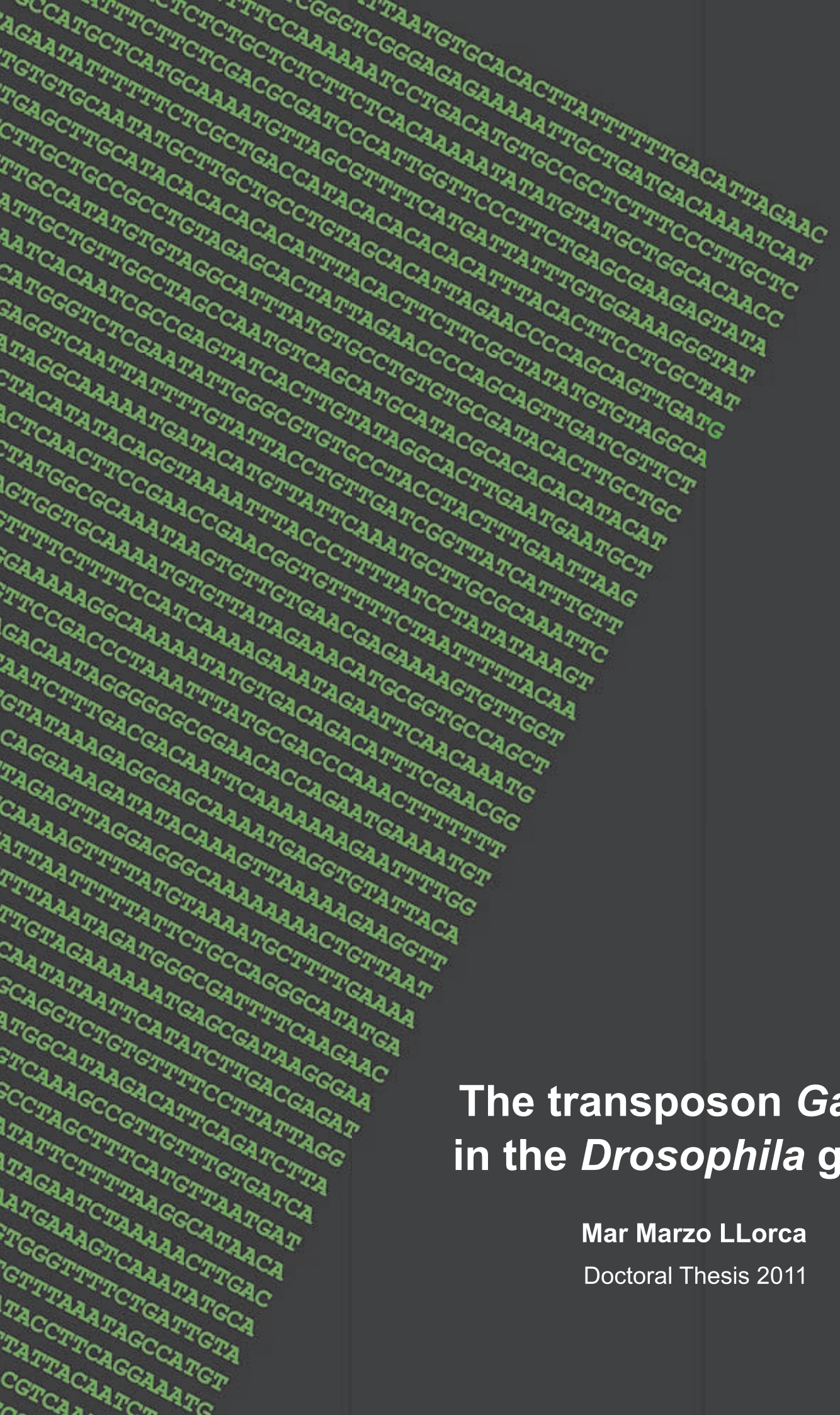


# The transposon *Galileo* in the *Drosophila* genus

**Mar Marzo LLorca**

Doctoral Thesis 2011

# The transposon *Galileo* in the *Drosophila* genus

Doctoral Thesis

Mar Marzo Llorca

Universitat Autònoma de Barcelona
Facultat de Biociències
Departament de Genètica i de Microbiologia
Bellaterra 2011

Memòria presentada per la Llicenciada en Biologia Mar Marzo Llorca per optar al grau de Doctora en Ciències Biològiques

Mar Marzo Llorca

Bellaterra, a 26 de Setembre de 2011

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

CERTIFICA: que la Mar Marzo Llorca ha dut a terme sota la seva direcció el treball de recerca realitzat al Departament de Genètica i de Microbiologia de la Facultat de Biociències de la Universitat Autònoma de Barcelona que ha portat a l'elaboració d'aquesta Tesi Doctoral, titulada "The transposon *Galileo* in the *Drosophila* genus".

I per què consti als efectes oportuns, signa el present certificat a Bellaterra, a 26 de Setembre de 2011.

Dr. Alfredo Ruiz Panadero

*To VSC*

# Table of Contents

# ABBREVIATIONS

| | |
|---|---|
| **Aa** | amino acid |
| **bp** | base pairs |
| **BS** | binding site |
| **CDS** | coding sequence |
| **Chr** | chromosome |
| **Dana** | *D. ananassae* |
| **Dbuz** | *D. buzzatii* |
| **Dere** | *D. erecta* |
| **Dgri** | *D. grimshawi* |
| **Dmel** | *D. melanogaster* |
| **Dmoj** | *D. mojavensis* |
| **DNA** | deoxyribonucleic acid |
| **Dper** | *D. persimilis* |
| **Dpse** | *D. pseudoobscura* |
| **Dsec** | *D. sechellia* |
| **Dsim** | *D. simulans* |
| **Dvir** | *D. virilis* |
| **Dwil** | *D. willistoni* |
| **Dyak** | *D. yakuba* |
| **EMSA** | electrophoretic mobility shift assay |
| **Enh** | enhancer |
| **F1** | transposon region between the TIR1 and the transposase coding region |
| **F2** | transposon region between the transposase coding region and the TIR2 |
| ***FB*** | Foldback |
| **GTP** | guanosine triphosphate |
| **HD** | hybrid dysgenesis |
| **kb** | kilobase |
| **MBP** | maltose binding protein |
| **LINE** | long interspersed element |
| **LTR** | long terminal repeat |
| **MITE** | miniature terminal inverted repeat element |
| **NAHR** | non-allelic homologous recom- bination |
| **NHEJ** | non-homologous end-joining |
| **nt** | nucleotide |
| **ORF** | open reading frame |
| **P** | probability (p-value) |
| **PCR** | polymerase chain reaction |
| **PLE** | Penelope-like element |
| **PK** | protein kinase |
| **Pol III** | RNA polymerase II I |
| **RNA** | ribonucleic acid |
| **RT** | retrotransciptase |
| **SDR** | split direct repeats |
| **SINE** | short interspersed element |
| **TE** | transposable element |
| **TIR** | terminal inverted repeat |
| **TSD** | target site duplication |

# Index of Tables

# Index of Figures

**ABSTRACT**. Transposable elements (TE) are repetitive sequences whose ability to change their location in the genome defines them. They made up a important proportion of the eukaryotic genomes, and although they are often considered as genetic parasites, it has been also argued that they might have some still unknown cellular function. Nevertheless, it is clear that they play a role as drivers of their host evolution, due to the fact that TEs generate genetic variability.

The TE *Galileo* is involved in the generation of adaptive chromosomal rearrangements in natural populations of *Drosophila buzzatii*, indicating that it would be a driver of adaptation in its host. Moreover, all *Galileo* elements found in previous works were incomplete – mainly composed by *Foldback*-like structures – and homology relationships could not be established with any known sequence. With this background, this thesis was proposed to characterise the mobile genetic element *Galileo* in different *Drosophila* species and analyse its evolutionary dynamics. Thus, in a first phase we searched for complete copies of *Galileo* in different species of the *Drosophila* genus: *D. buzzatii*, *D. mojavensis*, *D. virilis*, *D. willitoni*, *D. ananassae*, *D. pseudoobscura* and *D. persimilis*, using both bioinformatic and experimental methods (depending on whether the analysed genome was available or not). The copies found present long TIR (up to 1.2 Kb), high sequence identity with previously found *Galileo* sequences and, moreover, they harbour coding sequences that have allowed the classification of *Galileo* as a member of the *P-element* superfamily. Subsequently, by means of phylogenetic analyses, we have found that there are *Galileo* subfamilies in three different species (*D. buzzatii*, *D. mojavensis*, *D. virilis*) and evidence of recent transpositional activity (in *D. willitoni*, *D. ananassae*, *D. pseudoobscura, D. persimilis* and *D. mojavensis*). In a second phase of the thesis, we have conducted experiments with part of the *Galileo* protein and detected specific binding to the *Galileo* TIR, confirming that this sequence is responsible for the transposition reaction. Finally, we have thoroughly studied the *Galileo* variability in the *D. mojavensis* genome and found a striking structural variation, suggesting that the exchange of sequences among different *Galileo* copies might be quite common and important for TEs evolution.

**RESUM**. Els elements transposables (TEs) són seqüències repetitives amb el tret definitori de canviar la seva posició al genoma. Ocupen fraccions importants dels genomes eucariotes, y, tot i que solen considerar-se paràsits genètics, també s'especula amb la possibilitat de que tinguessin alguna funció cel·lular que encara ens és desconeguda. Tot i així, sembla evident que tenen un paper important com facilitadors de l'evolució, ja que generen variabilitat al genoma de l'hoste.

El TE *Galileo* està implicat en la generació de reordenacions cromosòmiques adaptatives naturals a l'espècie *Drosophila buzzatii*, en la que hauria generat variabilitat amb valor adaptatiu per a l'hoste. A més, tots els elements *Galileo* trobats en treballs anteriors eren defectius – composats bàsicament d'estructures similars a la dels elements *Foldback* – i no es van poder establir relacions d'homologia amb ninguna seqüència coneguda. Amb aquest rerefons, en aquesta tesi es va plantejar caracteritzar l'element genètic mòbil *Galileo* en diferents espècies de *Drosophila* i analitzar la seva dinàmica evolutiva. D'aquesta forma, en una primera fase es van buscar elements *Galileo* complets en diferents espècies del gènere *Drosophila*: *D. buzzatii*, *D. mojavensis*, *D. virilis*, *D. willitoni*, *D. ananassae*, *D. pseudoobscura* i *D. persimilis*, fent servir tant mètodes bioinformàtics com experimentals (depenent de si el genoma analitzat estava seqüenciat o no). Les còpies trobades presenten llargues Repeticions Invertides Terminals (TIR) de fins a 1,2 Kb, una elevada identitat amb seqüències de *Galileo* descrites anteriorment i, a més, contenen una zona codificant que ha permès classificar *Galileo* com a membre de la superfamília de l'*element P*. Posteriorment, mitjançant anàlisis filogenètiques, hem trobat l'existència de subfamílies de *Galileo* en tres espècies (*D. buzzatii*, *D. mojavensis*, *D. virilis*) i evidència d'activitat transposicional recent (*D. willitoni*, *D. ananassae*, *D. pseudoobscura, D. persimilis* i *D. mojavensis*). En una segona fase de la tesi, hem dut a terme experiments amb part de la proteïna que es codifica a *Galileo* i hem comprovat que interacciona amb les TIR de *Galileo*, confirmant que aquesta seqüència és la responsable de la reacció de transposició. Finalment, hem analitzat en detall la diversitat de *Galileo* al genoma de *D. mojavensis* i hem detectat una diversitat estructural molt important, on l'intercanvi de seqüències entre elements pareix força freqüent per l'evolució dels TEs.

**RESUMEN.** Los elementos transponibles (TEs) son secuencias repetitivas cuya característica definitoria es la capacidad de cambiar de posición en el genoma. Ocupan fracciones muy importantes de los genomas de eucariotas, y aunque se suelen considerar parásitos genéticos, también se especula con la posibilidad de que pudieran tener alguna función celular que aún nos es desconocida. No obstante, parece evidente que tienen un papel importante como facilitadores de la evolución, al generar variabilidad en el genoma del huésped.

El TE *Galileo* está implicado en la generación de reordenaciones cromosómicas adaptativas naturales en la especie *Drosophila buzzatii*, con lo que habría generado variabilidad adaptativa para el huésped. Además, todos los elementos *Galileo* encontrados en trabajos anteriores eran defectivos – compuestos básicamente de estructuras similares a las de los elementos *Foldback* – y no se pudieron establecer relaciones de homología con ninguna secuencia conocida. Con este trasfondo, en esta tesis se planteó caracterizar el elemento genético móvil *Galileo* en diferentes especies de *Drosophila* y analizar su dinámica evolutiva. De esta manera, en una primera fase se buscaron elementos *Galileo* completos en en diferentes especies del género *Drosophila*: *D. buzzatii*, *D. mojavensis*, *D. virilis*, *D. willitoni*, *D. ananassae*, *D. pseudoobscura* y *D. persimilis*, utilizando métodos tanto bioinformáticos como experimentales (dependiendo de si el genoma analizado estaba secuenciado o no). Las copias encontradas presentan largas Repeticiones Invertidas Terminals (TIR) de hasta 1,2 Kb, una elevada identidad con secuencias de *Galileo* descritas con anterioridad y, además, contienen una zona codificante que ha permitido clasificar *Galileo* como miembro de la superfamilia del *elemento P*. Posteriormente, mediante análisis filogenéticos, hemos encontrado la existencia de subfamilias de *Galileo* en tres especies (*D. buzzatii*, *D. mojavensis*, *D. virilis*) y evidencias de actividad transposicional reciente (*D. willitoni*, *D. ananassae*, *D. pseudoobscura, D. persimilis y D. mojavensis*). En una segunda fase de la tesis, hemos llevado a cabo experimentos con parte de la proteína que codifica *Galileo* y hemos comprobado que interacciona con las TIR de *Galileo*, confirmando que esta secuencia es la responsable de la reacción de transposición. Finalmente, hemos analizado en detalle la diversidad de *Galileo* en el genoma de *D. mojavensis* y hemos detectado una diversidad estructural muy importante, lo que sugiere que el intercambio de secuencias entre elementos podría ser bastante frecuente para la evolución de los TEs.

# I.- INTRODUCTION

# 1.- Transposable Elements

Transposable elements (TEs) are genetic entities with the capability of changing their location within the genome. They were discovered by Barbara McClintock in the 50s of the last century when she was exploring the origin and behaviour of mutable loci in maize (McClintock 1950, 1951). McClintock's discovery challenged the concept of the genome as a static set of instructions passed between generations, as genetic maps had shown. Thus, her theories about how changes in gene expression could appear in two successive generations were received with huge scepticism. Finally, since her observations and theories were corroborated in other organisms, she was awarded in 1983 with the Nobel Price of Physiology and Medicine for her discovery of transposition.

Usually, movement of TEs results in their multiplication, that can give rise to high copy numbers. TEs have been included in the fraction of middle repetitive DNA of the genome, as interspersed repeats (Britten & Kohne 1968). So far, TEs have been found in almost all studied species, prokaryotes and eukaryotes, except in the protozoan *Plasmodium falciparum* (Gardner et al. 2002). In all species, TEs make up a significant but variable proportion of the genome, e.g.: 12 % in *Caenorhabditis elegans* (The C. elegans Sequencing Consortium 1998), 14 % in *Arabidpsis thaliana* (Hua-Van et al. 2005), 16 % in *Drosophila melanogaster* (Kidwell 2002; *Drosophila* 12 Genomes Consortium et al. 2007), 45 % in humans (Lander et al. 2001) and 80% in some crops (Wicker et al. 2007).

TE activity in the genomes causes a broad range of mutations. Since their movement is often random, a priori, they can insert anywhere in the genome. By chance, they can insert in regions where they will not affect any function (heterochromatin, intergenic regions, etc), but likewise, they can interfere in the cell working machinery. For example, a gene can be inactivated because a TE insertion breaks the ORF or affects the splicing, or the TE impairs the expression of the gene. In addition, TE activity generates deletions, duplications and rearrangements in the genome. In summary, TEs generate a huge range of mutations with a broad impact on host fitness (Kidwell & Lisch 2002; Feschotte & Pritham 2007).

The expansive nature of TEs, occupying important fractions of genomes, along with their mutational activity due to its random movement, made them to be considered as selfish and/or junk DNA because no positive role for the cell was apparent (Doolittle & Sapienza 1980; Orgel & Crick 1980). Likewise, the broad distribution among species suggests they have a very successful parasitic strategy, although this broad distribution could be also be pointing out a putative role for the cell, as it has been seen in some cases; e.g. the telomere-length manteinance in *Drosophila* genus, which is carried out by the retrotransposons *HeT-A*, *TART* and *TAHRE* (Casacuberta & Pardue 2005; Pardue et al. 2005; Pardue & DeBaryshe 2011). Nevertheless, although most of the time the TE activity has deleterious effects, it also generates variability and even advantageous mutations, which indicates that they are facilitators of evolution (Kazazian 2004; Cordaux et al. 2006; Oliver & Greene 2009, 2011).

## 1.1.- The evolutionary life-cycle of transposable elements

TEs are dynamic entities which multiply, move, evolve and interact with the host. Their ability to invade genomes along with the fact that they do not play any cellular function in the host makes them to be considered parasitic sequences (Doolittle & Sapienza 1980; Orgel & Crick 1980). Thus, the evolutionary life-cycle of TEs has been suggested to be analogous to that of parasitic organisms, with a first phase characterised by the invasion and establishment of the host genome followed by a decrease of TE activity and a phase of coexistence of different mutant sequences until the disappearance of the mobile element (Figure 1) (Silva et al. 2004; Le Rouzic et al. 2007). During all this cycle, there are evidences of TE parasitism, such as their use of the cell machinery for spreading themselves and the host fitness decrease due to TE insertion mutations and chromosomal instability (Doolittle & Sapienza 1980; Orgel & Crick 1980).

The complex evolutionary dynamics of TEs has required the development of a theoretical framework based on population genetics models which provide a series of predictions that can be tested later on empirical grounds. In the 80s of the last century, several models were proposed to account for parasitic nature of TEs, such as the models of Brookfield (1982) and Hickey (1982). Afterwards, Charlesworth and Charlesworth (1983) modelled the dynamics of copy number taking into account the transposition rate

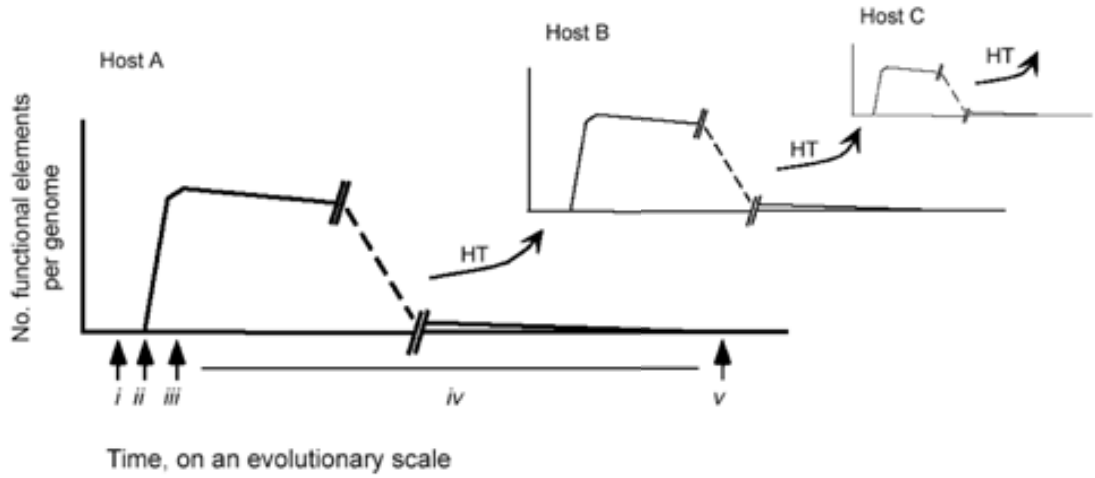



**Figure 1**. Schematic view of TE dynamics after entering the genome. HT means horizontal transfer of the TE to another host. The different steps are: (i) An element is transferred into a germline cell of host A. (ii) Transposition activity starts after a successful integration of the TE. There is a rapid increase in copy number. (iii) Repression of transposition arises the rate in copy number slows. (iv) Mutations accumulate in the different copies and the number of functional elements in the genome slowly decreases. This process that can take many millions of years (abbreviated period represented by a dashed line). (v) Finally, no functional elements are left in the genome of host A, and this TE lineage becomes extinct. Sometime between (ii) and (v), a functional element may be transferred horizontally (HT) to a new host and the process begins anew . Taken from Silva et al. (2004).

and the selective pressure against the TEs insertions. The exploration of the simulations stated that copy number should reach an equilibrium between these two forces, transposition and selection. This way, although element frequencies could change as a result of different phenomena (such as, replicative transposition, loss of elements from occupied sites, selection on copy number per individual, and genetic drift) the final balance would depend on a strong transposition control or a high selective pressure, or both (Charlesworth & Charlesworth 1983).

Purifying selection is a force opposing the spread of TEs, and it would act against (i) TE insertions which disrupt functional genetic units and (ii) TEs which generate deleterious products. Regarding these statements, it would be expected that the X chromosome, where selection is stronger than in autosomes due to the hemizygous state in males, would present a reduced number of TEs than the autosomes. This hypothesis was tested with three TE families of *D. melanogaster* and there was no evidence for any reduction in copy number for the X chromosome, leading to the suggestion that meiotic recombination between transposable elements at non-homologous sites would be responsible for the containment of TEs number in natural populations (Montgomery et al. 1987). Thus, a new model was proposed taking into account the distribution of TEs across genomic regions with different rates of unequal exchange or ectopic

recombination (Langley et al. 1988; Charlesworth & Langley 1991). This would mean that a TE insertion would disappear more quickly if it is located in a high recombination rate region because it would be more prone to recombine with a non-allelic homologous TE. This recombination event leads to the production of deleterious chromosomal rearrangements, thus lowering the fitness of individuals as a function of the number of elements carried. This model has been confirmed by some empirical data and seems to fit quite well with the actual distribution of TEs in natural populations (Charlesworth et al. 1992; Bartolomé et al. 2002; Petrov et al. 2003, 2010).

These models provide predictions for populations in which TEs have reached an asymptotic equilibrium state, but before this equilibrium is reached there are other steps in a TE cycle which are sensitive for the success or survival of mobile elements, such as the colonization of a new genome. Furthermore, the equilibrium could be affected by demographic events of the host or reactivation of a TE (such as stress responses or secondary contacts between geographically distant populations). Hence, not all genomes might be at equilibrium, rather they could be in an unstable TE-host state. Recently, new mathematical models have been proposed for predicting/modelling the whole cycle of TE. Le Rouzic and Capy have run simulations to predict the behaviour of TEs in different steps of the cycle: the invasion, the competition among subfamilies and the long-term evolution (Le Rouzic & Capy 2005, 2006, 2009; Le Rouzic et al. 2007). Their simulations predict that for a successful genome invasion, after a horizontal transfer event or a TE reactivation, a high transposition activity is needed followed by a tight control of it, which means a transposition burst. This way, the TE that arrived itself to a new genome would overcome the genetic drift and its extinction. After the establishment, TE activity starts to generate mutant copies, either transposition machinery-coding mutants or transposition efficiency

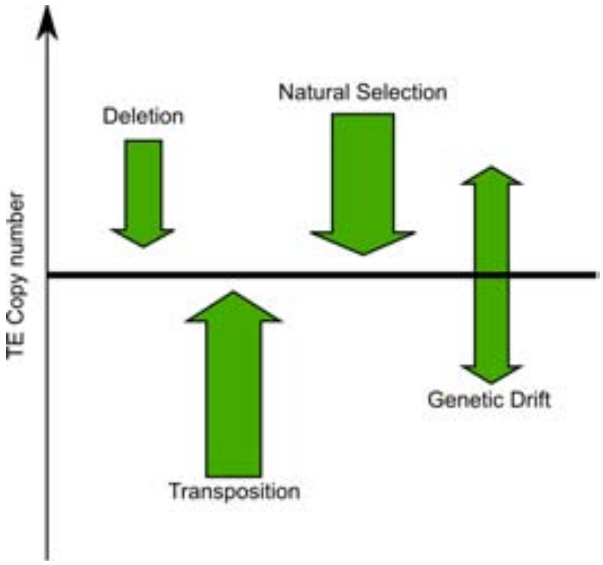



**Figure 2.** Simple representation of the different genomic forces which interact and affect TEs dynamics. The size of the arrows depicts an schematic contribution of each phenomenon to the TE copy number. Modified from Le Rouzic and Capy (2009)

mutants. Competition among these copies seems to prevent the system for achieving a stable transposition-selection equilibrium, rendering non-autonomous copies to multiply and spread at the expense of the autonomous elements (Leonardo & Nuzhdin 2002). This results in a mainly cyclic dynamics which highlight the similarities between genomic selfish DNA and host-parasite systems (Le Rouzic & Capy 2006). Furthermore, long-term evolution was explored introducing variability in both the effects of the insertion on host fitness and the production of functional transposition proteins, along with mutations in transposition efficiency of the copies (Le Rouzic et al. 2007). The most common dynamics was found to be the occurrence of one or more invasion-regression cycles (transposition bursts) followed by the definitive TE loss. This questions the likelihood of the sustainable long-term stable transposition-selection equilibrium of older models. Furthermore, TE domestication events could appear, allowing the survival and fixation of those TE copies that enhance the fitness of the host.

