABLE 1. Summary of genome properties of different organisms sequenced between 1996 and 2005. 6
TABLE 2. Summary of polymorphic inversions with characterized breakpoints in Drosophila and Anopheles. 27
TABLE 3. Comparison of different sequencing platforms. 203
TABLE 4. Genomic determinants of protein-coding gene evolution in Drosophila mojavensis and Drosophila buzzatii. 214
TABLE 5. Genes evolving under positive selection in D. mojavensis lineage involved in heterocycle catabolic processes. 220

Index of figures

FIGURE 1 8
FIGURE 3 12
FIGURE 4 15
FIGURE 5 17
FIGURE 6 19
FIGURE 7 22
FIGURE 8. 24
FIGURE 9. 26
FIGURE 10. 28
FIGURE 11. 29
FIGURE 12. 33
FIGURE 13. 36
FIGURE 14. 205
FIGURE 15. 209
FIGURE 16. 216
FIGURE 17.. 224
FIGURE 18 225

ACKNOWLEDGEMENTS

Quisiera dar las gracias a todas las personas que me han apoyado durante todos estos años, compañeros de trabajo, familiares y amigos. En primer lugar gracias a Alfredo por darme la oportunidad de trabajar en su grupo y descubrirme el mundo de la Genética Evolutiva. Gracias a David, Maite, Miquel y Nuria por hacer que los días de trabajo fueran más llevaderos, por vuestros consejos, vuestras correcciones y por nuestras conversaciones. Os deseo mucha suerte. Gracias a Alejandra por sus ánimos, sobre todo los recibidos en la etapa final. Y gracias también a Elena por facilitarnos tanto la vida resolviendo nuestros problemas burocráticos.

Sin duda todo habría sido mucho más difícil sin el apoyo de mis amigas, que comprenden tan bien el trabajo y las responsabilidades que conlleva esta profesión. Gracias Ana G, Ana M, Ari, Belén, Diana, Teresa y Mariaje. Nos quedan muchas tesis, viajes y celebraciones por delante. Gracias a ti también Maria, por conocerme tan bien. Ojalá compartamos juntas muchos logros. Thank you Flora for the time we spent together in Ithaca, it was great to meet you when I was so far away from home. I wish you the best. Gracias Victori y Bea por hacer que las últimas horas frente al ordenador fueran más divertidas con vuestras risas y karaokes de fondo.

Gracias a mi hermana por estar siempre ahí. Nunca dejarás de ser mi ejemplo a seguir. Gracias también a César por sus consejos y por atender mis dudas. Sois los dos grandes doctores y sobre todo grandes personas. Gracias Miguel por todos tus ánimos y tu apoyo incondicional. Has estado a mi lado cuando más lo necesitaba y espero compartir contigo muchos años de felicidad.

Y por último muchas gracias a mis padres, por cuidarme y por darme cariño en todo momento. Sé que siempre podré contar con vosotros.