When genomes are explored, the proportion of active copies is highly heterogeneous among species. For example, active copies account for: less than 20% in *D. melanogaster* (Bartolomé et al. 2002), less than 5% in *Schizosaccaromyces pombe (Bowen et al. 2003)*, and only 1% of LINEs in the human genome (Ostertag & Kazazian 2001). Le Rouzic et al. (2007) propose two hypotheses for this heterogeneity. On the one hand, different TE families and subfamilies are in different phases of their cycle, for example, some of them are actively colonising the genome whereas others are in the final step where there is no more mobilisation and the copies are accumulating mutations. On the other hand, long-term evolution of a TE family is affected by characteristics of the TE, the host and specific TE-host interactions, because slight changes in the parameters of the model (transposition rate, deletion rate, impact in host fitness, transposition activity and TE mutation) lead to distinct dynamics. Moreover, the two hypothesis are not mutually exclusive and its combination is likely to shape the TE ditribution observed in genomes (Le Rouzic et al. 2007).

In summary, although different models have been proposed, the TE dynamics are complex to infer, but it seems clear that the genetic drift and the purifying selection play a major role in TE control (Figure 2).

## 1.2.- Classification of transposable elements

The increasing amount of TEs being discovered makes necessary to develop a method of classifying and arranging all their information. Furthermore, the classification along with all the knowledge of TEs is a fundamental tool for the proper sequencing, assembly and annotation of the numerous genome projects that are being carried out (Edgar & Myers 2005; Han & Wessler 2010). One of the first methodical attempts to classify eukaryotic TEs was carried out by Finnegan (1989), who defined two main classes of TEs: Class I are TEs with a retrotranscription step, where a RNA state of the element is found and is retrotranscribed to DNA, while Class II are devoid of this step and are always found as DNA molecules (Finnegan 1989). More recently, Wicker et al. (2007) elaborated on this basic scheme and proposed different levels of classification, such as; subclass, order, superfamily and family. Subclass is used, within Class II, to distinguish elements that copy themselves for insertion, from those that leave the donor site to reintegrate elsewhere. It concomitantly reflects the number of DNA strands that are cut at the TE donor site. At the next level, order takes into account the element structure, for example, the existence of TIRs or LTR in the different classes. These structural traits reflects major differences in the insertion mechanism and, consequently, the overall organization and enzymology. The final levels are superfamily, family and subfamily, where phylogenetic relationship along with nucleotide identity are taken into account in each level of classification (Figure 3).

Class I

Class I of TEs, also known as retroelements, are characterised by a transposition reaction where an intermediate molecule of RNA is transcribed from the donor site and, afterwards, this RNA molecule will be retrotranscribed to DNA and inserted elsewhere in the genome. Thus, the main trait of this group is the retrotranscription step. It is noteworthy that this step is replicative (hence the “copy-and-paste” term often used to refer to this group). Consequently, retrotransposons may reach high copy numbers and are often the major contributors to the repetitive fraction in large genomes. Following the more detailed classification of Wicker et al. (2007) this class is subdivided in five orders on the basis of their mechanistic features, organization and reverse transcriptase phylogeny: LTR retrotransposons (Long Terminal Repeats), DIRS-like elements

(*Dictyostelium* intermediate repeat sequence; Cappello 1985), Penelope-like elements (PLEs), LINEs and SINEs. Prior to this classification, Class I elements were usually subdivided in LTR versus non-LTR elements (Kumar & Bennetzen 1999; Jurka et al. 2007).

LTR elements range in size from a few hundred base pairs up to, exceptionally, 25 kb (Wicker et al. 2007). The length of LTR range from a few hundred base pairs to

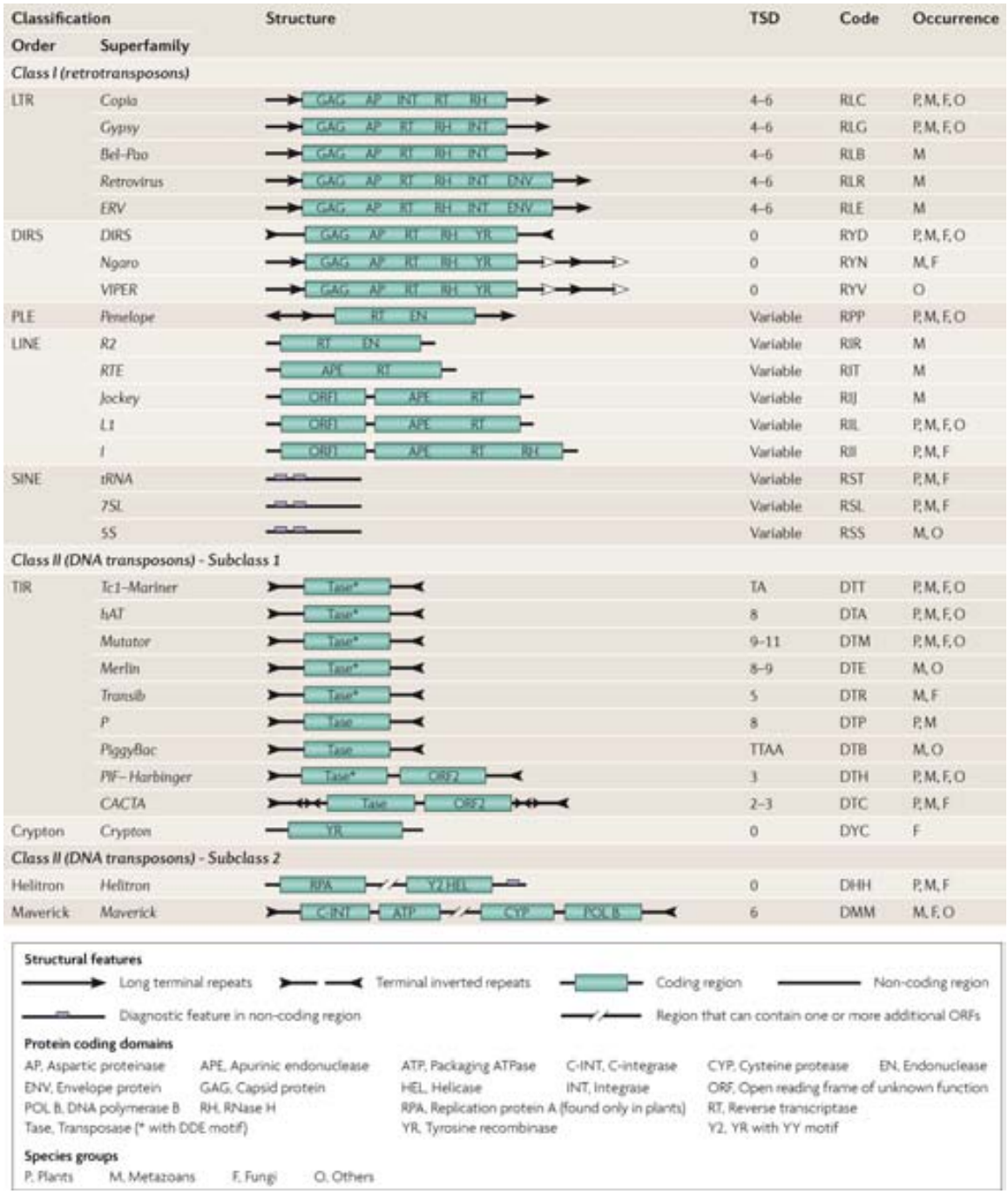



**Figure 3.** Classification of transposable elements proposed by Wicker et al (2007).

more than 5 kb, and start with 5'-TG-3' and end with 5'-CA-3'. Upon integration, LTR retrotransposons generate a target site duplication (TSD) of 4-6 bp. They typically contain ORFs for GAG, a structural protein for virus-like particles and for POL. *Pol* generally encodes an aspartic proteinase (AP), reverse transcriptase, RNaseH and DDE integrase (INT). Occasionally, there is an additional ORF of unknown function (Wicker et al. 2007).

DIRS-like elements contain a tyrosine recombinase gene instead of an integrase and, therefore, they do not generate TSD upon insertion. Their termini are unusual, resembling either split direct repeats (SDR) or inverted repeats. These features indicate a mechanism of integration that is different from that of LTR elements and LINEs. Nevertheless, their RT places them in Class I. Members of this order have been detected in diverse species, ranging from green algae to animals and fungi. Penelope-like elements (PLEs) encode a RT that is more closely related to telomerase than to the RT of LTR retrotransposons or LINEs. Furthermore, they code for an endonuclease that is related both to intron-encoded endonucleases and to the bacterial DNA repair protein UvrC. These elements also have LTR-like sequences that can be in direct or an inverse orientation (Wicker et al. 2007).

LINEs lack LTR, can reach several kilobases in length and encode at least a RT and a nuclease in their *pol* ORF for transposition. Sometimes there is also a *gag*-like ORF, and other containing RNaseH. LINEs generate TSDs of 7-20 bp length upon insertion, and usually they present truncated 5' ends as result from premature termination of their primed reverse transcription (Ostertag & Kazazian 2001). At their 3' end, they can display either a poly(A) tail, a tandem repeat or merely an A-rich region (Wicker et al. 2007). SINEs are non-autonomous elements but they are not deletion derivatives of autonomous ones; instead, they originate form accidental retrotransposition of various polymerase III (pol III) transcripts. Unlike retroprocessed pseudogenes, they possess internal Pol III promoters which allow them to be expressed. They rely on LINEs for trans-acting transposition functions such as RT. Some SINEs present a unique and obligatory partner whereas others are generalists. SINEs are small (80-500 bp) and generate TSDs (5-15bp). The Pol III promoter region defines SINE superfamilies and reveals their origin: tRNA, 7SL RNA and 5S RNA. SINE internal regions (50-200 bp) are family-specific and of variable origin, sometimes deriving from SINE dimerization

or trimerization (Kramerov & Vassetzky 2005). The best known SINE is the *Alu* element, which presents at least $>10^6$ copies in the human genome (Lander et al. 2001).

Class II

Class II elements are devoid of the retrotranscription step. In this class there are different strategies of transposition and some of them imply a direct replicative step. Two different subclasses have been proposed, one with the cut-and-paste elements and another one that entails replication without a double-stranded cleavage (Wicker et al. 2007). The first subclass is comprised by two orders: TIR containing elements and no-TIR elements (Crypton). TIR elements are subdivided in superfamilies but different proposed classifications do not agree in the number of them. For example, Feschotte and Pritham (2007) proposed 10 superfamilies of eukaryotic TIR transposons. However, Jurka et al. (2007) and Wicker et al. (2007) recognized 13 and 9 superfamilies, respectively. Recently, Yuan & Wessler (2011) have proposed to revise the number of cut-and-paste transposons because their phylogenetic analysis of the catalytic domain uncovered new relationships among the different groups. They propose 17 superfamilies clustered in three supergroups. Although the definition and number of superfamilies has not reached a consensus, these clusterings are very useful for uncovering the TEs in the different genome projects, because generally, they are searched by means of similarity tools for locating and annotating different TEs. The second subclass is split in two orders, *Helitrons* and *Mavericks*/*Polintons*. *Helitrons* replicate using a rolling-circle strategy, whereas transposition reaction for *Mavericks* is still unknown (Feschotte & Pritham 2007; Wicker et al. 2007).

The TEs studied in this thesis belong to the cut-and-paste class II transposons. In the sections below these elements are explained in detail.

## 2.- The *Drosophila P-element*

*The Drosophila P-element* is one of the best-studied eukaryotic mobile DNA elements. It was discovered in the late 1960s because it causes in *Drosophila melanogaster* a syndrome of genetic traits termed *hybrid dysgenesis* (HD) (Kidwell et al. 1977). HD is a term used to describe a collection of symptoms including high rates of sterility, mutation induction, male recombination and chromosomal abnormalities and rearrangements (Kidwell 1977; Kidwell & Novy 1979; Kidwell et al. 1977; Engels 1979). The unstable nature and reversibility of the mutations caused by hybrid dysgenic crosses first suggested that they might be caused by mobile element insertions (Kidwell et al. 1973). A detailed molecular analysis of hybrid dysgenesis-induced mutations at the *white* locus allowed the isolation and molecular cloning of the *P* transposable element (Bingham et al. 1982; Rubin et al. 1982). The characterization of *P-elements* rapidly let to the development of its use as a vector for efficient germ line transfer in *Drosophila* (Rubin & Spradling 1982; Spradling & Rubin 1982). Since then, *P-element* vectors have been widely used for transforming *D. melanogaster* (Figure 4).

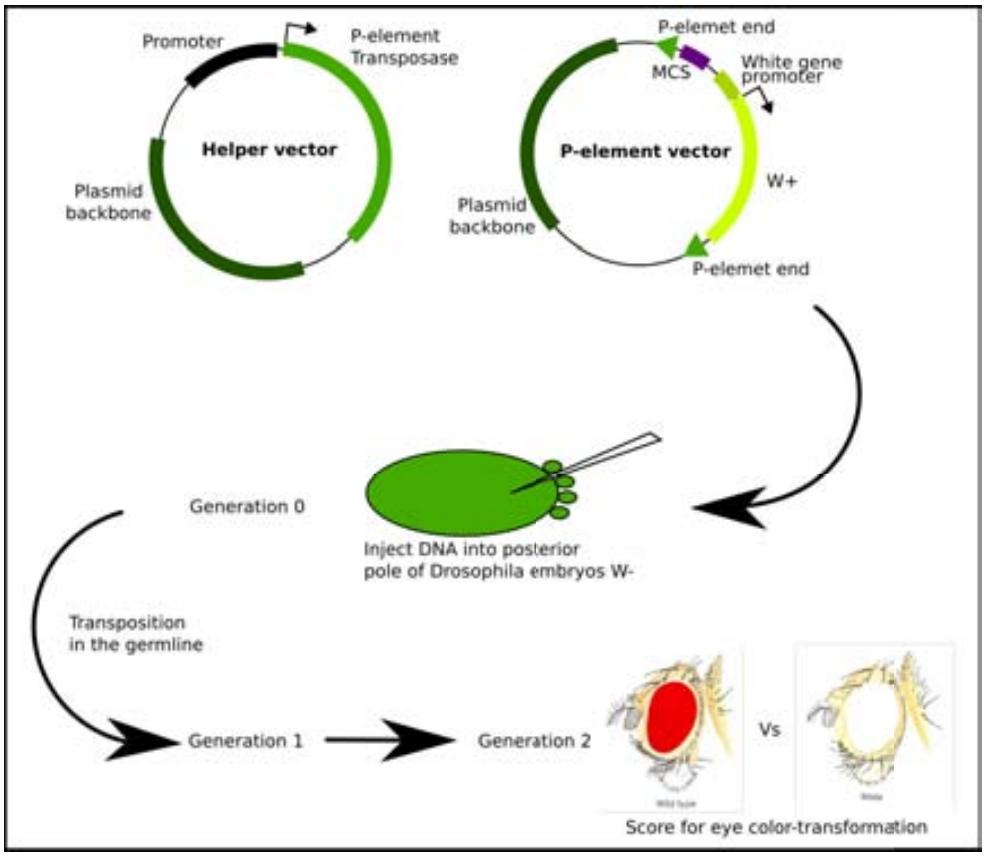



**Figure 4**. General procedure for *Drosophila* transformation using *P-element*-based vectors. General traits of vectors are shown on top. The procedure for *Drosophila* transformation is sketched as well. Adapted from Rio (2002).

Furthermore, these elements have found additional and critically important uses as the molecular genetics of *Drosophila* has evolved, such as, mutagenesis and gene-tagging, enhancer trapping, homologous gene targeting and gene replacement (Engels 1996; Rong & Golic 2000; Rubin et al. 2000). Nowadays, new vectors for transforming *Drosophila* are being developed and they are *P-element* based vectors, so germinal transformation is still the best choice for *Drosophila* transformation (Kondo et al. 2006; Bachmann & Knust 2008).

## 2.1.- *P-element* structure

The *P-element* is a cut-and-paste transposon from Class II of mobile elements (subclass I, TIR order, Wicker et al. 2007). The autonomous and complete copy is ~2.9 kb long and its structure consists of two 31-bp terminal inverted repeats (TIRs) surrounding an ORF encoding the transposase (Figure 5). This ORF comprises four exons and three introns and encodes the enzyme responsible for the transposition of the element. This protein is able to bind close to the ends of the transposon, join and cut them and insert the element in a new location (see below). Moreover, the alternative splicing of the transposase ORF generates a transposition inhibitor (KP protein), that directly binds to the transposase DNA binding sites and blocks the *P-element* DNA cleavage (Misra et al. 1993; Lee et al. 1998). Other important regions in the *P-element* are the binding sites, where the transposase binds (BS). The binding sites are not located inside the TIRs and are not equidistant from the transposon ends, one is 21 bp from the 5' TIR and the other is 9 bp from the 3'TIR (Rio 2002). These sequences are 10-bp long and correspond to GTTAAGTGGAT (3' end) and TTTAAGTGTAT (5' end) (Sabogal et al. 2010). Finally, there are two internal inverted repeats of 11 bp (ATTAACCCTTA) located 126 bp from the 5' end and 201 bp from the 3' end. Although not absolutely required for the transposition reaction, they act as transpositional enhancers (Rio 2002).

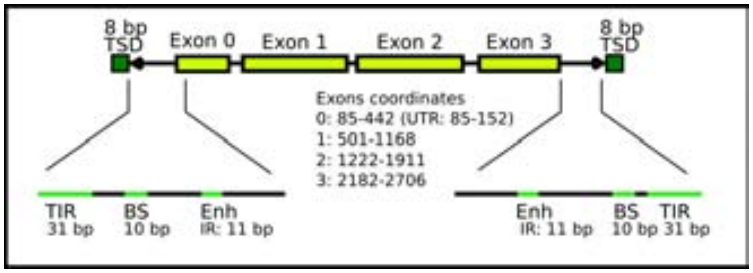



**Figure 5**. *D. melanogaster P-element* canonical sequence structure. Total length 2.9 kb. The binding sites (BS) of the transposase and the internal inverted repeats that act as transpositional enhancers (Enh) are shown. The transposase CDS is depicted with its structure of 4 exons and 3 introns. Adapted from Rio (2002).

## 2.2.- *P-element* transposase

The *P-element* transposase is a trans-acting protein of 87 kDa, 751 amino-acids, that catalyses the *P-element* mobilization through a cut-and-paste reaction. This protein has a modular structure with different domains that are responsible for different steps of the transposition reaction (Figure 6).


**Figure 6**. Structure of *D. melanogaster P-element* transposase. The different domains and their coordinates are depicted. Adapted from Rio (2002).

THAP domain

The DNA binding domain (DBD) of the transposase is located in the N-terminus and has been described as a special kind of zinc finger, the THAP domain (Roussigne et al. 2003; Clouaire et al. 2005). This domain is shared with other cellular proteins found in different animals, from *Drosophila* to humans, that are implicated in different pathways, such as, cellular cycle, apoptosis and chromatin-associated proteins among others (Figure 7) (Roussigne et al. 2003). This domain begins with a C2CH (cystein-cystein-cystein-histidine) zinc coordinating region and ends with an AVP (alanine-proline-valine) motif. Compared with the most common zinc fingers (e.g. C2H2 or C4-type, Lee et al. 1989; Pavletich & Pabo 1991) the THAP domain can be considered as a long domain.

Among the conserved features of the THAP domain are its location at the N-terminus of the proteins, its size about 90 residues and, most importantly, the presence of conserved sequence motifs. The defined THAP domain includes: a C2CH signature (consensus cystein-$Xaa_{2-4}$-cystein-$Xaa_{35-50}$-cystein-$Xaa_2$-histidine); three additional key residues that are strictly conserved in all THAP domains (proline (P), tryptophan (W), phenylalanine (F), see Figure 8); a C-terminal AVPTIF box (consensus: alanine(A)-valine(V)-proline(P)-threonine(T)-isoleucine(I)-phenylalanine(F)); and several other conserved amino acid positions with distinct physico-chemical properties (e.g. hydrophobic and polar) (Roussigne et al. 2003).

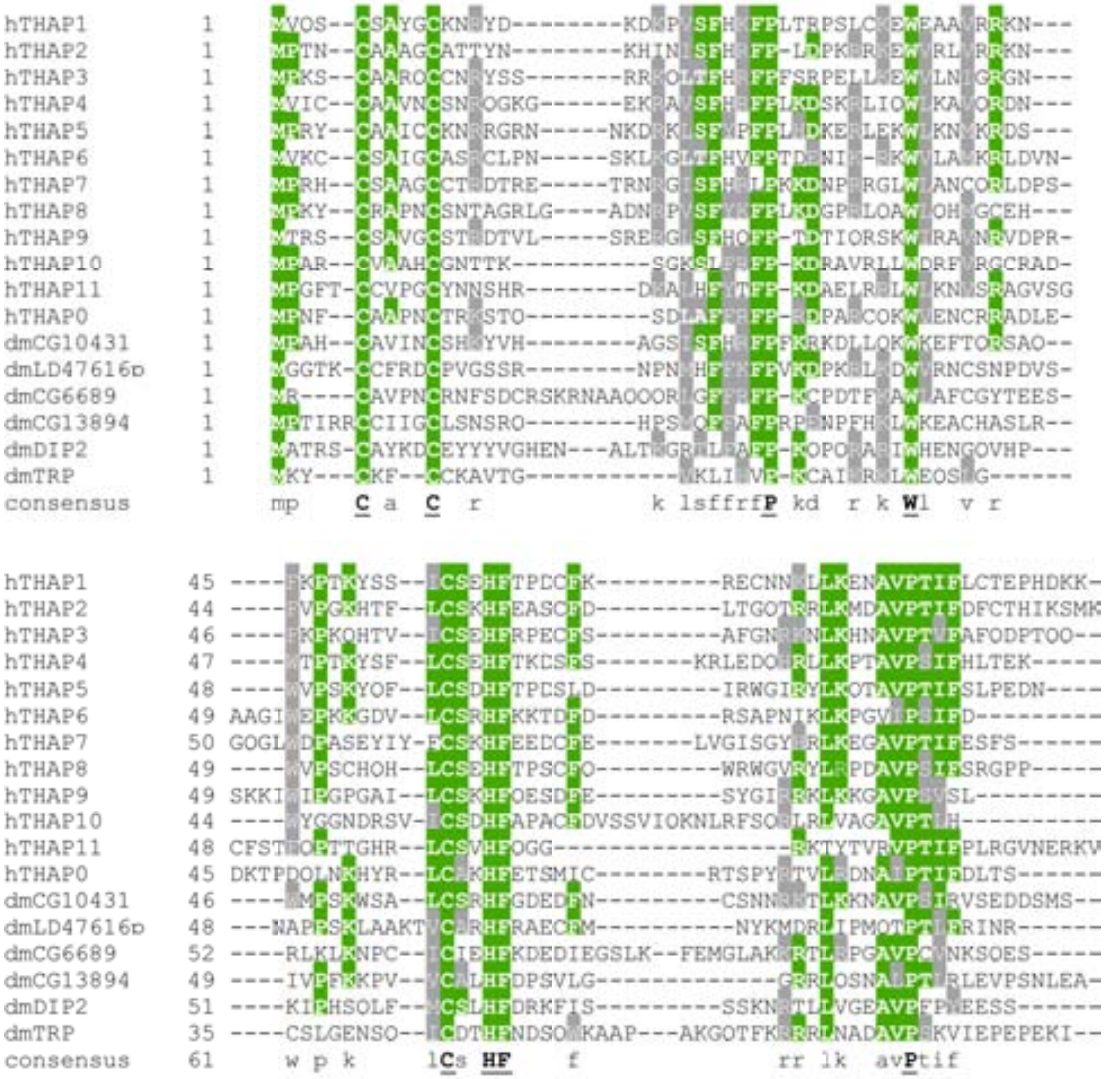


**Figure 7**. Alignment of different THAP domains from different proteins. dmTRP is *P-element* transposase THAP. The conserved key residues are underlined. Taken from Roussigne et al. (2003).

Recently, the three dimensional structures of two different THAP domains bound to DNA have been characterised: the human protein THAP1 and the *P-element* transposase (Figure 8) (Bessière et al. 2008; Sabogal et al. 2010). Despite the conservation of the key residues of the domain, the overall sequence conservation is very low. Nevertheless, the spatial conformation seems to be highly conserved and a new DNA interaction manner has been proposed: a β-sheet interacts with the target DNA through the major groove and a downstream loop in the domain interacts with the minor groove of the double helix. Since the DNA interaction is conserved, it has been proposed that the THAP DNA consensus binding sequence is TXXGGGX(A/T) or TXXXGGCA (the X are spacing sequence of variable length; (Clouaire et al. 2005; Campagne et al. 2010; Sabogal et al. 2010). It can be noticed that this two proposed

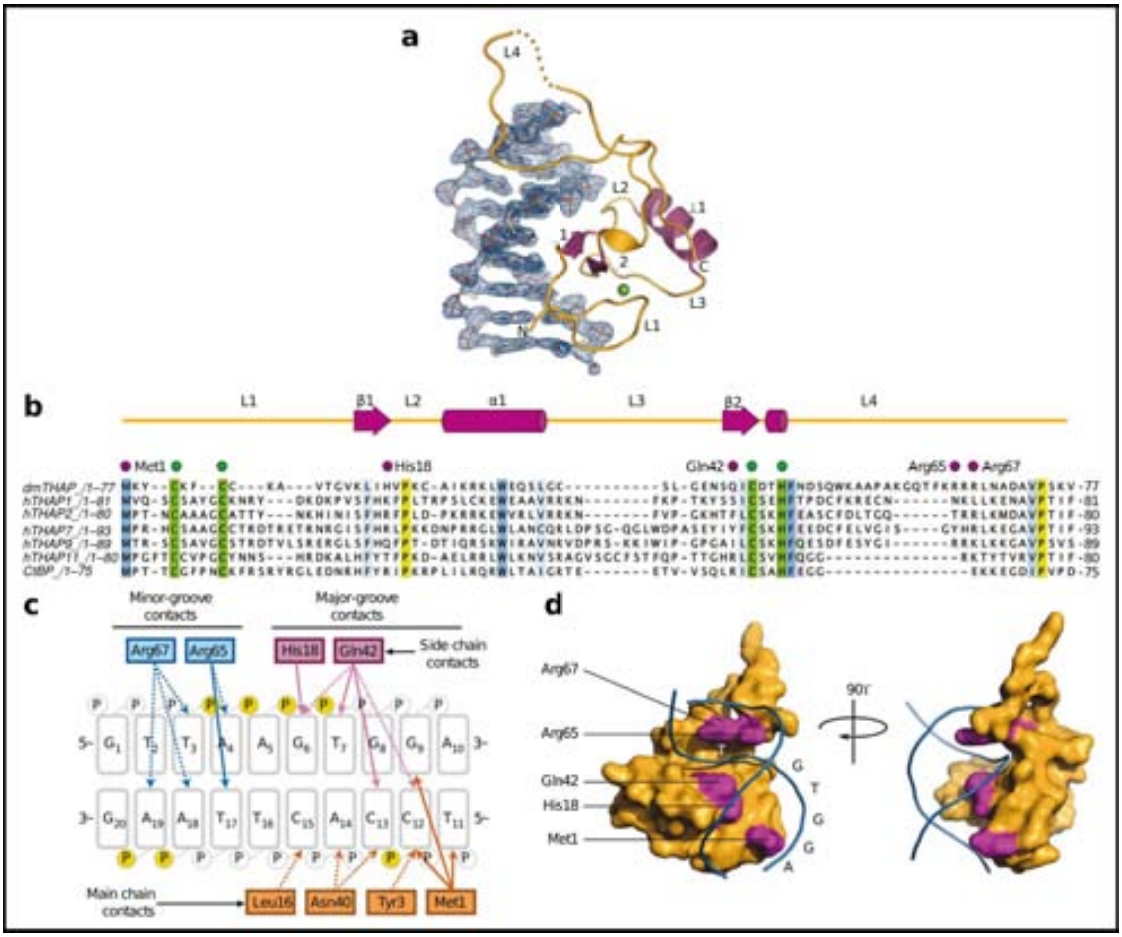



**Figure 8**. THAP domain 3D structure interacting with DNA. a) Protein-DNA interface b) Structure-based multiple sequence alignment of DmTHAP, human THAP1, THAP2, THAP7, THAP9 and THAP11 and *C. elegans* CtBP where conserved residues are highlighted; zinc-coordinating C2CH motif is highlighted in green; base-specific DNA-binding residues of DmTHAP are indicated by magenta. The secondary structure diagram is shown above the alignment. c) Schematic representation of all base-specific contacts in the major and minor grooves . d) Surface representation of DmTHAP. Sequence-specific DNA-binding residues are highlighted in magenta. DNA backbone is shown as lines with subsite positions labelled. Modified from Sabogal et al. (2010).

consensus binding sequences share similarities in sequence, such as the core of 3 GC base pairs (GGG or GGCA) which is the major groove interacting sequence, and a conserved AT base pair, which is the minor groove interacting sequence (Sabogal et al. 2010). Furthermore, the size of the two proposed consensus binding sequences are similar (~10 bp), although they correspond to a *Drosophila* and a human THAP1 protein, respectively.

Oligomerization region

After the DNA binding domain, there is an oligomerization region. It consists of a leucine zipper (Landschulz et al. 1988) responsible for the multimerization of the transposase. After this leucine zipper, there is a second oligomerization region

consisting of an unstructured region, possibly a coiled-coil region (Rio 2002; Sabogal et al. 2010). The multimerization is not necessary for the high-affinity site-specific DNA interaction, but it is essential for the transposition reaction (Rio 2002).

Putative regulatory domain

In the amino-terminal region, there is a regulatory domain that contains potential sites for phosphorylation by different kinases, such as the DNA repair-checkpoint phosphatidyl inositol-3-phosphate($PI_3$)-related protein kinases DNA-PK and ATM (Ataxia telangiectasia mutated, Ku p70 and Ku p80 in *Drosophila*). Alterations of these potential phosphorylations sites by mutagenesis to alanine result in both increased and decreased transposase activity *in vivo* and *in vitro*. In this sense, when the transposase is produced in bacteria, the enzyme is not active, due to the lack of phosphorylation. Similarly, transposases treated with phosphatases presented reduced activity (Rio 2002).

GTP-binding domain

The *P-element* transposase has a unique requirement for guanosine triphosphate (GTP) binding that distinguishes it from smaller transposases (e.g. those of Tn5 and Mu). However, GTP is known to take part as a cofactor in many diverse biochemical processes, such as Ras cellular signal transduction pathways, the assembly of dynamin in vesicle transport, and the self-splicing of group I introns, among other cellular functions (Bourne et al. 1991; Doudna & Cech 2002; Praefcke & McMahon 2004; Tang et al. 2005). Thus, it has been of interest to understand the role of GTP in a transposase, which has a very different function compared to the cellular proteins which need this nucleotide. The GTP molecule is considered to be an allosteric effector required for proper folding and domain positioning of the *P-element* transposase, because different experiments have shown that the GTP is not hydrolysed during the transposition reaction (Kaufman & Rio 1992). Without GTP, the transposase is not able to form the synaptic complex which is vital for the transposition reaction. The synaptic complex is the conformation when the transposase is bound to the two ends of the transposon (Rio 2002; Tang et al. 2005).

The GTP domain of the *P-element* transposase is a non-canonical version compared to the motifs found in the GPTase superfamily (Bourne et al. 1991; Rio 2002). Consequently, the boundaries of the domain could not been determined through

sequence comparison. However, the GTP binding domain of the *P-element* transposase has been recently characterised thanks to a green fluorescent protein (GFP) solubility screening in *E. coli* (Sabogal & Rio 2010). This assay has allowed to locate the whole region responsible for the GTP binding in coordinates from 275 to 409 of the transposase. The GTP domain is able to bind GTP itself, without need of the other protein domains or multimerization, thus it is a single and functional domain. Furthermore, no GTPase activity has been detected, which is in agreement with the observation that the GTP has a role of allosteric co-factor (Sabogal & Rio 2010).

Catalytic domain

The C-terminus of the *P-element* transposase protein contains many acidic residues which would make up the catalytic domain of the transposase. Mechanistically, this domain is thought to belong to the RNaseH-like superfamily of polynucleotidil transferases. This superfamily includes different transposases and integrases such as: the bacterial Tn5 transposase, the Mos1 transposase, the HIV integrase, the phage Mu transposase, the Holliday junction nuclease Ruv C and the RAG1 V(D)J recombinase, among other proteins (Capy et al. 1996; Nowotny 2009; Hickman et al. 2010). Although mechanistically the *P-element* transposase is related to this superfamily of proteins, sequence and structure-based alignments reveal little or no sequence similarity. Thus, it seems that the *P-element* transposase would have evolved from a different type of polynucleotidil transferase, that could be related to the nucleic acid polymerases or restriction-endonucleases (Rio 2002).

However, a recent sequence analysis of different transposases where no DDE motif was found, has uncovered the putative DDE motif in the *P-element* superfamily (Yuan & Wessler 2011). The proposed residues for the catalytic domain of the *P-element* would be located in D230, D303 and E531 (Figure 9). These residues appear conserved in the different transposases of the *P-element* superfamily along with surrounding residues. However, the residues proposed by Yuan and Wessler (2011) are in disagreement with those proposed previously by Rio (2002) (D444, D528, E531and D545/628) which were seemingly detected through random mutagenesis of the catalytic domain (Rio 2002). Experiments that could corroborate the residues proposed by Yuan

and Wessler (2011) would be very interesting for finally including all the eukaryotic transposases in the RNaseH superfamily of polynucleotidil transferases.

Regardless whether this catalytic domain harbours the DDE signature or not, this kind of enzymes, where the *P-element* transposase can be mechanistically included, use metal ion-mediated catalysis to hydrolyse the phosphodiester bond. The metal ion is bivalent, usually $Mg^{++}$, and it is coordinated with the protein through acidic protein residues. This essential co-factor is needed for both DNA strand cleavage and strand transfer, which means the double-strand breaks and the insertion of the transposon steps (Hickman et al. 2010).

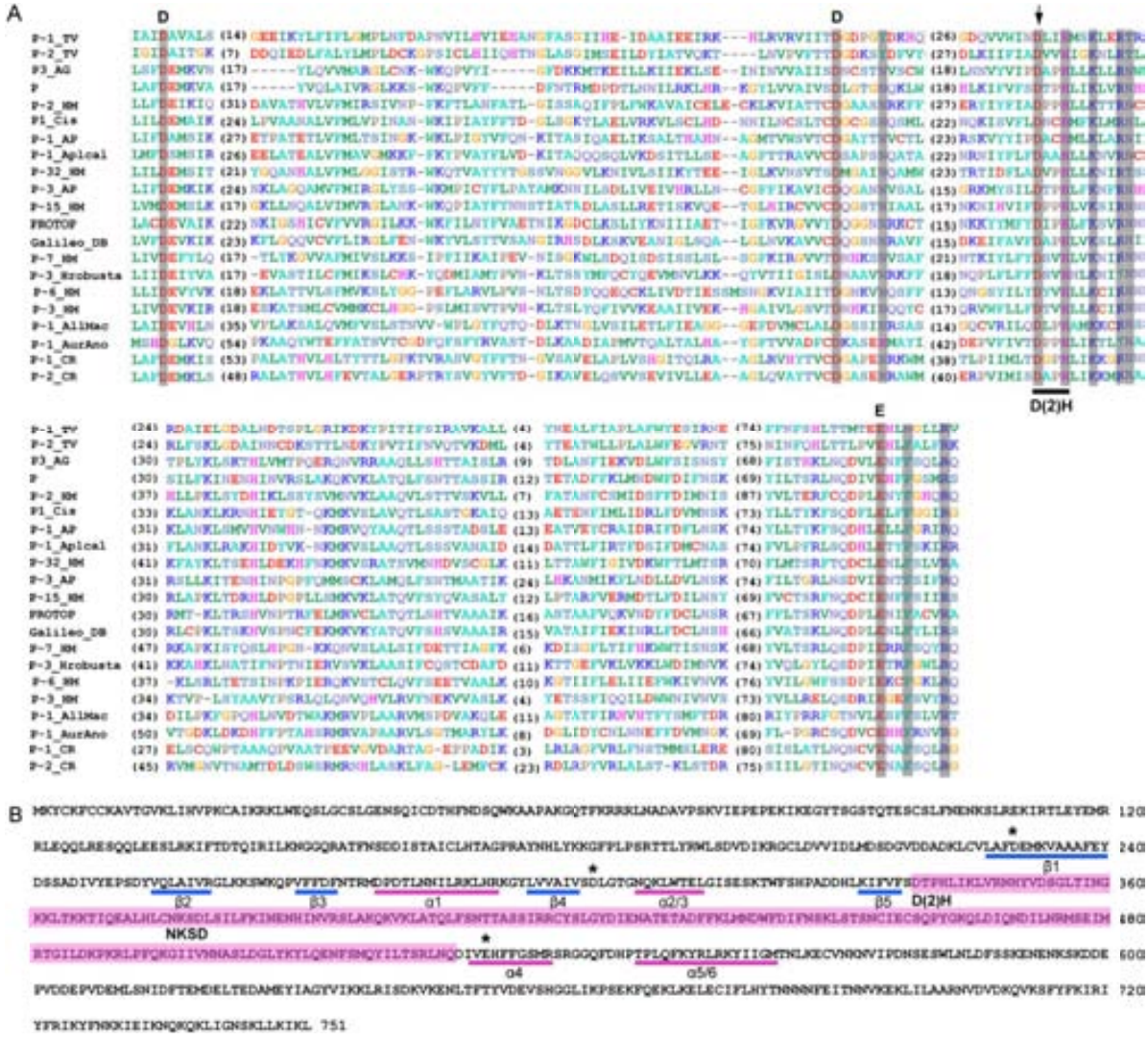



**Figure 9**. a) Alignment of the catalytic region of different transposases of the *P-element* superfamily. The conserved DDE residues are indicated. A part from the DDE residues, there is a region D(2)H which is conserved among all the transposases. b) Putative secondary structure of the *P-element* catalytic domain. The DDE residues are indicated with asterisks. Notice *D. buzzatii Galileo* element has been included. From Yuan and Wessler (2011).

## 2.3.- *P-element* transposition reaction

After transcription and translation of the *P-element* transposase ORF, the protein assembles itself as a tetramer (Tang et al. 2007). This pre-formed tetramer binds to one of the *P-element* ends and through a “looping" or intersegmental transfer (action helped by the GTP interaction) the tetramer binds the second binding site (synapsis) (Tang et al. 2007). After the binding, the transposase catalytic domains cut the transposon ends through a strand-transfer reaction. This is a staggered cut that leaves 17-bp overhangs at each 3' end. After that, the transpososome (transposon along with the transposition machinery) goes to a new location where there is a target insertion sequence. A staggered cut (8 bp length lag) is performed and the transposon inserts there. An eight base pair target site duplication (TSD) surrounds the element in its new location after the polymerase closes de remaining gaps (Rio 2002).

The gap left by the transposon jump, can generally have two different fates depending on the repairing pathway. On the one hand, the pathway may be non-homologous end joining repair (NHEJ), where the two 17 bp overhangs will be joined and a transposon footprint will appear surrounded by the 8-bp TSDs (Beall & Rio 1997; Dynan & Yoo 1998; Rio 2002). On the other hand, the repair may be done by the synthesis-dependent strand annealing pathway (SDSA), a gap repair process that uses the sister chromatid or the homologous chromosome as a template (Engels et al. 1990; Rio 2002). In that case, the whole *P-element* would be copied again in the location where it jumped from. This last step would be the responsible of the replicative transposition of the element and the rapid spread of *P-elements* in wild populations. If this repair synthesis is interrupted, this could give rise to the internally deleted *P-elements* observed naturally (Rio 2002).

## 2.4.- Insertional preference of the *P-element*

The initial DNA sequence analysis of several cloned *P-element* insertions revealed that 8-bp duplications of the target site (TSD) were found flanking all the *P-elements* analysed. Comparisons of these target site sequences revealed a general high GC base composition in the 8-bp sites, with the consensus sequence being 5'-GTCCGGAC-3' (O’Hare & Rubin 1983). Another study analysed 2266 *P-element* insertion sites from the Berkeley *Drosophila* Genome Project and showed that the 8-bp GC-rich TSD was

centred in a longer 14-bp palindromic target sequence (Liao et al. 2000). Recently, a more exhaustive bioinformatic analysis of the *P-element* insertion sites (over 10000 *P-element* insertions) has uncovered the putative consensus sequence for this 14-bp target palindrome (Linheiro & Bergman 2008). This sequence is 5'-ATR**GTCCGGAC**WAT-3' where the 8-bp palindromic target site duplication is shown in bold characters. All the positions of the motif presented strong statistical support deviating significantly from the overall *D. melanogaster* base composition. Strikingly, in this work from Linheiro and Bergman (2008), they found that the sequence of the *P-element* TIR restores the 14 bp palindrome after insertion. This suggests a mechanistic link between staggered-cut palindromic target sites and the structure of the transposon TIRs, specially involving the terminal nucleotides of the TIR. Moreover, this special role for terminal nucleotides in the *P-element* TIRs could explain the strong conservation of only the first 3 bp of the TIRs among the *P-element* family members in insects and vertebrates (see below). A *P-element* insertion becomes a new site for another *P-element* insertion. The fact that the sequence recognized by the transposase is a palindrome is consistent with the transposase acting as an homomultimeric complex with the target DNA (Linheiro & Bergman 2008).

## 2.5.- *D. melanogaster P-element* origin

To study the evolutionary origin and history of mobile elements a survey of phylogenetic distribution is very useful. These studies revealed *P-element* homologous sequences were distributed throughout the species groups that comprise the subgenus *Sophophora*, but were absent from the species most closely related to *D. melanogaster* (Brookfield et al. 1984; Anxolabehere et al. 1985; Lansman et al. 1985; Daniels & Strausbaugh 1986). This fact together with the *P-element* absence in old laboratory strains of *D. melanogaster,* suggested *P-element* might had entered in *D. melanogaster* through horizontal transfer from a distantly related member of the genus (Bingham et al. 1982; Anxolabéhère et al. 1988).

An exhaustive screening using Southern blot of 136 species of *Drosophila* genus uncovered a broad distribution of *P-element* in the *Sophophora* subgenus and a lack in the *Drosophila* subgenus. Furthermore, the strongest signals were found in the *willistoni* and *saltans* species group (Daniels et al. 1990). The candidate source species

for the putative horizontal transfer of the *P-element* were narrowed taking into account the species in sympatry with *D. melanogaster*. Finally, a whole *P-element* from *D. willistoni* was isolated and presented only one base-pair missmatch with *D. melanogaster P-element* canonical sequence (Daniels et al. 1990). Given the time lapse between the first collection of the stock flies and the new captures, the horizontal transfer event of the *P-element* into the *D. melanogaster* genome and its immediate spreading into different populations would have happened in the very short span of 40 years.

## 2.6.- *P-element* in other species

The *P-element* was first isolated in *D. melanogaster* (Bingham et al. 1982), but further investigations led to the discovery of *P* homologs in many *Drosophila* species (Clark & Kidwell 1997; Pinsker et al. 2001) and even in closely related genera like *Scaptomyza* (Simonelig & Anxolabéhère 1991). Sequences homologous to the *P-element* have also been detected in other Diptera, like *Musca domestica* (Lee et al. 1999), *Lucilia cuprina* (Perkins & Howells 1992), or *Anopheles* (Sarkar 2003; Oliveira de Carvalho et al. 2004) and have been detected in humans as well (Hagemann & Pinsker 2001). The study of *P-element* distribution reveals several discontinuities suggesting the occurrence of horizontal gene transfer or differential loss of the element (Pinsker et al. 2001).

Moreover, recent studies have uncovered the presence of sequences similar to *P-element* homologous sequences in different vertebrates besides humans, such as *Danio rerio*, *Gallus gallus*, mouse and rat (Quesneville et al. 2005). These sequences, except for that of *Danio rerio*, seem to be located in an orthologous position and that could be the result of an ancient *P-element* domestication (Hammer et al. 2005). Finally, Kimbacher et al. (2009) looked for *P-element* homology in the *Ciona sp*. genome. This organism is a direct descendant of the chordate ancestor, *urochordata*, located phylogenetically at the base of the chordate lineage. The finding of *P-element* sequences with the typical transposon traits (TIRs and TSDs) revealed that this TE could have existed already in the base of vertebrate evolution. Likewise, the stable integration of this *P-element* into the genome in higher vertebrates could be result of a molecular domestication event during evolution of these animals (Kimbacher et al. 2009).

Besides the sequence diversity and subfamilies of *P-element* found in different species (for example, the *P-element* clades in Clark & Kidwell (1997)), a structural dynamism in the copies has been observed as well. Incomplete copies are found that have lost part of the middle region, were the transposase is located. This seems to have an explanation. When a *P-element* has jumped from the donor site, this site has a DSB which needs to be repaired. As mentioned above, this repairing can be done by NHEJ or homologous recombination (gap repair). In this last case, if the synthesis of the new copy of the transposon is accidentally stopped, as the DNA synthesis is triggered from the transposon ends, the central part of the transposon is more prone to disappear from the new copy of the transposon (synthesis-dependent strand annealing SDSA) (Rio 2002). Furthermore, it seems that the shorter a transposon is the higher is its transpositional efficiency, so this accidental shortening might favour the spreading of the short and non-autonomous copies (Atkinson & Chalmers 2010).

In this sense, in some genomes were the *P-element* has been studied with more depth, these short copies, which are called MITEs, have been detected. Usually these shortest copies outnumber the longest and complete ones. For example, in *Anopheles gambiae*, the length of these *P-element* MITEs covers from 205 bp to 2450 bp (Quesneville et al. 2006). MITEs have been found in other transposon superfamilies, and since sometimes their relationship with the whole copies is not very clear, it could be possible that its origin would be by chance through recombination (Gonzalez & Petrov 2009).

## 2.7.- *P-element*-related elements: *1360*

Element *1360* (also referred to as *Hoppel* by Reiss et al. 2003 and as *Proto-P* by Kapitonov & Jurkal 2003) was discovered in the 80s in a region of the long arm of the Y chromosome of *D. melanogaster* (Kholodilov et al. 1988). This sequence was found to harbour terminal inverted repeats and it was repetitive and variable among different strains. In the 90s, more *1360*-like elements were found in the *D. melanogaster* genome. Although none of the copies harboured a coding region, the TIR and TSD structure along with the repetitiveness in the genome, indicated that this was a class II transposable element (Kurenova et al. 1990). The lack of a coding region prevented the element to be assigned to a known superfamily of transposons.

The sequencing of the *D. melanogaster* (Adams et al. 2000), provided the opportunity to look for *P-element* related sequences. The reason for this searches was that, after the discovery of the *P-element* in *D. melanogaster*, this transposon was found to have a wide distribution in the *Sophophora* subgenus, with the exception of the *D. melanogaster* sugroup. This wide distribution suggested the existence of a *P-element* in the ancestor of this subgenus and when the *D. melanogaster* genome sequence was available, different research groups searched for *P-element* sequences descendants of this putative subgenus ancestor. These searches were fruitful and confirmed the hypothesis, mainly thanks to the use of the *P* transposase sequence as query in similarity searches (Kapitonov & Jurka 2003; Reiss et al. 2003).

The *P* related element found turned out to be *1360* elements longer than those characterised in the 90s, encoding a truncated transposase sequence which made possible to place *1360* or *Hoppel* in the *P-element* superfamily of DNA transposons. All the longest *1360* copies harboured truncated transposase sequences and seemed incomplete, but a consensus sequence generated with the different copies pointed out that the putative complete copy would be 4480 bp long, with 31-bp TIR and ~2.6 kb of putative coding region (Kapitonov & Jurka 2003). Although the putative *1360* transposon encodes the same protein domains present in the *P-element* transposase with similarity values of about 40%, *1360* do not harbour any intron (Reiss et al. 2003). Another difference between these two elements is the length of the TSD: 8 bp in the *P-element* and 7 bp in the *1360* element, but this kind of differences among members of the same superfamily is not uncommon (Kapitonov & Jurka 2003).

Furthermore, *1360* element is the most abundant DNA cut-and-paste transposon of the *D. melanogaster* euchromatic genome fraction, reaching a total of 105 copies in the sequenced strain (Kaminker et al. 2002). These copies harbour different deletions and most of them could be considered as non-autonomous elements. Moreover, the *1360* element has been correlated with variegation through iRNA dependent mechanism in *D. melanogaster,* providing insights into a role for TEs in sequence-specific heterochromatic silencing (Haynes et al. 2006). This fact, along with the high copy number of this transposon suggests an important role in genomic regulation and host evolution of TEs.

# 3.- The *Foldback* element

*Foldback* elements are a special group of TEs with a common structural trait, namely, very long and internally repetitive TIRs. Although the existence of terminal inverted repeats and TSD suggest they could be classified as class II elements, the fact that they did not present sequence homology to known transposons and most of them did not harbour any coding sequence, made them to be included in a putative class III of TEs (Capy 1998). After the first *foldback* element was discovered in *D. melanogaster*, structurally similar elements were found in different species, in both animals and plants, such as, sea urchin, *Chironomus thummi*, rice, tomato, *Arabidopsis*, and rye (Hoffman-Lieberman et al. 1989; Hankeln & Schmidt 1990; Rebatchouk & Narita 1997; Cheng et al. 2000; Alves et al. 2005; Daskalova et al. 2005; Marquez & Pritham 2010). All these elements only share structural features, never share similarities in their proteins or DNA sequences. This observation suggests that this group is a kind of hotchpotch where elements from different origins have been put together.

The first *foldback* element (*FB*) was discovered in *D. melanogaster* in the last 80s. Since at this time sequencing techniques were expensive and laborious, indirect techniques to uncover the nature of the DNA sequences were used, such as the search of inverted repeat structures through electron microscopy (Potter et al. 1980). After the de-naturalization and re-naturalization of the DNA, stem-and-loop structures appeared because of the presence of inverted repeats. The detailed study of the sequences that had "folded back" (this is the origin of the name of this class of elements), uncovered the unusual highly repetitive structure of the *FB* TIRs: where a 10 bp sequence is repeated generating a longer repetitive unit in the TIR (Figure 10) (Truett et al. 1981). The sequences of the TIRs are

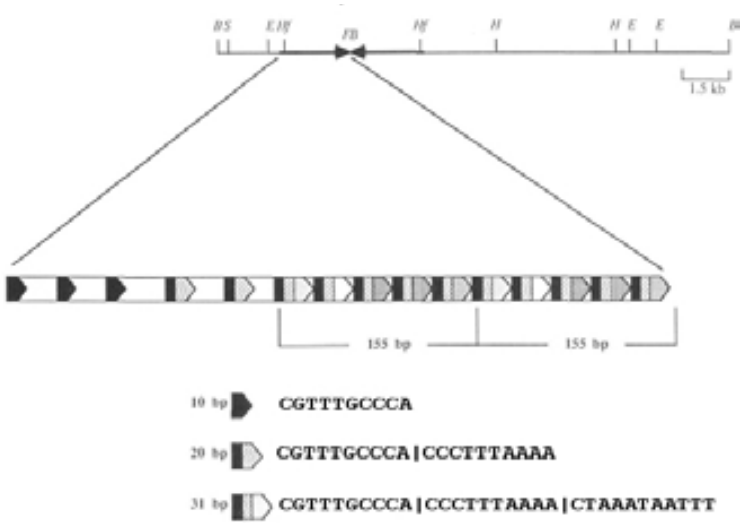



**Figure 10**. Restriction enzyme maps of a *FB* element containing clone. The repetitive structure of the *FB* TIR is depicted. Different repetitive motifs are found along the TIR sequence. From Harden and Ashburner (1990).

similar, but not identical; some sequences are longer than others because the numbers of repetitive units in the TIRs are variable. Likewise, the central region of the *FB* element could be TIR sequence that is missing in the other TIR, because there is an important length difference between the two TIR. However, in some copies of *FB* an extra sequence with putative coding capabilities was found. It was named *NOF* and presented no similarity to other known transposases or proteins, rendering the transposition reaction of these elements as a mystery.

It has been proposed that *NOF* would be an independent transposon with insertion preference for *FB*, because *NOF* is present in few copies of the *FB* element and possesses its own TIR of 308 bp along with a putative coding region with 1 to 3 ORF depending on the *FB-NOF* copy observed (Templeton & Potter 1989; Harden & Ashburner 1990; Badal et al. 2006b). However, the ratio of autonomous to non-autonomous elements (if *NOF* were the *FB* transposase), is similar to other TEs. Furthermore, a *NOF* element without *FB* TIRs has never been found. Thus, it seems reasonable to consider that *NOF* is the transposase-coding ORF of *FB*. Recently, since the TEs catalogue has greatly increased it has been possible to locate the *FB-NOF* protein within a the *MuDR* superfamily of DNA transposons (class II, subclass I, TIR elements order (Feschotte & Pritham 2007; Wicker et al. 2007).

The contribution of *FB* and *FB-NOF* elements to genome plasticity is well known since they are able to promote all sort of genomic rearrangements: inversions, duplications and translocations involving pairs of *FB* elements have been described (Collins & Rubin 1984; Moschetti et al. 2004; Badal et al. 2006a). Likewise, *FB* elements have been reported in the molecular descriptions of different *D. melanogaster* unstable eye mutants. In this sense, *FB* elements have been found responsible for the *white crimson* phenotype in the *white* locus. In these cases the instability has been found to be due to the precise excision of *FB* which originates phenotype revertants (Collins & Rubin 1983; Paro et al. 1983). Nevertheless, there are other cases where interaction with *zeste*1 mutants is the responsible for the eye colour instability (Bingham & Zachar 1985; Rasmuson-Lestander & Ekström 1996; Badal et al. 2006a). Thus, the *FB* transposon generates instability due to both processes, transposition activity and ectopic recombination.

## 4.- The *Galileo* element

The *Galileo* element was discovered when the breakpoints of the *2j* polymorphic chromosomal inversion of *Drosophila buzzatii* were isolated and annotated (Cáceres et al. 1999, 2001). A *Galileo* copy was found in each of the inversion breakpoints. These two *Galileo* copies presented exchanged TSD, which would be a sign of ectopic recombination responsible for the chromosomal inversion (Figure 11). This was the first time a transposon was directly involved in the generation of a chromosomal inversion in natural population. Previously, other inversions were known to have been generated by transposable elements but in laboratory experimental populations (Engels & Preston 1984; Schneuwly et al. 1987; Lim & Simmons 1994). Furthermore, the *2j* inversion presents an adaptive effect in *D. buzzatii*, because different pieces of evidence have been found, such as, (I) the clinal variation of the inversion frequencies along latitunial and altitudinal geographic gradients or (ii) its effect on the adult fly size and the development time (Ruiz et al. 1991; Hasson et al. 1995; Betrán et al. 1998).

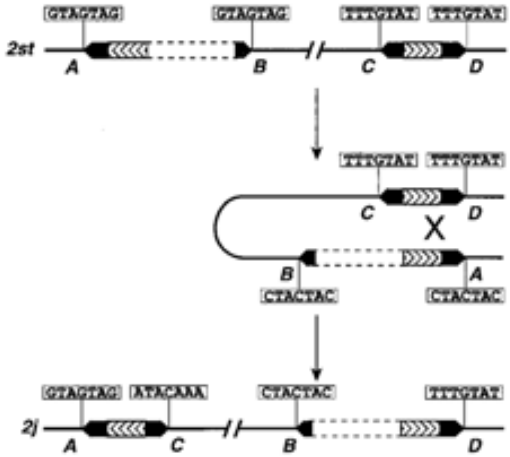



**Figure 11**. Schematic model for the generation of *2j* chromosomal inversion in *D. buzzatii* through ectopic recombination between two *Galileo* copies. The model explains why the TSD of the *Galileo* elements have been exchanged. From Cáceres et al (1999).

In the last decade, our research group has analysed the breakpoints of another two *D. buzzatii* polymorphic inversions, $2q^7$ and $2z^3$ (Casals et al. 2003; Delprat et al. 2009). These two inversions were generated by the same transposable element and the same mechanism, i.e. *Galileo* was the substrate for the ectopic recombination event that generated the inversion. The fact that the same element is involved in three different inversions is noteworthy and suggests *Galileo* unusual structure and/or its transpositional activity contribute to its ability of generate chromosomal inversions (Delprat et al. 2009).

The *Galileo* copies found in the inversion breakpoints were seemingly incomplete because they did not contain any significant coding regions. In a subsequent study in our group (Casals et al. 2005), new *Galileo* copies were isolated from *D. buzzatii* (total

length ranging from 392 to 2304 bp) which corroborated the long TIR of *Galileo* (lengths up to 1115 bp) and its internally repetitive structure with tandem repeats of 136 bp (three and a half repetitions). Furthermore, *Galileo* elements presented target sites duplications of 7 bp, with the palindromic consensus sequence GTAGTAC (Cáceres et al. 2001; Casals et al. 2005). Since *Galileo* copies did not present any similarity to known transposons, it was tentatively classified as a *Foldback* element, using structural criteria because of its main trait: long and internally repetitive TIR (Cáceres et al. 2001; Casals et al. 2005). Furthermore, the study of the breakpoints variability of the *2j* inversion in different *D. buzzatii* strains, uncovered the existence of two closely related elements, which were named *Kepler* and *Newton* (Figure 12). These elements also harboured long TIRs, along with an average 73% sequence identity to *Galileo* TIR, identical 40 bp of the terminal TIR region and TSD of 7 bp long (Cáceres et al. 2001). These traits suggested these elements belonged to the same family, because they shared both structure and sequence identity (Casals et al. 2005).

In neither *Galileo*, *Kepler* and *Newton* copies a putative ORF that could encode the element transposase was found, although in some *Galileo* copies there was a short region encoding a putative protein product with low similarity to the transposase of *1360* (*Hoppel*) element (Casals et al. 2005). Therefore, the *Galileo* copies isolated seemed to be non-autonomous elements in which the coding region could have been deleted and longer *Galileo* copies could exist in the genome with whole coding capability.

The abundance of *Galileo* elements in *D. buzzatii* was assessed by Southern blot and *in situ* hybridization. Southern blot yielded from 21 to 29 *Galileo* copies/genome, with an average of 26.7 copies/genome and no significant different means among the different *D. buzzatii* strains (Casals et al. 2005). In situ hybridization yielded a somewhat higher copy number with no differences among strains but a significant accumulation in the pericentromeric regions and dot chromosome (Casals et al. 2005). Furthermore, when the presence of *Galileo* was explored in other species of the *repleta* group, it was detected only in species closely related to *D. buzzatii* of the *buzzatii*, *martensis* and *stalkeri* clusters. No *Galileo* signal was detected in other more distant species from the *repleta* group, such as *D. mulleri* or *D. repleta*. This could be due to a narrow species distribution of *Galileo* elements or it could be due to the fact that the

sequence divergence of the elements makes them undetectable with the techniques used (Casals et al. 2005).


**Figure 12**. *Galileo*, *Kepler* and *Newton* schematic structure. The TIR region are the different segments considered inverted repeat (IR). The tandem repeats are the dashed rectangles, where the number depicts the number of repetitions. The short region that presented homology with *1360* transposase is depicted. Taken from Casals et al. (2005).

## 5.- *Drosophila* as a model organism

One of the most studied eukaryotes is the fruit fly *Drosophila melanogaster* which has been used as model organism since the beginning of the last century (Figure 13). Thomas H. Morgan was the first scientist to use this fly systematically for Genetics studies, because of its short generation time (10 days), along with the numerous offspring individuals and the phenotypic mutations easy to detect. All these traits made *Drosophila* of exceptional utility for detecting and studying the inheritance of mutations. Furthermore, since *D. melanogaster* is an organism easy to handle and cheap to maintain, its use has been extended to other Biology fields, such as, development, behaviour, physiology, immunology, neuroscience, along with evolution and population genetics. It is worth to mention that 75% of the genes that are involved in human illnesses possess an ortholog gene in *D. melanogaster* genome, a fact that emphasises the importance of the generated knowledge in these flies and encourages further studies (Rubin et al. 2000).

Furthermore, because of its historical importance, large research community, and powerful research tools, as well as its modest genome size (~180 Mb), *Drosophila* was chosen as a test system to explore the applicability of whole-genome shotgun (WGS) sequencing for large and complex eukaryotic genomes (Venter et al. 1998; Adams et al. 2000). This way, the genome of *D. melanogaster* was the second animal genome to be sequenced and annotated. This fact made *D. melanogaster* a model organism for genomics as well, providing the foundation for a new era of sophisticated functional studies and the set up of tools for whole-genome analysis for more complex genomes.

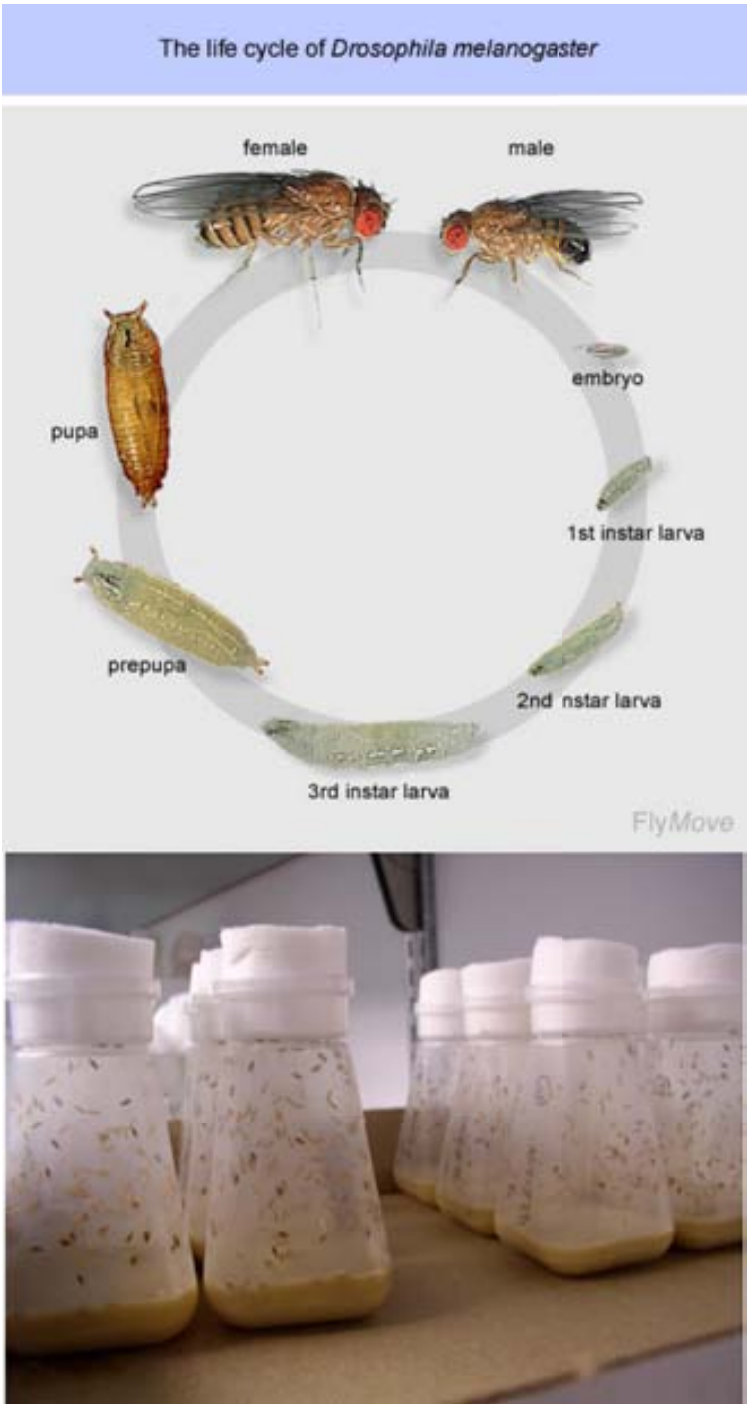



**Figure 13**. a) *Drosophila melanogaster* 10 days life cycle. b) Media flask where *Drosophila* are kept. This media is cheap and easy to handle. Pictures taken from http://www.hoxfulmonsters.com and http://en.wikipedia.org.

## 5.1.- The *Drosophila* genus

The genus *Drosophila* is a very large group of well over 2000 described species that belong to the family Drosophilidae (Markow & O'Grady 2007). Its members are usually called fruit flies (or vinegar fly) because some of its species linger around overripe or rotting fruit. Currently, *Drosophila* is divided into ten subgenera, the largest of which is undoubtedly the subgenus *Drosophila*. The subgenus *Sophophora*, with over 300 described species, is the second largest. Together, the subgenera *Drosophila* and *Sophophora* account for roughly 90 per cent of the diversity in the genus *Drosophila*. Generally, *Drosophila* phylogenetic studies have focused on different groups or species complexes of this genus, which imply that few studies have worked with the whole genus and many aspects of drosophilid phylogeny are controversial or poorly studied (Ashburner et al. 2005; Markow & O'Grady 2006). However, recent molecular systematic studies have shown that this genus is comprised of at least three independent lineages and that several other genera are actually embedded within *Drosophila* (O'Grady & Markow 2009; van der Linde et al. 2010). Since the phylogenetic basis of the genus are not in total agreement with the developed *Drosophila* taxonomy, some *Drosophila* researchers are advocating dividing this genus into three or more separate genera, but others favour maintaining *Drosophila* as a single large genus (Figure 14) (Markow & O'Grady 2006; O'Grady & Markow 2009; van der Linde et al. 2010). The large number of species, along with the huge variability in the ecological habitats and geographical regions where these flies are found, are probably a reflection of the age of the genus, estimated in 40 to 60 myr (Russo et al. 1995; Tamura et al. 2004).

Although *D. melanogaster* is the most studied species of this genus, the other groups of species have been of interest as well, because they are good models for studying speciation patterns, adaptation and relationship with latitudinal gradients, chromosomal evolution and morphology evolution. For example, one of the most eye-catching groups is the Hawaiian *Drosophila* flies, which show a huge variability in size, colour and shapes, along with behaviour (for an example of wing diversity see Edwards et al. 2007). This group comprise a radiation of approximately 1000 species and it seems to be the result of a single colonist lineage that arrived in the islands 25 myr ago (Russo et al. 1995; Markow & O'Grady 2006). This species diversity is a putative result of

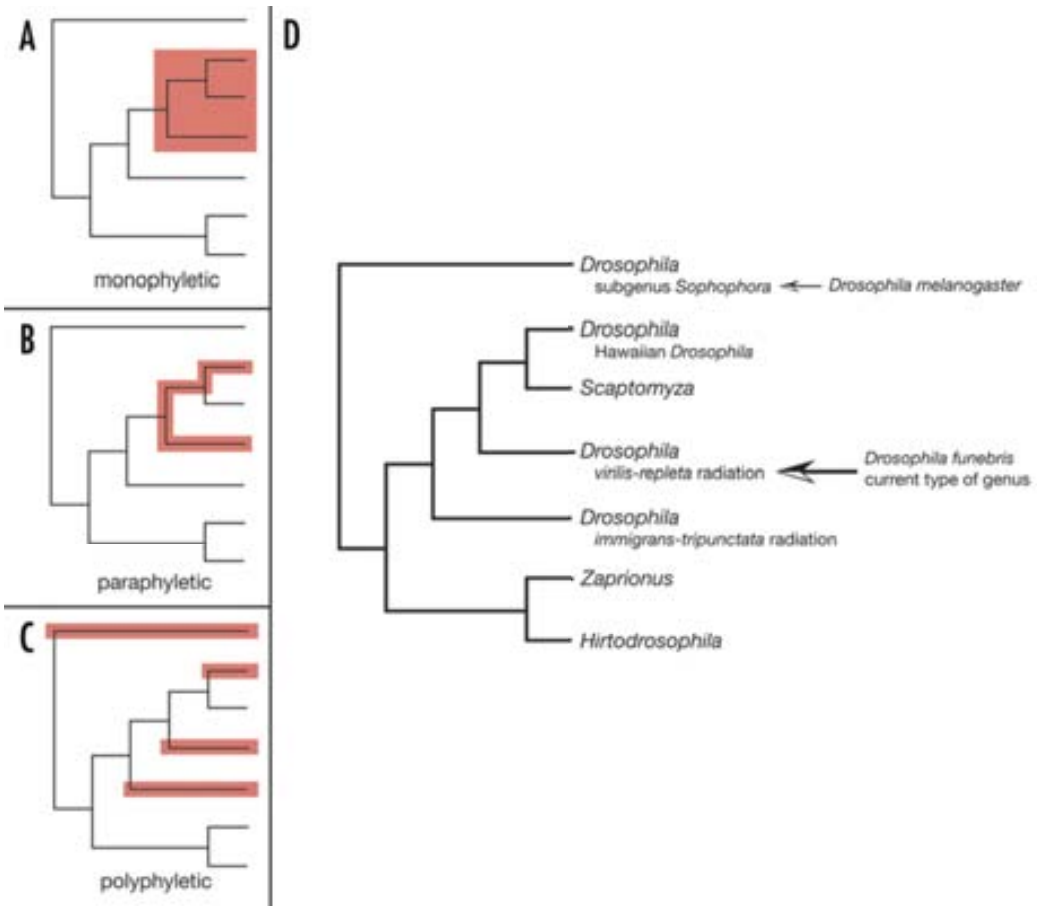



**Figure 14**. Genus *Drosophila* phylogenetic trees showing: a) monophyletic, b) paraphyletic, c) polyphiletic groups in pink d) simplified version of phylogenetic relationships to illustrate the polyphyly of the genus *Drosophila*. Taken from O'Grady and Markow (2009).

different factors, such as, sexual selection, geographic isolation, host plant specialization and morphological innovation (Craddock 2000; Boake 2005). Other species groups which have been studied by ecologist and evolutionary biologist are, for example, the *obscura* group, where we find *D. pseudoobscura*, a well known species studied by Dobzhansky and colleagues. Another example is the virilis group whose speciation and chromosome evolution has been studied broadly (Popadić & Anderson 1994; Caletka & McAllister 2004).

Another important *Drosophila* species group is the *repleta* group. This group is one of the largest and most extensively studied groups in the subgenus *Drosophila*, with more than 90 species classified in six species subgroups – *mulleri*, *hydei*, *mercatorum*, *repleta*, *fasciola*, and *inca*. (Markow & O'Grady 2006; Bächli 2007). Many species of the *repleta* group are adapted to arid or semiarid places and are cactophilic, feeding and breeding on the rotting cactus tissues (Ruiz et al. 1990; Wasserman 1992). The *repleta*

group has served as a model system for evolutionary and ecological studies. Some species have been studied regarding their plant-insect interactions or insect-plant-microbe interactions, along with adaptation to extreme environments (Barker & Starmer 1982; Ruiz & Heed 1988; Barker et al. 1990; Etges et al. 1999; Matzkin & Markow 2009). Furthermore, detailed polytene chromosome maps were conducted for almost all the species of the group and more than 296 inversions were mapped. Several of the chromosomal inversions were variable among closely related species which provided a valuable tool for understanding the phylogeny of this group (Wasserman 1982, 1992). The availability of molecular data offered the opportunity to test and complete the phylogeny provided by the cytological studies. Although some molecular works did not support the monophyletic nature of the *repleta* group, more recent data seem to point in the opposite direction (Durando et al. 2000; Oliveira et al. 2011).

*Drosophila buzzatii* is a cactophilic species that breeds and feeds in the necroses of *Cactaceae*, mainly *Opuntia* and secondarily *Trichocereus* (Hasson et al. 1992). It has an American origin and has recently spread reaching a sub-cosmopolitan distribution which covers South America, South Europe, North Africa and Australia (Fontdevila et al. 1981, 1982; Barker & Starmer 1982). Different aspects of *D. buzzatii* evolutionary biology have been studied such as: geographical patterns of inversion frequencies in both the original species range and the colonizing population of the Old World (Fontdevila et al. 1982; Hasson et al. 1995); latitudinal and altitudinal clines in inversion frequencies (Hasson et al. 1995); the relationship between second chromosome inversions and different phenotypic traits, such as, body size, developmental time, viability and longevity (Ruiz et al. 1991; Betrán et al. 1995; Rodriguez et al. 1999; Fernandez Iriarte et al. 2003); and natural selection in the wild because the knowledge of its breeding sites allows the assessment of changes of inversion frequencies during life cycle (Ruiz et al. 1986; Hasson et al. 1991). This species names its own species complex, the *buzzatii* complex, which belongs to the *mulleri* subgroup in the *repleta* group in the *Drosophila* subgenus (Wasserman 1992; Ruiz & Wasserman 1993).

## 5.2.- *Drosophila* 12 genomes consortium

The extraordinary diversity of *Drosophila* has led to widespread use of species in this genus as model systems for many aspects of genetics, ecology, evolutionary biology, and comparative biology. The existence of the extraordinarily well-annotated genome of *D. melanogaster* (Adams et al. 2000) embedded in the context of a species group with a long history of biological research, immediately motivated the development of comparative genomics in this genus. The *D. pseudoobscura* genome was sequenced in 2005, triggering comparative genomics studies in the *Drosophila* genus (Richards et al. 2005). Afterwards, 10 more *Drosophila* species were chosen to generate a set of 12 *Drosophila* sequenced genomes: *D. melanogaster*, *D. simulans, D. sechellia, D. erecta, D. yakuba, D. pseudoobscura, D. persimilis, D. willistoni, D. virilis, D. mojavensis* and *D. grimshawi* (Figure 15). These genome sequences provide an unprecedented dataset to contrast genome structure, genome content, and evolutionary dynamics across the well-defined phylogeny of the sequenced species (Clark et al. 2003; Singh et al. 2009).

The group of 12 sequenced species, capture a range of evolutionary distances, from closely related sibling species pairs such as *D. simulans* and *D. sechellia*, to more distantly related species defined by the subgenera of *Sophophora* and *Drosophila*. Furthermore, there are species with broad distribution, such as the cosmopolitan species *D. melanogaster* and *D. simulans*, as well as species with highly restricted geographic ranges such as *D. sechellia*, whose distribution is limited to the Seychelles Islands (Indian Ocean). Moreover, generalist and specialist species are multiply represented, a large range of body sizes is encompassed, and a remarkable array of courtship and other behaviours are sampled, as are divergent life histories (Powell 1997; Markow & O'Grady 2007). Besides the common traits, these differences among the 12 *Drosophila* species would be studied in depth thanks to the availability of the sequenced genomes and it allows *Drosophila* researchers to place their questions in a phylogenetic context.

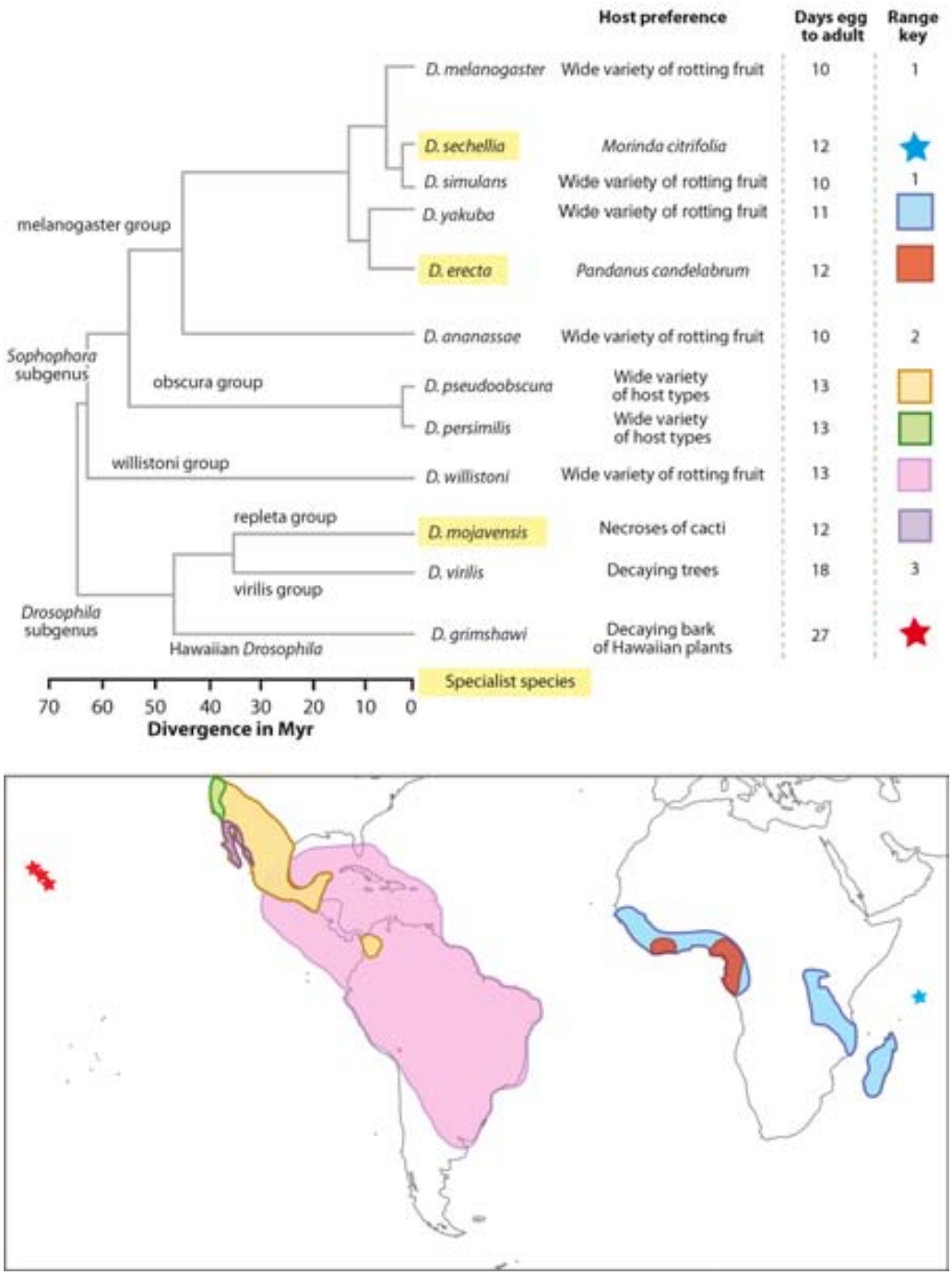



**Figure 15**. Phylogeny of the 12 sequenced species of *Drosophila* showing host preference for oviposition, developmental time from egg to adult in days, and the approximate geographic ranges of the species. Divergence times between species are in millions of years (Tamura et al. 2004) . Geographic ranges of different species (the ones with a "range jey") are depicted. Modified from Signh et al (2009).

The 12 *Drosophila* genomes provide a tool to study the evolution of other types of DNA sequences besides the protein-coding genes, such as the TEs. These genomes provide a landscape where the relationship among the different genomes and TEs could be studied, not only in one species, but rather from a phylogenetic perspective. Genomic TE content is a variable trait that differ among the species. Some TEs appeared to be in

the genus from the beginning, such as the telomeric retrotransposons (Villasante et al. 2007). Other present a patchy distribution among the species, which could be a result of genomic losses or horizontal transfer events (Loreto et al. 2008). Furthermore, different classes of transposable elements can vary in abundance owing to a variety of host factors, motivating an analysis of the intragenomic ecology of transposable elements in the 12 genomes. Although comprehensive analysis of the structural and evolutionary relationships among families of transposable elements in the 12 genomes remains a major challenge for *Drosophila* genomics, some initial insights can be gleaned from analysis of particularly well-characterised transposable element families. The use of these 12 genomes also facilitated the discovery of transposable element lineages not yet documented in *Drosophila*, and a deeper study of the already known (*Drosophila* 12 Genomes Consortium et al. 2007; Singh et al. 2009).

# II.- OBJECTIVES

The *Galileo* element has been directly involved in the generation of three different natural chromosomal inversions in *D. buzzatii*. All copies found in the inversion breakpoints as well as other copies isolated in our research group were incomplete copies with no significant similarity to any known TE neither any known protein. Hence, the *Galileo* element was worth to study in more depth due to its implication in the *D. buzzatii* chromosomal evolution and its unknown nature as TE. Furthermore, the availability of the 12 sequenced *Drosophila* genomes provided a very useful tool, not only to look for *Galileo*-like elements, but also for studying the TE from a genomic perspective.

The main objective of this thesis is to fully characterise the transposon *Galileo* along with its classification based on functional means, such as the putative *Galileo* mobilization proteins. Moreover, the classification allows the comparison of *Galileo* with related transposons. Furthermore, other objectives of this thesis are to analyse the *Galileo* copies found in different genomes and compare them inter-species and intra-species, to test biochemically that the detected transposase interacts with *Galileo* TIR sequences and, finally, characterise and study the dynamics of the *Galileo* long TIR.

This thesis is divided in three chapters. Each of them has different specific objectives that are in part a consequence of previous results.

In the first chapter, the objectives are:

- To find a complete or nearly-complete copy of *Galileo* (which means a copy with a protein-coding ORF) in the genome where *Galileo* was discovered, *D. buzzatii*.
- To look for similar elements in the publicly available sequenced genomes of 12 *Drosophila* species.
- To unequivocally classify *Galileo*.
- To compare *Galileo* with other related TEs.
- To analyse the different *Galileo* elements found in each genome

In the second chapter, the objectives are:

- To reconstruct nucleotide coding for a functional *Galileo* transposase in *D. buzzatii* and nucleotide coding sequences for the transposase DNA binding domain in three different species (*D. buzzatii*, *D. mojavensis* and *D. ananassae*).
- To express and purify the transposase DNA binding domains and *in vitro* test its binding properties
- To isolate and determine the nucleotide sequence of the binding site of the transposase binding domain in *D. buzzatii*
- To test *Galileo* whole transposition reaction in *D. melanogaster* through plasmid transformation of embryos and fly crosses.

In the third chapter of this thesis the objectives are:

- To isolate all *Galileo* copies in the *D. mojavensis* sequenced genome.
- To carefully annotate all the regions in each *Galileo* copy.
- To study the phylogenetic relationship among the elements taking into account the TIR and the transposase sequence and compare the results.
- To study the *Galileo* chormosomal distribution and its relation with *D. mojavensis* genes
- To study the composition and the cause of variation in *Galileo* TIR length and propose mechanism responsible for it.

# III.- MATERIALS AND METHODS

# 1.- *Drosophila* strains

In this work the following *Drosophila* strains have been used for molecular work:

- *D. buzzatii* st-1, Maz-4, j-9, jq7-4, jz3-2, jq7-1, Sar-9 and j-4.
- *D. mojavensis* 15081-1352.22, Tucson Stock Center. This is the stock used for genome sequencing (*Drosophila* 12 genomes consortium 2007).
- *D. melanogaster white* strain (w1118)

The 12 sequenced *Drosophila* genomes have been used for *in silico* analyses. For the genomes of *D. melanogaster* (strain reference: 10421-0231.36, Tucson Stock Center) and *D. simulans* (strain reference: 10421-0251.195, Tucson Stock Center) the assembly which has been analysed corresponds to CAF1 chromosomes. For the rest of species *D. sechellia* (strain reference: 10421-0248.25, Tucson Stock Center), *D. yakuba* (strain reference: 10421-0231.36, Tucson Stock Center), *D. erecta* (strain reference: 10421-0224.01, Tucson Stock Center), *D. ananassae* (strain reference: 10421-0371.13, Tucson Stock Center), *D. pseudoobscura* (strain reference: 10421-0121.94, Tucson Stock Center), *D. persimilis* (strain reference: 10421-0111.49, Tucson Stock Center), *D. willistoni* (strain reference: 10421-0811.24, Tucson Stock Center), *D. virilis* (strain reference: 10421-1051.87, Tucson Stock Center) and *D. grimshawi* (strain reference: 10421-2541.00, Tucson Stock Center) the CAF1 contigs assembly was analysed. In the case of *D. mojavensis* (strain reference: 10421-1352.22, Tucson Stock Center), both CAF1 contigs and scaffolds assemblies have been explored.

# 2.- Molecular techniques

## 2.1.- Nucleic acids isolation (Genomic and plasmid)

Total genomic DNA was extracted from 0.2 g of adult flies following the protocol described by Piñol et al. (1988). Plasmid DNA was extracted using standard methods (Sambrook et al. 1989). The quality of the purified DNA was checked with an agarose gel.

## 2.2.- PCR

PCRs were performed in a total volume of 25 µl, including 1 µl of cDNA or 100-200 ng of genomic DNA, 10 pmol of each primer, 200 µM dNTPs, 1.5 mM MgCl2, and 1.5 units of Taq DNA polymerase (Roche or Bioron) or Phusion enzyme (Finnzymes). Typical cycling conditions were 30 rounds of 30 sec at 94ºC, 30 sec at 55-60ºC (depending on the primer pair used), and 60 sec at 72ºC. The PCR products were loaded in an agarose gel and purified with QiaQuick kit (Qiagen).

## 2.3.- Plasmid generation

For testing the transposition reaction of *Galileo in vivo*, a two plasmid system was generated consiting in a helper plasmid, where *Galileo* transposase was cloned, and a donor plasmid, where the mini*white* gene was contained in between two *Galileo* TIRs with TSD. The co-injection of these two plasmids in *Drosophila white* embryos and the posterior screening of the F1 generation should show when the transposition reaction has happened because individuals with coloured eyes shall appear. In this experiment the *P-element* transformation vectors were used as positive control, whereas the donor plasmid alone was used as negative control. The details of the generation of the plasmids are found in the second chapter of results.

## 2.4.- Protein assays

Protein expression and purification

Different ORF of the putative DNA binding domain proteins inferred (see below) were cloned in expression vectors (N-ter MBP-tag vector from The Oxford Protein Production Facility, UK) and transformed in *Escherichia. coli* BL21 (DE3) expression cell strain. The protein expression was induced in DO680 =0.5 LB cultures with 100

ug/ml ampicillin cultures, 1mM of IPTG and 100uM of $ZnCl_2$ at 16ºC over night. The cells were harvested by centrifugation and resuspended in HSG buffer (50mM HEPES pH 7.5, 200mM NaCl, 2mM dithiothreitol (DTT), 5mM EDTA and 10% glycerol). The cells were lysed in a French press and centrifuged at 25000g for 30 min. The supernatant was loaded onto an amylose resin column (New England Biolabs). The column was washed several times with HSG buffer and the protein eluted with HGS buffer plus 10mM maltose. The fractions containing MBP transposase were pooled and aliquots were stored at -80ºC.

Electrophoretic mobility shift assay

This assay was performed to test the binding activity of the expressed *Galileo* protein domains. The purified recombinant THAP domains were incubated for 2 hours at room temperature with the labelled TIR in 20 ul reaction of binding buffer (20 mM Tris-HCl, pH 7.5, 100 mM KCl, 100 g/ml bovine serum albumin, 2.5 mM DTT, 5% glycerol). Different conditions were tested: different protein concentration (1, 1:100, 1:10000 from the stock protein solution (5ug/µL or 94 µM), addition of ZnCl2 (100 µM final concentration) and addition of unspecific competitor DNA (pBlueScript, ~500ng/reaction). The reactions were loaded in a 4% TAE-polyacrilamide gel and run for 2 hours at 300V at 4ºC.

Footprint assay

A sample of the EMSA reaction was digested by 0.05U of DNase I for 1 minute at room temperature. The enzyme was diluted to 1U/µL with dilution buffer (5 mM MgCl2, 0.5 mM CaCl2). The reaction was stopped using 1 µL of 500 mM EDTA. DNA was purified by phenol-chloroform extraction and ethanol precipitation. The cleavage pattern was analysed by electrophoresis on a 5% polyacrylamide sequencing gel. DMS/piperidin reactions were performed following standard procedures to reveal G positions and were used to localize the DNase I protected regions.

## 3.- Sequence analysis

The sequences obtained in the different PCRs were assembled with Geneious and aligned with Muscle 3.6 software (Edgar 2004; Drummond et al. 2010). The sequences were compared to previous ones using Blast searches and alingments (Altschul et al. 1997; Katoh et al. 2002; Edgar 2004).

The 12 genomes searches were performed with Blast algorithms, using tBlastn for looking for putative ORF and Blastn for non-coding sequences. Different thresholds of scores have been used in the different searches during this thesis: an e-value of $10^{-20}$ (which corresponds to a fragment of at least ~200 amino-acids with a ~30 % of identity for tBlastn searches); an e-value of $10^{-3}$ for Blastn searches (which corresponds to 21-22 identical consecutive nucleotides); and an 80-80 criteria, where at least an 80% of the length of the query was found along with a minimum of 80% identity. Different sequences have been used as query, such as *Galileo* TIR, *Galileo* transposase, *Galileo* whole element of each species. In each of the results chapters, these details are specified. The parameters of the different Blast searches have been used as they are set by default.

The sequences detected with the different Blast searches have been thoroughly annotated using a group of different tools, most of them implemented in the Geneious software, such as dotplot graphics for detecting repetitions and its span, different alignment algorithms and custom Blast searches with specific *Galileo* and *Drosophila* TEs databases (Drummond et al. 2010). All the *Galileo* copies found have been classified regarding identity and phylogenetic inference in different subfamilies, and the internal structure of each copy has been explored, annotating TIR regions, transposase regions, F1 and F2 spacing regions, tandem repeat regions and insertions.

The putative ORF found in this work have been conceptually translated. In all copies *Galileo* ORF presented frame-shift and premature stop codons mutations. In these cases a consensus was reconstructed using all the sequences available and a majority rule. The obtained sequences have been analysed using Blastp and domains have been detected with Domain Conserve Search, InterProScan and Coils servers (Lupas et al. 1991; Zdobnov & Apweiler 2001; Marchler-Bauer et al. 2005).

The MEGA software have been used for calculation of the pairwise number of differences among different sets of sequences (p-distance) (Tamura et al. 2004). These nucleotide differences have been transformed to absolute time using the *Drosophila* evolutionary rates of 0.016 changes/position/myr and 0.011 changes/position/myr (Li 1997; Tamura et al. 2004).

The different set of sequences have been aligned and filtered with Gblocks using relaxed parameters (Talavera & Castresana 2007). jModelTest was run to find the best evolutionary model for the different sets of sequences and phylogenetic trees were inferred. For these inferences, different computer programs have been used, such as MEGA 4 for Neighbor-joining trees, PhyML and RAxML for maximum-likelihood inferences and BEAST for Bayesian inferences (Guindon & Gascuel 2003; Stamatakis 2006; Drummond & Rambaut 2007; Tamura et al. 2007).

*Ad hoc* perl scripts have been used to analyse the inter-chromosomal and intrachromosome distribution of *Galileo* and to compare their position to the predicted genes in the genome. The software package JMP 8.0.2 (SAS Institute Inc. 2009) has been used for performing statistical tests.

# IV.- RESULTS

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

Mar Marzo, Marta Puig and Alfredo Ruiz
Departament de Genètica i Microbiologia, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain

E-mails: mar.marzo@uab.cat, marta.puig@uab.cat, alfredo.ruiz@uab.cat
Corresponding author: alfredo.ruiz@uab.cat

Keywords: Transposable Element, *P-element*, *Galileo*, *Foldback*, THAP domain, TIR, DNA binding, transposase, reconstruction

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

**Mar Marzo, Marta Puig, and Alfredo Ruiz***

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

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

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

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

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

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

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

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

The authors declare no conflict of interest.

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

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

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

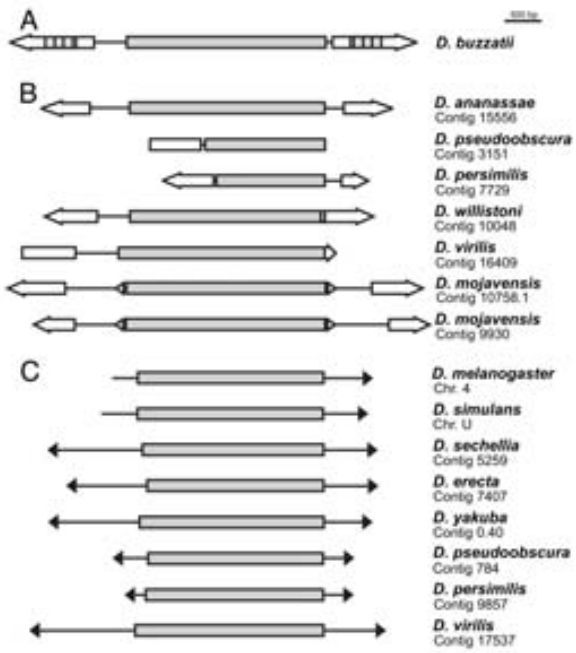



**Fig. 1.** Most complete copies of *Galileo* and *1360* found in this work. (*A*) Putative complete *Galileo* copy from the *D. buzzatii* genome. (*B*) Most complete copies of *Galileo* found in the 12 sequenced genomes. (*C*) Most complete copies of *1360*. TIRs are represented as arrows and TPases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

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

## Results

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

A search using BLASTX revealed significant similarity of the *Dbuz\Galileo* TPase to those of the related *D. melanogaster 1360* and *P* elements (25, 26) [AAN39288, E-value = 1e−95; Q7M3K2, E-value = 3e−25]. The *Dbuz\Galileo* TPase includes a THAP domain near the N terminus (amino acids 27–104) similar to the DNA binding domain of *P* element TPase (27–30). A copy of *1360* located in chromosome 4 of *D. melanogaster* (31) encodes a TPase (854 aa) longer than that in the National Center for Biotechnology Information database (25), including a THAP domain near the N terminus (after curation of a 1-bp frameshift mutation). A global alignment of the *Dbuz\Galileo* TPase with those of *Dmel\1360* and *Dmel\P* yielded 34.5% and 27.6% identity, respectively. No significant similarity was found between the *Dbuz\Galileo* TPase and the proteins encoded by *Dmel\FB* (14, 15).

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

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

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

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

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


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

evidence for more recent transpositional events was found because a subgroup of 13 copies shows an average divergence of 0.36% equivalent to a divergence time of only 0.225 myr. Similar observations were made in *D. pseudoobscura*, *D. persimilis*, and *D. willistoni* (SI Table 6). In each case, subgroups with ~1% average divergence (implying divergence times ~0.6 myr) were found. In *D. virilis*, analysis of 13 short nonautonomous copies uncovered two highly divergent groups that we named A and B (SI Fig. 5). Copies within each group were aligned and analyzed separately (SI Table 6). The average pairwise divergence within groups A and B was 4.6 and 5.7%, implying divergence times of 2.9 and 3.6 myr, respectively. Inclusion in the analysis of the longest copy found in the species (contig 16409) indicated unequivocally that it is a member of group A (SI Fig. 5). In *D. mojavensis*, analysis of 20 short nonautonomous copies revealed the presence of four well defined groups, here named C–F. We included in the analysis nine of the long copies containing the two TIRs and generated a phylogenetic tree with the 29 copies (Fig. 2). Groups C and D correspond to the two groups previously detected when the long, nearly complete, copies were analyzed. Copies within each group were separately aligned and analyzed. Average pairwise divergences within groups C through F were 2.2%, 2.3%, 2.4%, and 8.9%, respectively, indicating divergence times ranging from 1.4 to 5.5 myr (SI Table 6). The two and four *Galileo* groups or subfamilies found in *D. virilis* and *D. mojavensis*, respectively, seemingly represent relatively old tranposition bursts in these genomes. We suggest that the *Newton* and *Kepler* elements previously found in the *D. buzzatii* genome (18, 21) should likewise be considered only as different groups or subfamilies of *Galileo* in this species.

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

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

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

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

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

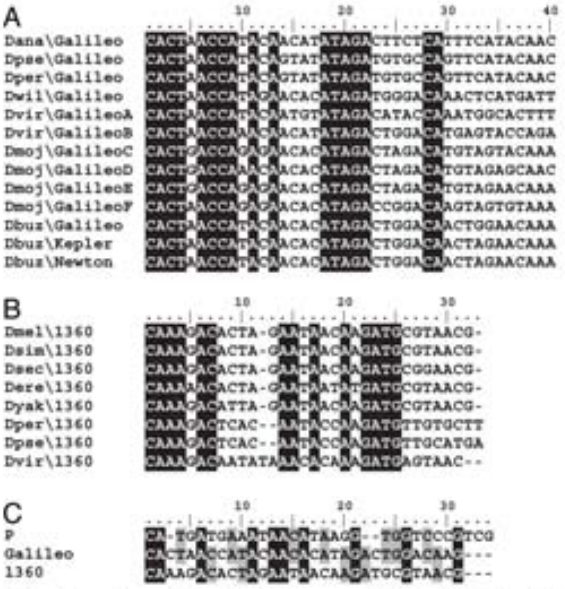



**Fig. 3.** Comparison of TIR ends. (*A*) Alignment of 40 bp of the TIR end of *Galileo*. A consensus sequence was constructed for *Galileo* TIRs in each TE subfamily and species. (*B*) Alignment of the 31-bp TIR of *1360*. A representative TIR from a single copy of the TE is included. (*C*) Comparison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

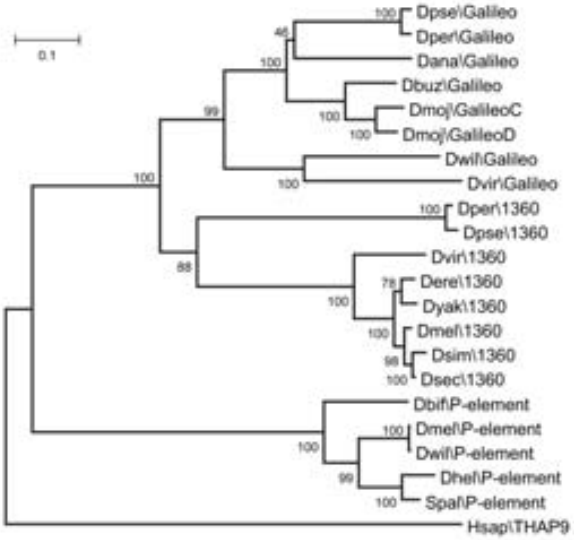



**Fig. 4.** Neighbor-joining phylogenetic tree constructed with the eight consensus *Galileo* TPases, eight consensus *1360* TPases, and five TPases from representative *P* elements. The human *P*-like THAP9 protein is included as an outgroup. The complete alignment without Gblocks filtering is shown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

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

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

## Discussion

We characterized a seemingly complete copy of *Galileo* from the genome of *D. buzzatii* that contains a 2,738-bp ORF encoding a TPase. Three observations indicate that this is the true *Galileo* TPase instead of that of another TE accidentally associated with the long *Galileo* TIRs. (*i*) Two previously isolated *Galileo* copies bear a 141-bp portion of the same ORF in the right position and orientation (21), suggesting that all previously isolated *Galileo* copies are defective versions of the complete structure reported here. (*ii*) Our bioinformatic searches uncovered TEs structurally similar to *Galileo* in the genomes of six phylogenetically distant *Drosophila* species. These searches were carried out by using as queries the *Dbuz\Galileo* and *Dmel\1360* TPases, and a careful scrutiny of the contigs producing significant hits led to the finding of the TIRs associated with the TPase segment and the characterization of the elements as either *Galileo* or *1360*. No other TIRs besides those of these two TEs were found flanking the hits (but note that in *Dmoj\Galileo* 160-bp internal inverted repeats bound the TPase; Fig. 1*B*). The persistent association (over tens of myr) of this TPase with the same type of TIRs renders the possibility of an accidental association extremely unlikely. (*iii*) The presence of multiple *Galileo* copies comprising both TIRs and TPase-coding segments in seven *Drosophila* genomes suggests that these are integral components of the same elements, and these elements are (or have been) able to replicate and transpose within these genomes.

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

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

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

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

The origin of the numerous chromosomal inversions in *Drosophila* and other Dipterans is still an open question and very few species have been investigated in this regard. Strong evidence implicating TE-mediated ectopic exchange has been found in four polymorphic inversions only, including the two *D. buzzatii* inversions generated by *Galileo* (39). In *D. melanogaster* and its close relatives, no TEs have been involved in the origin of three polymorphic inversions and only 2 of 29 fixed inversions contain repetitive sequences inverted with respect to each other at both breakpoints, pointing to a completely different mechanism for inversion generation (39). The fact that *Galileo* generated two independent inversions in *D. buzzatii* suggests that *Galileo* is not a passive substrate where ectopic recombination operates but may be actively generating inversions as a byproduct of its transposition mechanism. If this is correct, to create inversions, *Galileo* has to be active in a genome and a recent transpositional activity would be a necessary condition for *Galileo* to have any role in the generation of current inversions. We have not found any functional TPase in any of these species but only one genome was sequenced in each case, so they could still exist in unsequenced genomic regions, other genomes, and/or other natural populations. However, we have provided evidence of recent (<1 myr) transpositional activity of *Galileo* in *D. ananassae*, *D. persimilis*, *D. pseudoobscura*, and *D. willistoni*. These four are among the most polymorphic species of the genus with 24, 28, 13, and 50 inversions, respectively (40). In *D. mojavensis*, with fewer inversions (41), the most recent transpositional activity of *Galileo* seems somewhat older (≈1.5 myr). Finally, *D. virilis* with the oldest *Galileo* activity (≈3 myr) is chromosomally monomorphic (40). Therefore, there is a qualitative correlation between the number of inversions and the time of the most recent activity of *Galileo* in this small group of species. This correlation is suggestive but might be only coincidental. However, the detection of chimerical copies that may be the result of chromosomal rearrangements (19) indicates that, indeed, *Galileo* might have been involved in the origin of inversions, at least in some other species besides *D. buzzatii*.

## Methods

**PCR Amplification and DNA Sequencing.** Genomic DNA from *D. buzzatii* (strain st-1) and *D. mojavensis* (strain 15081-1352.22, Tucson *Drosophila* Stock Center) (as control) was used as template for PCR amplification of *Galileo* copies. Primers located in the TIRs were designed based on *D. buzzatii* known incomplete copies of *Galileo* (21), whereas primers inside the TPase were designed on the *D. mojavensis* putative complete TPases found in a preliminary bioinformatic search (SI Fig. 7). Primers in the TIRs were always used in combination with primers anchored in the TPase to avoid multiple bands generated by the highly repetitive primer alone or the amplification of defective copies without TPase. PCRs were carried out in a total volume of 25 μl including 100–200 ng of genomic DNA, 20 pmol of each primer, 200 μM dNTPs, 1.5 mM $MgCl_2$, and 1–1.5 units of Taq DNA polymerase. PCR products were gel-purified by using QIAquick Gel Extraction kit (Qiagen) and sequenced directly with the amplification primers and sequencing primers designed over the end sequences to close gaps (SI Fig. 7). Sequences were aligned and assembled by using multialign software MUSCLE 3.6 (42).

**Bioinformatic Searches.** BLAST searches were performed on the chromosome assemblies of *D. melanogaster* and *D. simulans* and the contig CAF1 assemblies of the other ten publicly available *Drosophila* genomes (http://rana.lbl.gov/drosophila). We used BLAST algorithm version 2.2.2 (43) implemented in the Drosophila Polymorphism Database server (http://bioinformatica.uab.es/dpdb) with default parameters. TBLASTN searches in the different species were performed by using as queries the TPases of *Dbuz\Galileo* and *Dmel\1360* (SI Table 1). Hits with an E-value ≤ $10^{-20}$ (which in the conditions of our searches amounts roughly to ≈30% identity over a stretch of 200 aa) were considered significant. BLASTN searches were also carried out with the 40 terminal bp of *Dbuz\Galileo* and the 31 bp of the *Dmel\1360* TIR (SI Table 1). The cutoff in this case was an E-value ≤ $10^{-3}$ (that requires ≈21–22 consecutive identical base pairs).

Contigs producing significant hits with the *Dbuz\Galileo* and *Dmel\1360* TPases in each species were scrutinized to characterize the different copies of both TEs. TIRs and TSDs were searched around the putative TPases by using Dotlet 1.5 (44) to define the boundaries of each copy. Insertions of other TEs inside Galileo were identified by aligning the different *Galileo* copies found in the same species and further analyzing the sequences present in only one of them. Significant contigs <1 kb long and those that were found to contain complex clusters of several TE insertions (likely of heterochromatic origin) were not further investigated.

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

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

**Consensus Sequences.** The consensus sequences for *Galileo* and *T360* TPases and *Galileo* TIRs were generated by using BioEdit 7.0.5 (45) after aligning the respective nucleotide sequences (SI Table 9) with MUSCLE 3.6 software (42). In the case of TPases, this consensus sequence was then translated into protein to allow the comparison among different species (SI Fig. 6). Conserved protein domains were detected by using InterProScan (46) and Conserved Domain Search (47). Coiled-coil regions were predicted by using the Coils server (48).

**Phylogenetic Analyses.** TPase sequences were aligned with MUSCLE 3.6 (42) and the alignment was filtered with Gblocks version 0.91b (49) to remove the poorly aligned and highly divergent segments. Gblocks was used with the default parameters except for the maximum number of contiguous nonconserved positions = 15, the minimum length of a block = 6, and allowed gap position = half. These parameters were fixed so that the conserved THAP domain was included in the filtered alignment. All phylogenetic trees were constructed with MEGA 3.1 (50) by using the neighbor-joining method with complete deletion and 500 replicates to generate bootstrap values. Poissor correction and Kimura 2 parameters were used as substitution models for amino acid and nucleotide sequences, respectively. We dated the most recent transposition events within each species by dividing the average pairwise divergence between the elements in the same group or subgroup by the *Drosophila* synonymous substitution rate, 0.016 substitutions per nucleotide myr (21). To date the divergence between different groups or subfamilies we calibrated the tree with the same substitution rate by using the appropriate option in MEGA (50). Time estimates for TEs should be taken with caution; if the synonymous substitution rate were an underestimate of the true mutation rate for TEs, our time estimates would provide an upper bound for the true values.

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

1. Kidwell MG, Lisch DR (2002) in *Mobile DNA II*, eds Craig NL, Craigie R, Gellert M, Lambowitz AM (American Society for Microbiology, Washington, DC), pp 59–89.
2. Capy P, Bazin C, Higuet D, Langin T (1998) *Dynamics and Evolution of Transposable Elements* (Springer, Heidelberg).
3. Feschotte C, Pritham EJ (2007) DNA transposons and the evolution of eukaryotic genomes. *Annu Rev Genet* 41:331–368.
4. Haren L, Ton-Hoang B, Chandler M (1999) Integrating DNA: transposases and retroviral integrases. *Annu Rev Microbiol* 53:245–281.
5. Rio DC (2002) in *Mobile DNA II*, eds Craig NL, Craigie R, Gellert M, Lambowitz AM (American Society for Microbiology, Washington, DC), pp 484–518.
6. Hartl DL, Lohe AR, Lozovskaya ER (1997) Modern thoughts on an ancient marinere: Function, evolution, regulation. *Annu Rev Genet* 31:337–358.
7. Potter S, Truett M, Phillips M, Maher A (1980) Eucaryotic transposable genetic elements with inverted terminal repeats. *Cell* 20:639–647.
8. Truett MA, Jones RS, Potter SS (1981) Unusual structure of the FB family of transposable elements in Drosophila. *Cell* 24:753–763.
9. Liebermann D, *et al.* (1983) An unusual transposon with long terminal inverted repeats in the sea urchin Strongylocentrotus purpuratus. *Nature* 306:342–347.
10. Rebatchouk D, Narita JO (1997) Foldback transposable elements in plants. *Plant Mol Biol* 34:831–835.
11. Ade J, Belzile FJ (1999) Hairpin elements, the first family of foldback transposons (FTs) in Arabidopsis thaliana. *Plant J* 19:591–597.
12. Simmen MW, Bird A (2000) Sequence analysis of transposable elements in the sea squirt, Ciona intestinalis. *Mol Biol Evol* 17:1685–1694.
13. Windsor AJ, Waddell CS (2000) FARE, a new family of foldback transposons in Arabidopsis. *Genetics* 156:1983–1995.
14. Templeton NS, Potter SS (1989) Complete foldback transposable elements encode a novel protein found in Drosophila melanogaster. *EMBO J* 8:1887–1894.
15. Harden N, Ashburner M (1990) Characterization of the FB-NOF transposable element of Drosophila melanogaster. *Genetics* 126:387–400.
16. Pritham EJ, Feschotte C, Wessler SR (2005) Unexpected diversity and differential success of DNA transposons in four species of entamoeba protozoans. *Mol Biol Evol* 22:1751–1763.
17. 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.
18. 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.
19. Casals F, Caceres 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.
20. Lim JK, Simmons MJ (1994) Gross chromosome rearrangements mediated by transposable elements in Drosophila melanogaster. *BioEssays* 16:269–275.
21. Casals F, Caceres M, Manfrin MH, Gonzalez J, Ruiz A (2005) Molecular characterization and chromosomal distribution of Galileo, Kepler and Newton, three foldback transposable elements of the Drosophila buzzatii species complex. *Genetics* 169:2047–2059.
22. Adams MD, *et al.* (2000) The genome sequence of Drosophila melanogaster. *Science* 287:2185–2195.
23. Richards S, *et al.* (2005) Comparative genome sequencing of Drosophila pseudoobscura: Chromosomal, gene, and cis-element evolution. *Genome Res* 15:1–18.
24. Drosophila 12 Genomes Consortium (2007) Evolution of genes and genomes on the Drosophila phylogeny. *Nature* 480:203–218.
25. Reiss D, Quesneville H, Nouaud D, Andrieu O, Anxolabehere D (2003) Hoppel, a P-like element without introns: a P-element ancestral structure or a retrotranscription derivative? *Mol Biol Evol* 20:869–879.
26. Laski FA, Rio DC, Rubin GM (1986) Tissue specificity of Drosophila P element transposition is regulated at the level of mRNA splicing. *Cell* 44:7–19.
27. Lee CC, Beall EL, Rio DC (1998) DNA binding by the KP repressor protein inhibits P-element transposase activity in vitro. *EMBO J* 17:4166–4174.
28. Clouaire T, *et al.* (2005) The THAP domain of THAP1 is a large C2CH module with zinc-dependent sequence-specific DNA-binding activity. *Proc Natl Acad Sci USA* 102:6907–6912.
29. Roussigne M, *et al.* (2003) The THAP domain: A novel protein motif with similarity to the DNA-binding domain of P element transposase. *Trends Biochem Sci* 28:66–69.
30. Quesneville H, Nouaud D, Anxolabehere D (2005) Recurrent recruitment of the THAP DNA-binding domain and molecular domestication of the P-transposable element. *Mol Biol Evol* 22:741–746.
31. Kapitonov VV, Jurka J (2003) Molecular paleontology of transposable elements in the Drosophila melanogaster genome. *Proc Natl Acad Sci USA* 100:6569–6574.
32. Hagemann S, Pinsker W (2001) Drosophila P transposons in the human genome? *Mol Biol Evol* 18:1979–1982.
33. Richardson C, Moynahan ME, Jasin M (1998) Double-strand break repair by interchromosomal recombination: suppression of chromosomal translocations. *Genes Dev* 12:3831–3842.
34. Kapitonov VV, Jurka J (2006) Self-synthesizing DNA transposons in eukaryotes. *Proc Natl Acad Sci USA* 103:4540–4545.
35. Page RD, Charleston MA (1998) Trees within trees: Phylogeny and historical associations. *Trends Ecol Evol* 13:356–359.
36. Silva JC, Loreto EL, Clark JB (2004) Factors that affect the horizontal transfer of transposable elements. *Curr Issues Mol Biol* 6:57–71.
37. Clark JB, Kidwell MG (1997) A phylogenetic perspective on P transposable element evolution in Drosophila. *Proc Natl Acad Sci USA* 94:11428–11433.
38. Silva JC, Kidwell MG (2000) Horizontal transfer and selection in the evolution of P elements. *Mol Biol Evol* 17:1542–1557.
39. Ranz JM, *et al.* (2007) Principles of genome evolution in the Drosophila melanogaster species group. *PLoS Biol* 5:e152.
40. Sperlich D, Pfriem P (1986) in *The Genetics and Biology of Drosophila*, eds Ashburner M, Carson HL, Thompson JNJ (Academic, London), pp 257–309.
41. 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.
42. Edgar RC (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792–1797.
43. Altschul SF, *et al.* (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402.
44. Junier T, Pagni M (2000) Dotlet: Diagonal plots in a web browser. *Bioinformatics* 16:178–179.
45. Hall TA (1999) BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp Ser* 41:95–98.
46. Zdobnov EM, Apweiler R (2001) InterProScan-an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* 17:847–848.
47. Marchler-Bauer A, *et al.* (2005) CDD: A Conserved Domain Database for protein classification. *Nucleic Acids Res* 33:D192–D196.
48. Lupas A, Van Dyke M, Stock J (1991) Predicting coiled coils from protein sequences. *Science* 252:1162–1164.
49. Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* 17:540–552.
50. Kumar S, Tamura K, Nei M (2004) MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment. *Brief Bioinform* 5:150–163.

## 1.1.- Supplementary material

Supporting figures list:

SI Figure 5

SI Figure 6

SI Figure 7

Supporting tables list:

SI Table 1.1

SI Table 1.2

SI Table 1.3

SI Table 1.4

SI Table 1.5

SI Table 1.6

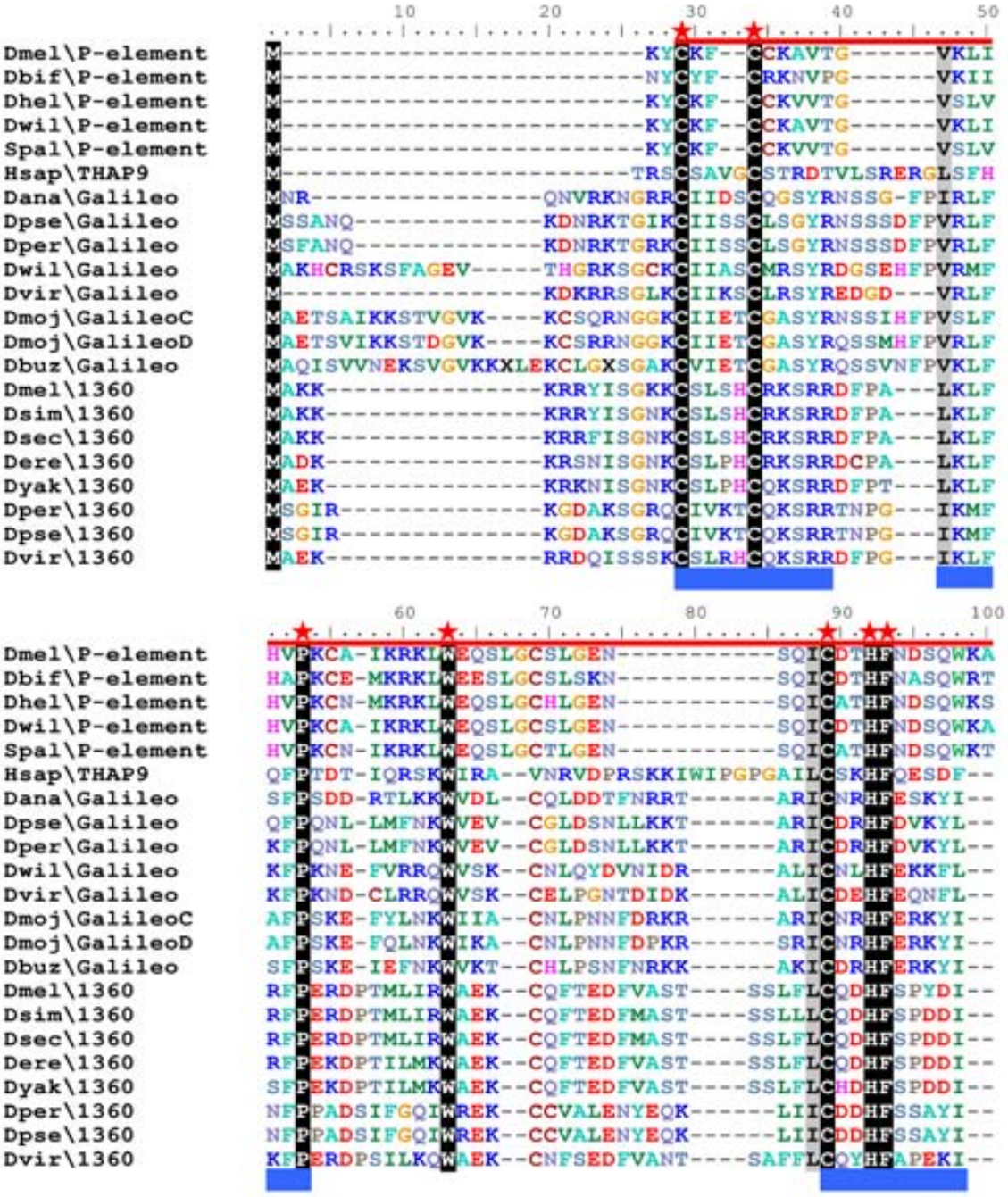
SI Table 1.7

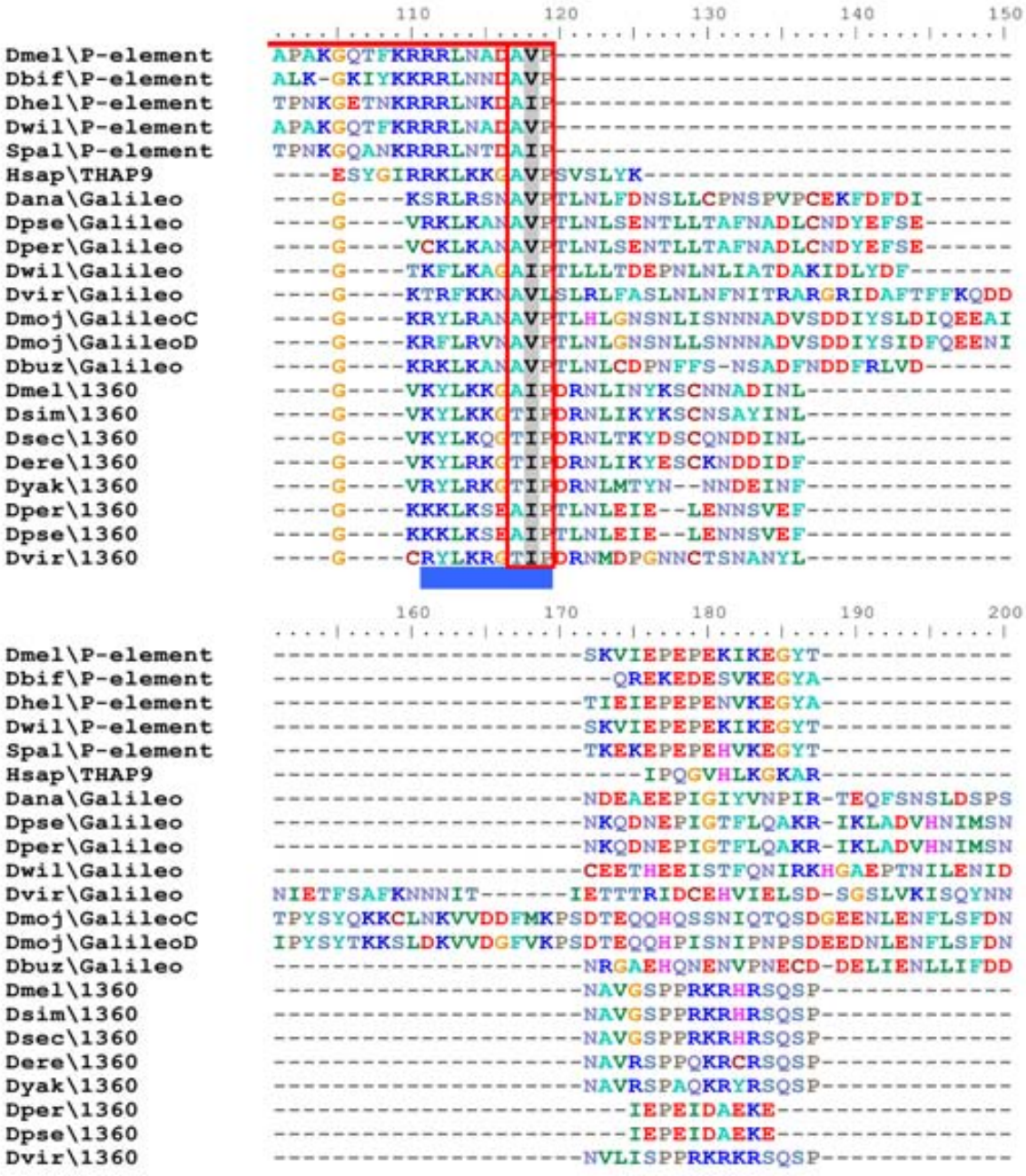
SI Table 1.8

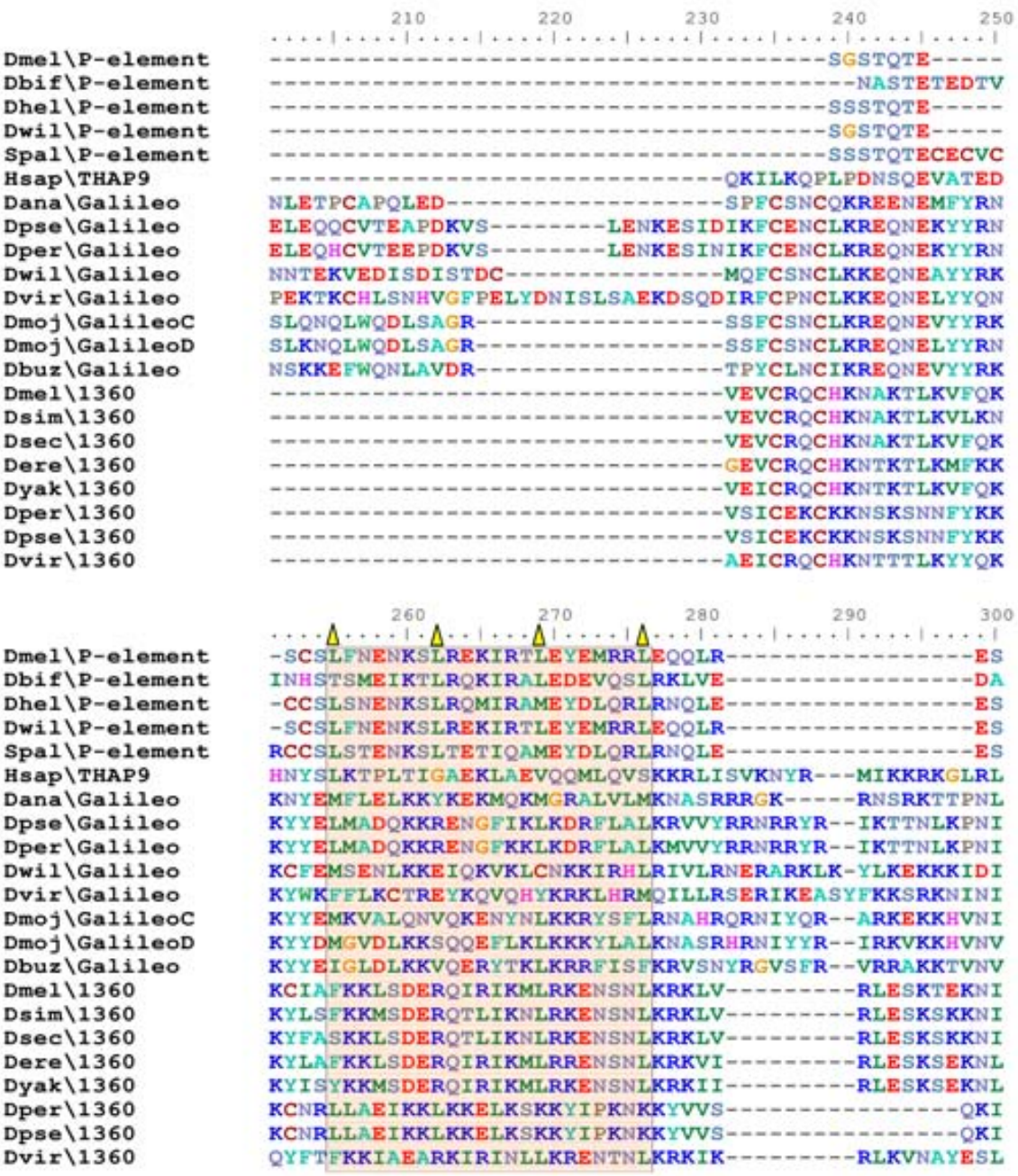
SI Table 1.9

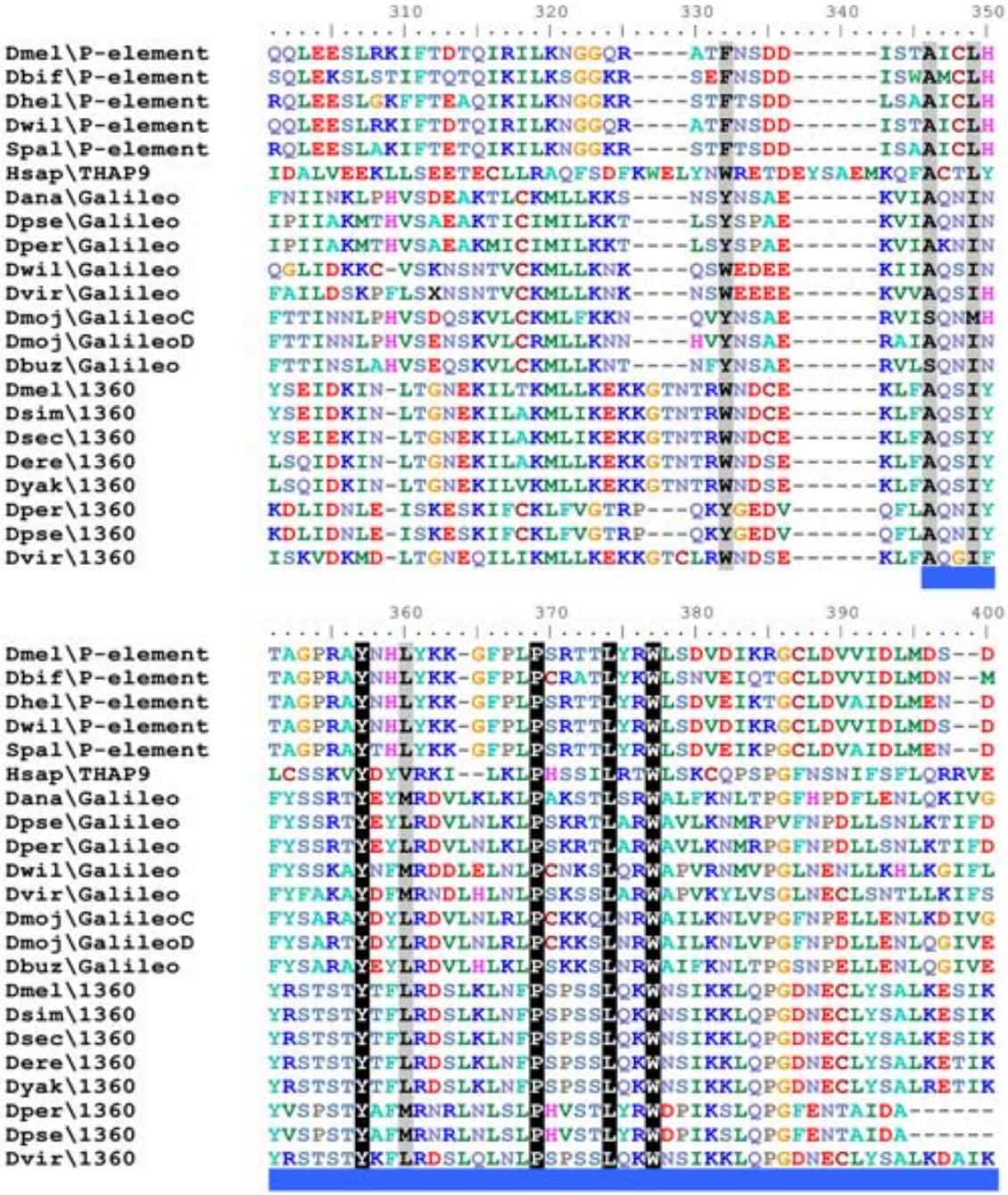

**SI Figure 5.** Neighbor-joining phylogenetic tree built with 14 *Galileo* copies found in the D. virilis genome by using MEGA (16409 is the most-complete copy, see Fig. 1B). The two TIRs of each copy were included in the tree as separate sequences to allow their comparison within and between copies. TIRa is the TIR located at 5' from the TPase or the first TIR that appears in the contig if the copy could not be oriented. The complete deletion option was used leaving 76 informative sites (an almost identical tree results when omitting some of the shortest sequences, increasing the number of informative sites to 258). Bootstrap values of main nodes are shown. Groups A and B show a ~68% divergence indicating ~20 myr of separation. Details of these *Galileo* copies are given in SI Tables 4 and 5.

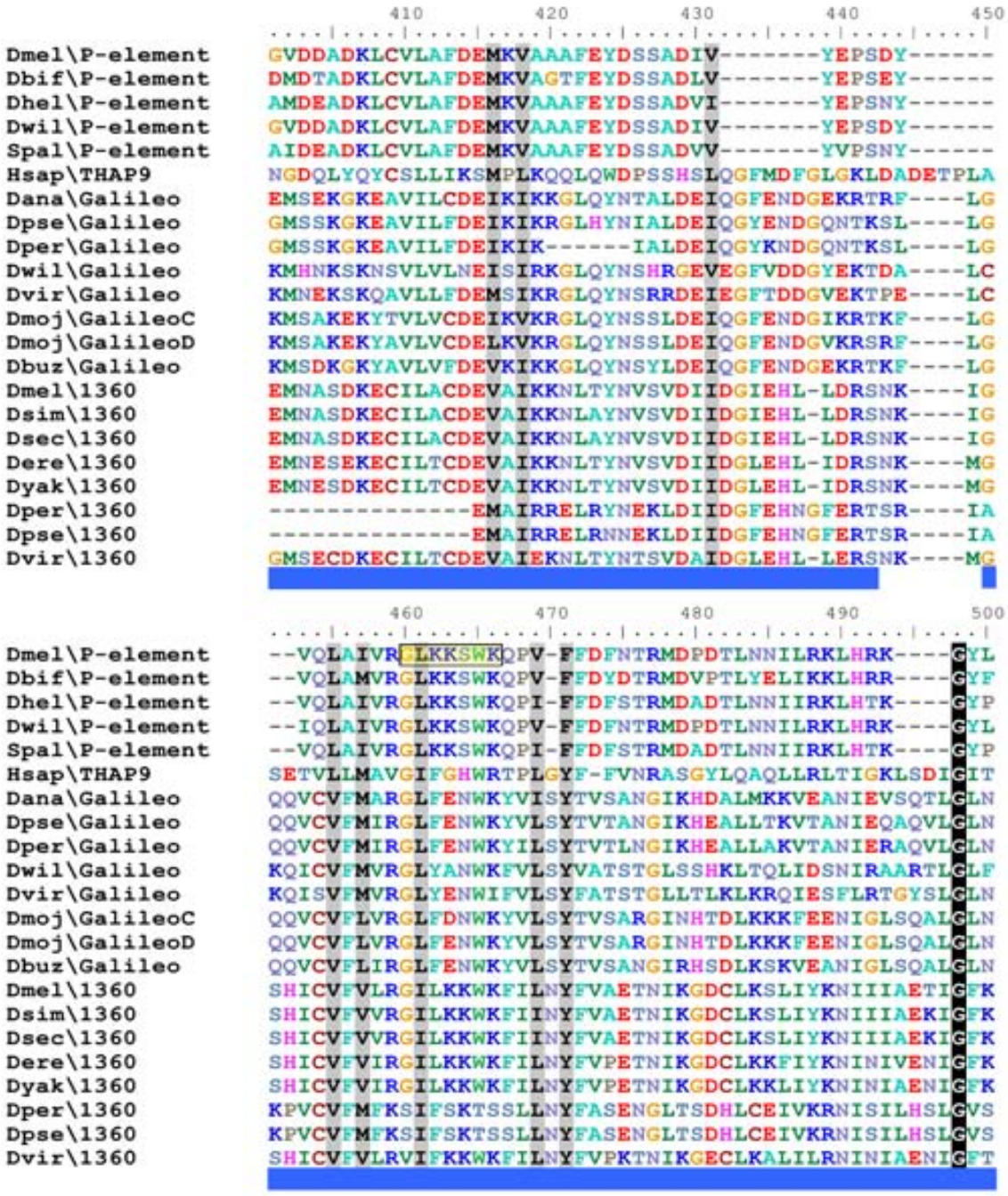
**SI Figure 6.** Multiple alignment of 22 proteins: eight *Galileo* TPases, eigth *1360* TPases, five representative *P-element* TPases and THAP9 protein from *Homo sapiens*. The alignment of the THAP domain region was corrected by hand to align the functional and conserved amino acids of the domain. Conserved blocks selected with Gblocks are marked with a blue box. Identical positions are black-shaded and the positions with similar amino acids are gray-shaded. THAP domain conserved residues are marked with a red star and the three final residues (AVP) are included in a red box. A red line marks the entire THAP domain region. The coiled-coil region is marked with an orange-filled box. The Leucine amino acids of the Leucine zipper coiled-coil motif of the *Dmel\P* TPase are marked with a yellow triangle. GTP binding sites of the *Dmel\P* TPase are marked with a yellow-filled box. The catalytic amino acids are labeled with a green star. The fourth acidic catalytic amino acid of the *P-element* transposase that is not conserved in the TPases of *Galileo* and *1360* is indicated with a gray star. Accession numbers for *P-element* TPases and THAP9 are: *Dmel\P*: Q7M3K2, *Dbif\P*: AAB31526, *Dhel\P*: AAK08181, *Dwil\P*: AAT96022, *Spal\P*: M63341, THAP9: NP_078948.

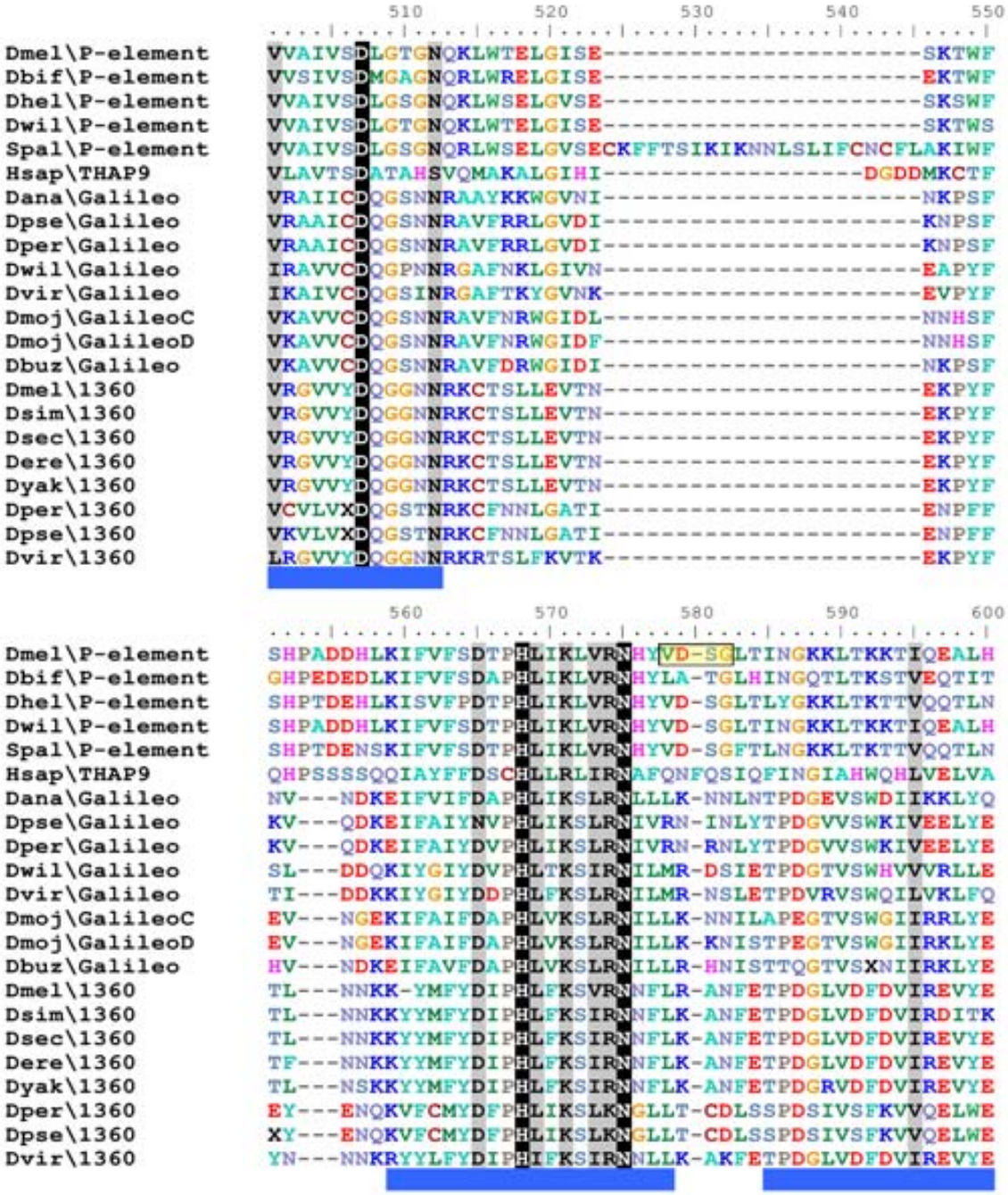

10
20
30
40
50
Dmel\P-element
Dbif\P-element
Dhel\P-element
Dwil\P-element
Spal\P-element
Hsap\THAP9
Dana\Galileo
Dpse\Galileo
Dper\Galileo
Dwil\Galileo
Dvir\Galileo
Dmoj\GalileoC
Dmoj\GalileoD
Dbuz\Galileo
Dmel\1360
Dsim\1360
Dsec\1360
Dere\1360
Dyak\1360
Dper\1360
Dpse\1360
Dvir\1360
60
70
80
90
100
Dmel\P-element
Dbif\P-element
Dhel\P-element
Dwil\P-element
Spal\P-element
Hsap\THAP9
Dana\Galileo
Dpse\Galileo
Dper\Galileo
Dwil\Galileo
Dvir\Galileo
Dmoj\GalileoC
Dmoj\GalileoD
Dbuz\Galileo
Dmel\1360
Dsim\1360
Dsec\1360
Dere\1360
Dyak\1360
Dper\1360
Dpse\1360
Dvir\1360

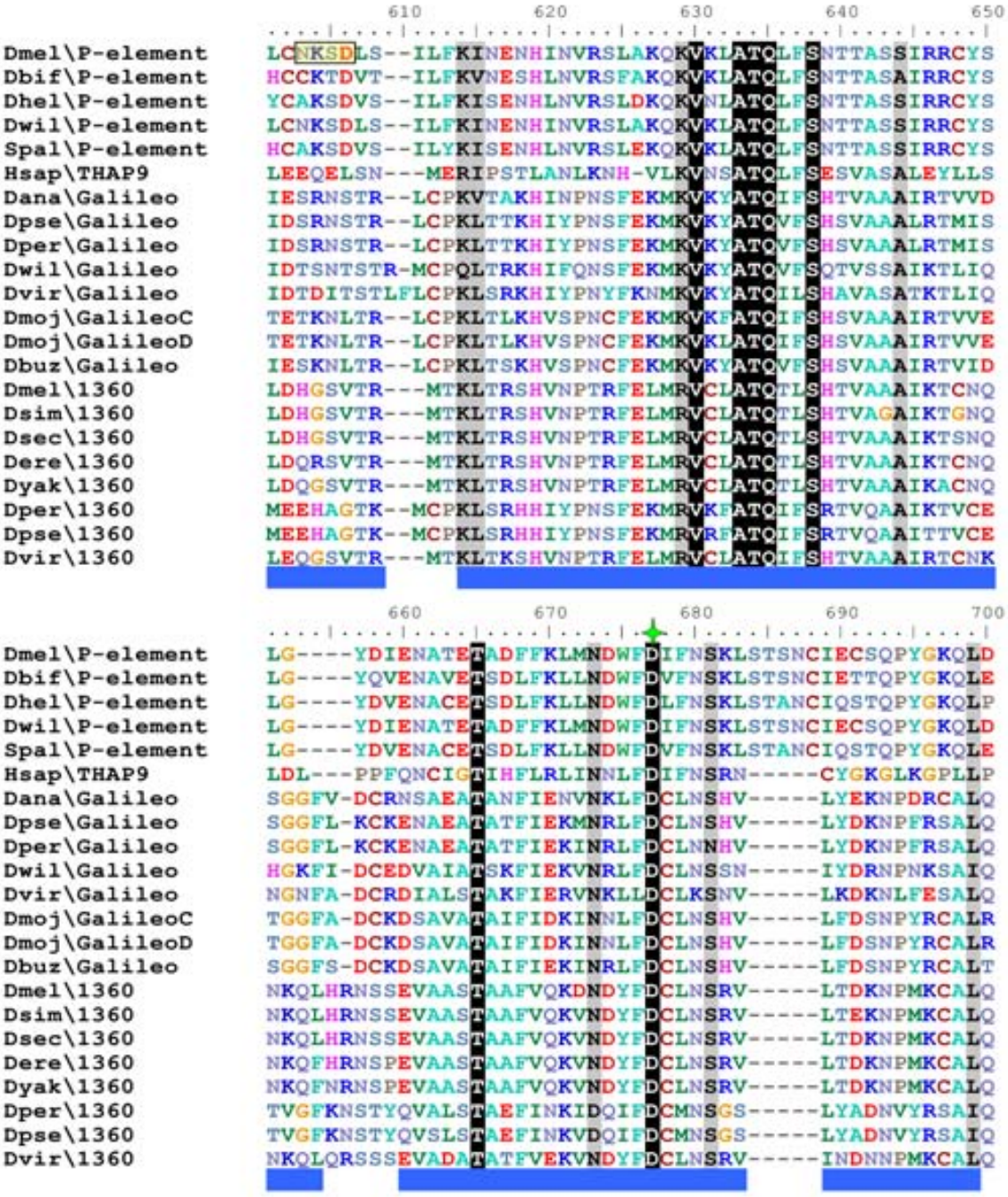

110 120 130 140 150
Dmel\P-element APAKGQTFKRRRLNADAVP----------
Dbif\P-element ALK-GKIYKKRRLNNDAVP----------
Dhel\P-element TPNKGETNKRRRLNKDAIP----------
Dwil\P-element APAKGQTFKRRRLNADAVP----------
Spal\P-element TPNKGQANKRRRLNTDAIP----------
Hsap\THAP9 ----ESYGIRRKLKKGAVPSVSLYK----------
Dana\Galileo ----G----KSRLRSNAVPTLNLFDNSLLCPNSPVPCEKFDFDI------
Dpse\Galileo ----G----VRKLKANAVPTLNLSENTLLTAFNADLCNDYEFSE------
Dper\Galileo ----G----VCKLKANAVPTLNLSENTLLTAFNADLCNDYEFSE------
Dwil\Galileo ----G----TKFLKAGAIPTLLLTDEPNLNLIATDAKIDLYDF-------
Dvir\Galileo ----G----KTRFKKNAVLSLRLFASLNLNFNITRARGRIDAFTFFKQDD
Dmoj\GalileoC ----G----KRYLRANAVPTLHLGNSNLISNNNADVSDDIYSLDIQEEAI
Dmoj\GalileoD ----G----KRFLRVNAVPTLNLGNSNLLSNNNADVSDDIYSIDFQEENI
Dbuz\Galileo ----G----KRKLKANAVPTLNLCDPNFFS-NSADFNDDFRLVD------
Dmel\1360 ----G----VKYLKKGAIPDRNLINYKSCNNADINL----------
Dsim\1360 ----G----VKYLKKGTIPDRNLIKYKSCNSAYINL----------
Dsec\1360 ----G----VKYLKQGTIPDRNLTKYDSCQNDDINL----------
Dere\1360 ----G----VKYLRKGTIPDRNLIKYESCKNDDIDF----------
Dyak\1360 ----G----VRYLRKGTIPDRNLMTYN--NNDEINF----------
Dper\1360 ----G----KKKLKSEAIPTLNLEIE--LENNSVEF----------
Dpse\1360 ----G----KKKLKSEAIPTLNLEIE--LENNSVEF----------
Dvir\1360 ----G----CRYLKRGTIPDRNMDPGNNCTSNANYL----------
160 170 180 190 200
Dmel\P-element ----------SKVIEPEPEKIKEGYT----------
Dbif\P-element ----------QREKEDESVKEGYA----------
Dhel\P-element ----------TIEIEPEPENVKEGYA----------
Dwil\P-element ----------SKVIEPEPEKIKEGYT----------
Spal\P-element ----------TKEKEPEPEHVKEGYT----------
Hsap\THAP9 ----------IPQGVHLKGKAR----------
Dana\Galileo ----------NDEAEEPIGIYVNPIR-TEQFSNSLDSPS
Dpse\Galileo ----------NKQDNEPIGTFLQAKR-IKLADVHNIMSN
Dper\Galileo ----------NKQDNEPIGTFLQAKR-IKLADVHNIMSN
Dwil\Galileo ----------CEETHEEISTFQNIRKHGAEPTNILENID
Dvir\Galileo NIETFSAFKNNNIT------IETTTRIDCEHVIELSD-SGSLVKISQYNN
Dmoj\GalileoC TPYSYQKKCLNKVVDDFMKPSDTEQQHQSSNIQTQSDGEENLENFLSFDN
Dmoj\GalileoD IPYSYTKKSLDKVVDGFVKPSDTEQQHPISNIPNPSDEEDNLENFLSFDN
Dbuz\Galileo ----------NRGAEHQNENVPNECD-DELIENLLIFDD
Dmel\1360 ----------NAVGSPPRKRHRSQSP----------
Dsim\1360 ----------NAVGSPPRKRHRSQSP----------
Dsec\1360 ----------NAVGSPPRKRHRSQSP----------
Dere\1360 ----------NAVRSPPQKRCRSQSP----------
Dyak\1360 ----------NAVRSPAQKRYRSQSP----------
Dper\1360 ----------IEPEIDAEKE----------
Dpse\1360 ----------IEPEIDAEKE----------
Dvir\1360 ----------NVLISPPRKRKRSQSP----------

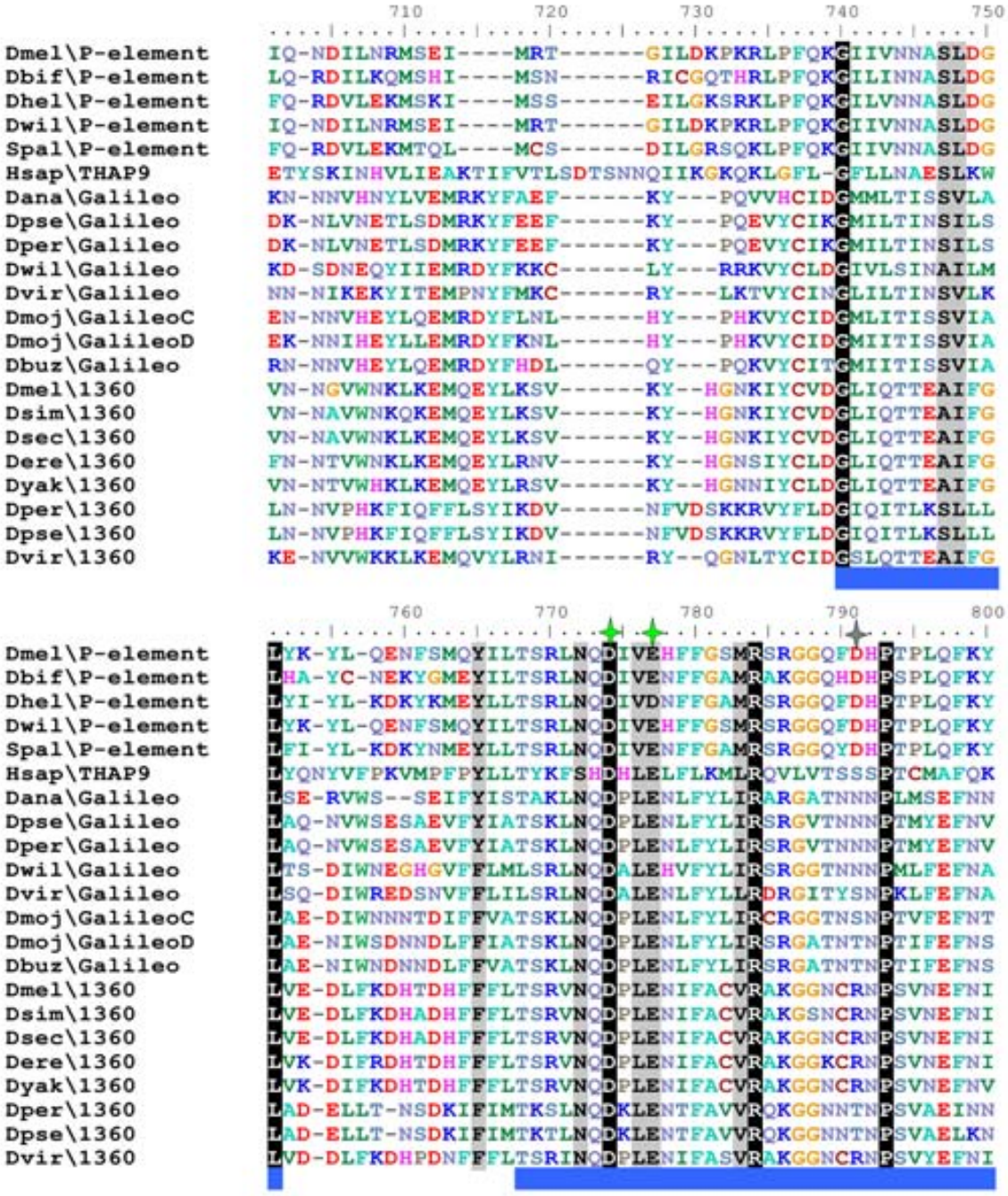

210 220 230 240 250
Dmel\P-element -----------SGSTQTE-----
Dbif\P-element -----------NASTETEDTV
Dhel\P-element -----------SSSTQTE-----
Dwil\P-element -----------SGSTQTE-----
Spal\P-element -----------SSSTQTECECVC
Hsap\THAP9 -----------QKILKQPLPDNSQEVATED
Dana\Galileo NLETPCAPQLED----------SPFCSNCQKREENEMFYRN
Dpse\Galileo ELEQQCVTEAPDKVS--------LENKESIDIKFCENCLKREQNEKYYRN
Dper\Galileo ELEQHCVTEEPDKVS--------LENKESINIKFCENCLKREQNEKYYRN
Dwil\Galileo NNTEKVEDISDISTDC----------MQFCSNCLKKEQNEAYYRK
Dvir\Galileo PEKTKCHLSNHVGFPELYDNISLSAEKDSQDIRFCPNCLKKEQNELYYQN
DmojGalileoC SLQNQLWQDLSAGR----------SSFCSNCLKREQNEVYYRK
DmojGalileoD SLKNQLWQDLSAGR----------SSFCSNCLKREQNELYYRN
Dbuz\Galileo NSKKEFWQNLAVDR----------TPYCLNCIKREQNEVYYRK
Dmel\1360 -----------VEVCRQCHKNAKTLKVFQK
Dsim\1360 -----------VEVCRQCHKNAKTLKVLKN
Dsec\1360 -----------VEVCRQCHKNAKTLKVFQK
Dere\1360 -----------GEVCRQCHKNTKTLKMFKK
Dyak\1360 -----------VEICRQCHKNTKTLKVFQK
Dper\1360 -----------VSICEKCKKNSKSNNFYKK
Dpse\1360 -----------VSICEKCKKNSKSNNFYKK
Dvir\1360 -----------AEICRQCHKNTTTLKYYQK
260 270 280 290 300
Dmel\P-element -SCSLFNENKSLREKIRTLEYEMRRLEQQLR----------ES
Dbif\P-element INHSTSMEIKTLRQKIRALEDEVQSLRKLVE----------DA
Dhel\P-element -CCSLSNENKSLRQMIRAMEYDLQRLRNQLE----------ES
Dwil\P-element -SCSLFNENKSLREKIRTLEYEMRRLEQQLR----------ES
Spal\P-element RCCSLSTENKSLTETIQAMEYDLQRLRNQLE----------ES
Hsap\THAP9 HNYSLKTPLTIGAEKLAEVQQMLQVSKKRLISVKNYR---MIKKRKGLRL
Dana\Galileo KNYEMFLELKKYKEKMQKMGRALVLMKNASRRRGK-----RNSRKTTPNL
Dpse\Galileo KYYELMADQKKRENGFIKLKDRFLALKRVVYRRNRRYR--IKTTNLKPNI
Dper\Galileo KYYELMADQKKRENGFKKLKDRFLALKMVVYRRNRRYR--IKTTNLKPNI
Dwil\Galileo KCFEMSENLKKEIQKVKLCNKKIRHLRIVLRNERARKLK-YLKEKKKIDI
Dvir\Galileo KYWKFFLKCTREYKQVQHYKRKLHRMQILLRSERIKEASYFKKSRKNINI
DmojGalileoC KYYEMKVALQNVQKENYNLKKRYSFLRNAHRQRNIYQR--ARKEKKHVNI
DmojGalileoD KYYDMGVDLKKSQQEFLKLKKKYLALKNASRHRNIYYR--IRKVKKHVNV
Dbuz\Galileo KYYEIGLDLKKVQERYTKLKRRFISFKRVSNYRGVSFR--VRRAKKTVNV
Dmel\1360 KCIAFKKLSDERQIRIKMLRKENSNLKRKLV---------RLESKTEKNI
Dsim\1360 KYLSFKKMSDERQTLIKNLRKENSNLKRKLV---------RLESKSKKNI
Dsec\1360 KYFASKKLSDERQTLIKNLRKENSNLKRKLV---------RLESKSKKNI
Dere\1360 KYLAFKKLSDERQIRIKMLRRENSNLKRKVI---------RLESKSEKNL
Dyak\1360 KYISYKKMSDERQIRIKMLRKENSNLKRKII---------RLESKSEKNL
Dper\1360 KCNRLLAEIKKLKKELKSKKYIPKNKKYVVS----------QKI
Dpse\1360 KCNRLLAEIKKLKKELKSKKYIPKNKKYVVS----------QKI
Dvir\1360 QYFTFKKIAEARKIRINLLKRENTNLKRKIK---------RLKVNAYESL

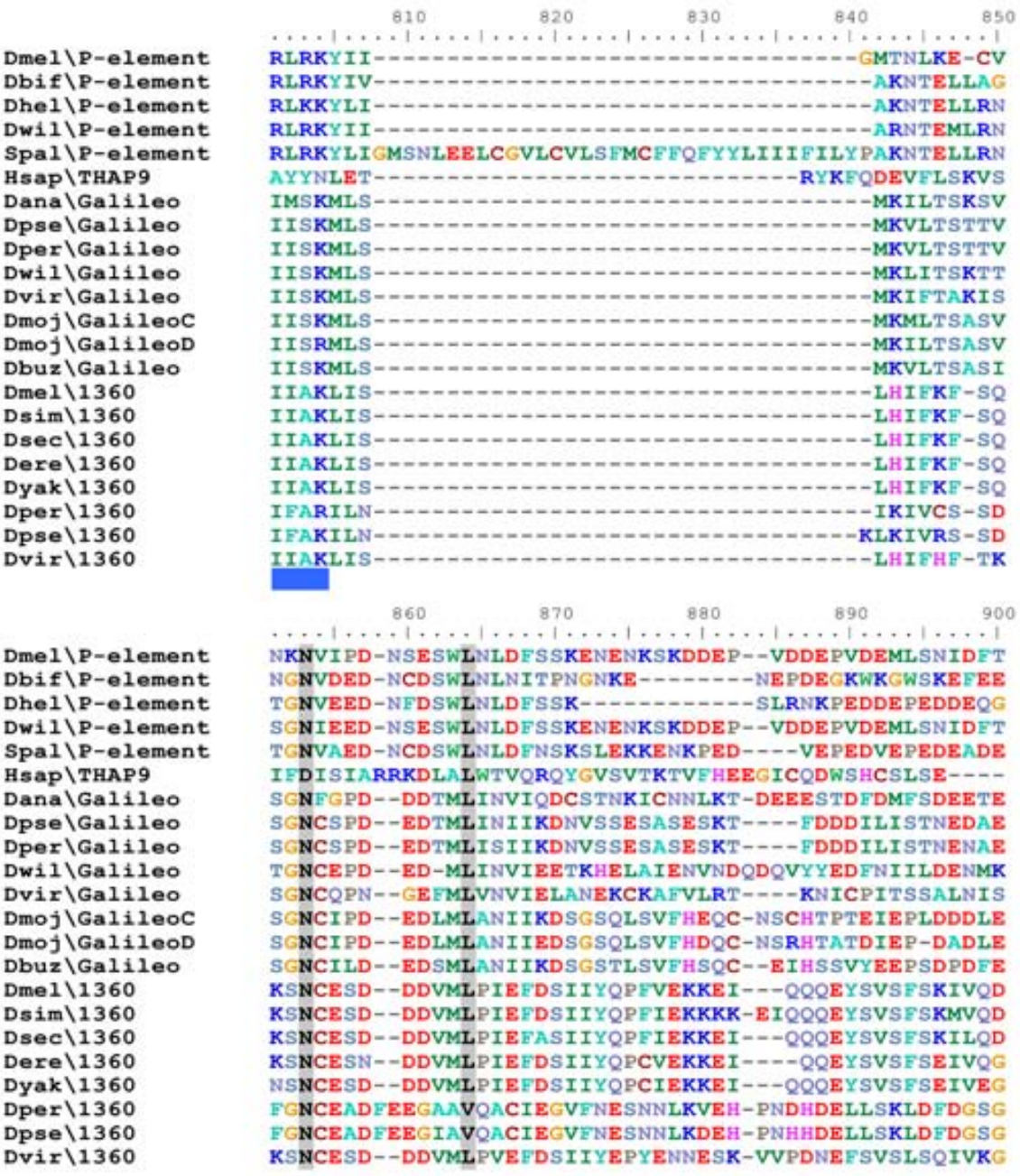

310 320 330 340 350
Dmel\P-element QQLEESLRKIFTDTQIRILKNGGQR----ATFNSDD------ISTAICLH
Dbif\P-element SQLEKSLSTIFTQTQIKILKSGGKR----SEFNSDD------ISWAMCLH
Dhel\P-element RQLEESLGKFFTEAQIKILKNGGKR----STFTSDD------LSAAICLH
Dwil\P-element QQLEESLRKIFTDTQIRILKNGGQR----ATFNSDD------ISTAICLH
Spal\P-element RQLEESLAKIFTETQIKILKNGGKR----STFTSDD------ISAAICLH
Hsap\THAP9 IDALVEEKLLSEETECLLRAQFSDFKWELYNWRETDEYSAEMKQFACTLY
Dana\Galileo FNIINKLPHVSDEAKTLCKMLLKKS----NSYNSAE------KVIAQNIN
Dpse\Galileo IPIIAKMTHVSAEAKTICIMILKKT----LSYSPAE------KVIAQNIN
Dper\Galileo IPIIAKMTHVSAEAKMICIMILKKT----LSYSPAE------KVIAKNIN
Dwil\Galileo QGLIDKKC-VSKNSNTVCKMLLKNK----QSWEDEE------KIIAQSIN
Dvir\Galileo FAILDSKPFLSXNSNTVCKMLLKNK----NSWEEEE------KVVAQSIH
DmojGalileoC FTTINNLPHVSDQSKVLCKMLFKKN----QVYNSAE------RVISQNMH
DmojGalileoD FTTINNLPHVSENSKVLCRMLLKNN----HVYNSAE------RAIAQNIN
Dbuz\Galileo FTTINSLAHVSEQSKVLCKMLLKNT----NFYNSAE------RVLSQNIN
Dmel\1360 YSEIDKIN-LTGNEKILTKMLLKEKKGTNTRWNDCE------KLFAQSIY
Dsim\1360 YSEIDKIN-LTGNEKILAKMLIKEKKGTNTRWNDCE------KLFAQSIY
Dsec\1360 YSEIEKIN-LTGNEKILAKMLIKEKKGTNTRWNDCE------KLFAQSIY
Dere\1360 LSQIDKIN-LTGNEKILAKMLLKEKKGTNTRWNDSE------KLFAQSIY
Dyak\1360 LSQIDKIN-LTGNEKILVKMLLKEKKGTNTRWNDSE------KLFAQSIY
Dper\1360 KDLIDNLE-ISKESKIFCKLFVGTRP---QKYGEDV------QFLAQNIY
Dpse\1360 KDLIDNLE-ISKESKIFCKLFVGTRP---QKYGEDV------QFLAQNIY
Dvir\1360 ISKVDKMD-LTGNEQILIKMLLKEKKGTCLRWNDSE------KLFAQGIF
360 370 380 390 400
Dmel\P-element TAGPRAYNHLYKK-GFPLPSRTTLYRWLSDVDIKRGCLDVVIDLMDS--D
Dbif\P-element TAGPRAYNHLYKK-GFPLPCRATLYKWLSNVEIQTGCLDVVIDLMDN--M
Dhel\P-element TAGPRAYNHLYKK-GFPLPSRTTLYRWLSDVEIKTGCLDVAIDLMEN--D
Dwil\P-element TAGPRAYNHLYKK-GFPLPSRTTLYRWLSDVDIKRGCLDVVIDLMDS--D
Spal\P-element TAGPRAYTHLYKK-GFPLPSRTTLYRWLSDVEIKPGCLDVAIDLMEN--D
Hsap\THAP9 LCSSKVYDYVRKI--LKLPHSSIERTWLSKCQPSPGFNSNIFSFLQRRVE
Dana\Galileo FYSSRTYEYMRDVLKLKLPAKSTLSRWALFKNLTPGFHPDFLENLQKIVG
Dpse\Galileo FYSSRTYEYLRDVLNLKLPSKRTLARWAVLKNMRPVFNPDLLSNLKTIFD
Dper\Galileo FYSSRTYEYLRDVLNLKLPSKRTLARWAVLKNMRPGFNPDLLSNLKTIFD
Dwil\Galileo FYSSKAYNFMRDDLELNLPCNKSLQRWAPVRNMVPGLNENLLKHLKGIFL
Dvir\Galileo FYFAKAYDFMRNDLHLNLPSKSSLARWAPVKYLVSGLNECLSNTLLKIFS
DmojGalileoC FYSARAYDYLRDVLNLRLPCKKQLNRWAILKNLVPGFNPELLENLKDIVG
DmojGalileoD FYSARTYDYLRDVLNLRLPCKKSLNRWAILKNLVPGFNPDLLENLQGIVE
Dbuz\Galileo FYSARAYEYLRDVLHLKLPSKKSLNRWAIFKNLTPGSNPELLENLQGIVE
Dmel\1360 YRSTSTYTFLRDSLKLNFPSPSSLQKWNSIKKLQPGDNECLYSALKESIK
Dsim\1360 YRSTSTYTFLRDSLKLNFPSPSSLQKWNSIKKLQPGDNECLYSALKESIK
Dsec\1360 YRSTSTYTFLRDSLKLNFPSPSSLQKWNSIKKLQPGDNECLYSALKESIK
Dere\1360 YRSTSTYTFLRDSLKLNFPSPSSLQKWNSIKKLQPGDNECLYSALKETIK
Dyak\1360 YRSTSTYTFLRDSLKLNFPSPSSLQKWNSIKKLQPGDNECLYSALRETIK
Dper\1360 YVSPSTYAFMRNRLNLSLPHVSTLYRWDPIKSLQPGFENTAIDA------
Dpse\1360 YVSPSTYAFMRNRLNLSLPHVSTLYRWDPIKSLQPGFENTAIDA------
Dvir\1360 YRSTSTYKFLRDSLQLNLPSPSSLQKWNSIKKLQPGDNECLYSALKDAIK

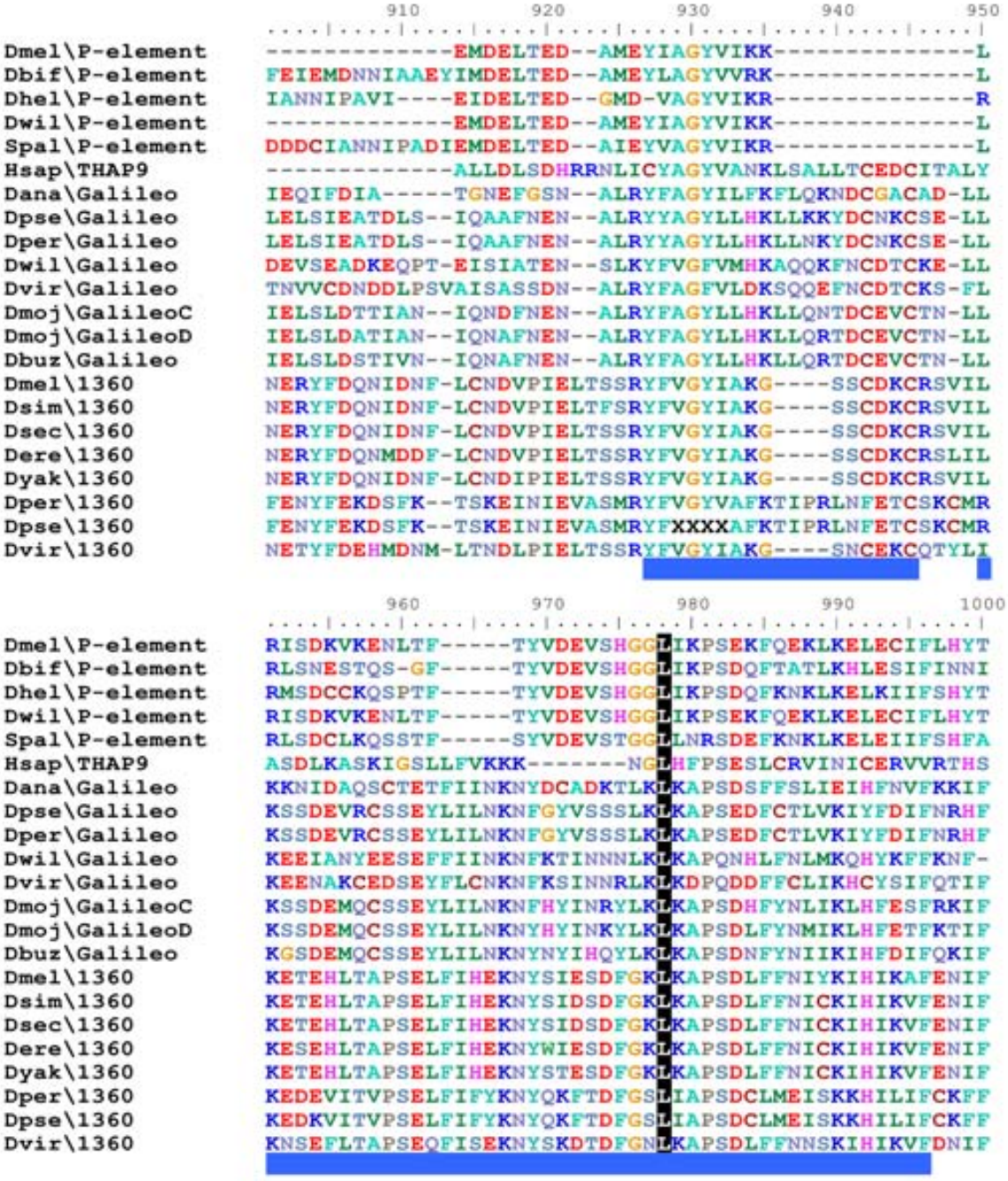

410 420 430 440 450
Dmel\P-element GVDDADKLCVLAFDEMKVAAAFEYDSSADIV-------YEPSDY------
Dbif\P-element DMDTADKLCVLAFDEMKVAGTFEYDSSADLV-------YEPSEY------
Dhel\P-element AMDEADKLCVLAFDEMKVAAAFEYDSSADVI-------YEPSNY------
Dwil\P-element GVDDADKLCVLAFDEMKVAAAFEYDSSADIV-------YEPSDY------
Spal\P-element AIDEADKLCVLAFDEMKVAAAFEYDSSADVV-------YVPSNY------
Hsap\THAP9 NGDQLYQYCSLLIKSMPLKQQLQWDPSSHSLQGFMDFGLGKLDADETPLA
Dana\Galileo EMSEKGKEAVILCDEIKIKKGLQYNTALDEIQGFENDGEKRTRF----LG
Dpse\Galileo GMSSKGKEAVILFDEIKIKRGLHYNIALDEIQGYENDGQNTKSL----LG
Dper\Galileo GMSSKGKEAVILFDEIKIK------IALDEIQGYKNDGQNTKSL----LG
Dwil\Galileo KMHNKSKNSVLVLNEISIRKGLQYNSHRGEVEGFVDDGYEKTDA----LC
Dvir\Galileo KMNEKSKQAVLLFDEMSIKRGLQYNSRRDEIEGFTDDGVEKTPE----LC
Dmoj\GalileoC KMSAKEKYTVLVCDEIKVKRGLQYNSSLDEIQGFENDGIKRTKF----LG
Dmoj\GalileoD KMSAKEKYAVLVCDELKVKRGLQYNSSLDEIQGFENDGVKRSRF----LG
Dbuz\Galileo KMSDKGKYAVLVFDEVKIKKGLQYNSYLDEIQGFENDGEKRTKF----LG
Dmel\1360 EMNASDKECILACDEVAIKKNLTYNVSVDIIDGIEHL-LDRSNK----IG
Dsim\1360 EMNASDKECILACDEVAIKKNLAYNVSVDIIDGIEHL-LDRSNK----IG
Dsec\1360 EMNASDKECILACDEVAIKKNLAYNVSVDIIDGIEHL-LDRSNK----IG
Dere\1360 EMNESEKECILTCDEVAIKKNLTYNVSVDIIDGLEHL-IDRSNK----MG
Dyak\1360 EMNESDKECILTCDEVAIKKNLTYNVSVDIIDGLEHL-IDRSNK----MG
Dper\1360 --------------EMAIRRELRYNEKLDIIDGFEHNGFERTSR----IA
Dpse\1360 --------------EMAIRRELRNNEKLDIIDGFEHNGFERTSR----IA
Dvir\1360 GMSECDKECILTCDEVAIEKNLTYNTSVDAIDGLEHL-LERSNK----MG
460 470 480 490 500
Dmel\P-element --VQLAIVRGLKKSWKQPV-FFDFNTRMDPDTLNNILRKLHRK----GYL
Dbif\P-element --VQLAMVRGLKKSWKQPV-FFDYDTRMDVPTLYELIKKLHRR----GYF
Dhel\P-element --VQLAIVRGLKKSWKQPI-FFDFSTRMDADTLNNIIRKLHTK----GYP
Dwil\P-element --IQLAIVRGLKKSWKQPV-FFDFNTRMDPDTLNNILRKLHRK----GYL
Spal\P-element --VQLAIVRGLKKSWKQPI-FFDFSTRMDADTLNNIIRKLHTK----GYP
Hsap\THAP9 SETVLLMAVGIFGHWRTPLGYF-FVNRASGYLQAQLLRLTIGKLSDIGIT
Dana\Galileo QQVCVFMARGLFENWKYVISYTVSANGIKHDALMKKVEANIEVSQTLGLN
Dpse\Galileo QQVCVFMIRGLFENWKYVLSYTVTANGIKHEALLTKVTANIEQAQVLGLN
Dper\Galileo QQVCVFMIRGLFENWKYILSYTVTLNGIKHEALLAKVTANIERAQVLGLN
Dwil\Galileo KQICVFMVRGLYANWKFVLSYVATSTGLSSHKLTQLIDSNIRAARTLGLF
Dvir\Galileo KQISVFMVRGLYENWIFVLSYFATSTGLLTLKLKRQIESFLRTGYSLGLN
Dmoj\GalileoC QQVCVFLVRGLFDNWKYVLSYTVSARGINHTDLKKKFEENIGLSQALGLN
Dmoj\GalileoD QQVCVFLVRGLFENWKYVLSYTVSARGINHTDLKKKFEENIGLSQALGLN
Dbuz\Galileo QQVCVFLIRGLFENWKYVLSYTVSANGIRHSDLKSKVEANIGLSQALGLN
Dmel\1360 SHICVFVLRGILKKWKFILNYFVAETNIKGDCLKSLIYKNIIIAETIGFK
Dsim\1360 SHICVFVVRGILKKWKFIINYFVAETNIKGDCLKSLIYKNIIIAEKIGFK
Dsec\1360 SHICVFVVRGILKKWKFIINYFVAETNIKGDCLKSLIYKNIIIAEKIGFK
Dere\1360 SHICVFVIRGILKKWKFILNYFVPETNIKGDCLKKFIYKNINIVENIGFK
Dyak\1360 SHICVFVIRGILKKWKFILNYFVPETNIKGDCLKKLIYKNINIAENIGFK
Dper\1360 KPVCVFMFKSIFSKTSSLLNYFASENGLTSDHLCEIVKRNISILHSLGVS
Dpse\1360 KPVCVFMFKSIFSKTSSLLNYFASENGLTSDHLCEIVKRNISILHSLGVS
Dvir\1360 SHICVFVLRVIFKKWKFILNYFVPKTNIKGECLKALILRNINIAENIGFT

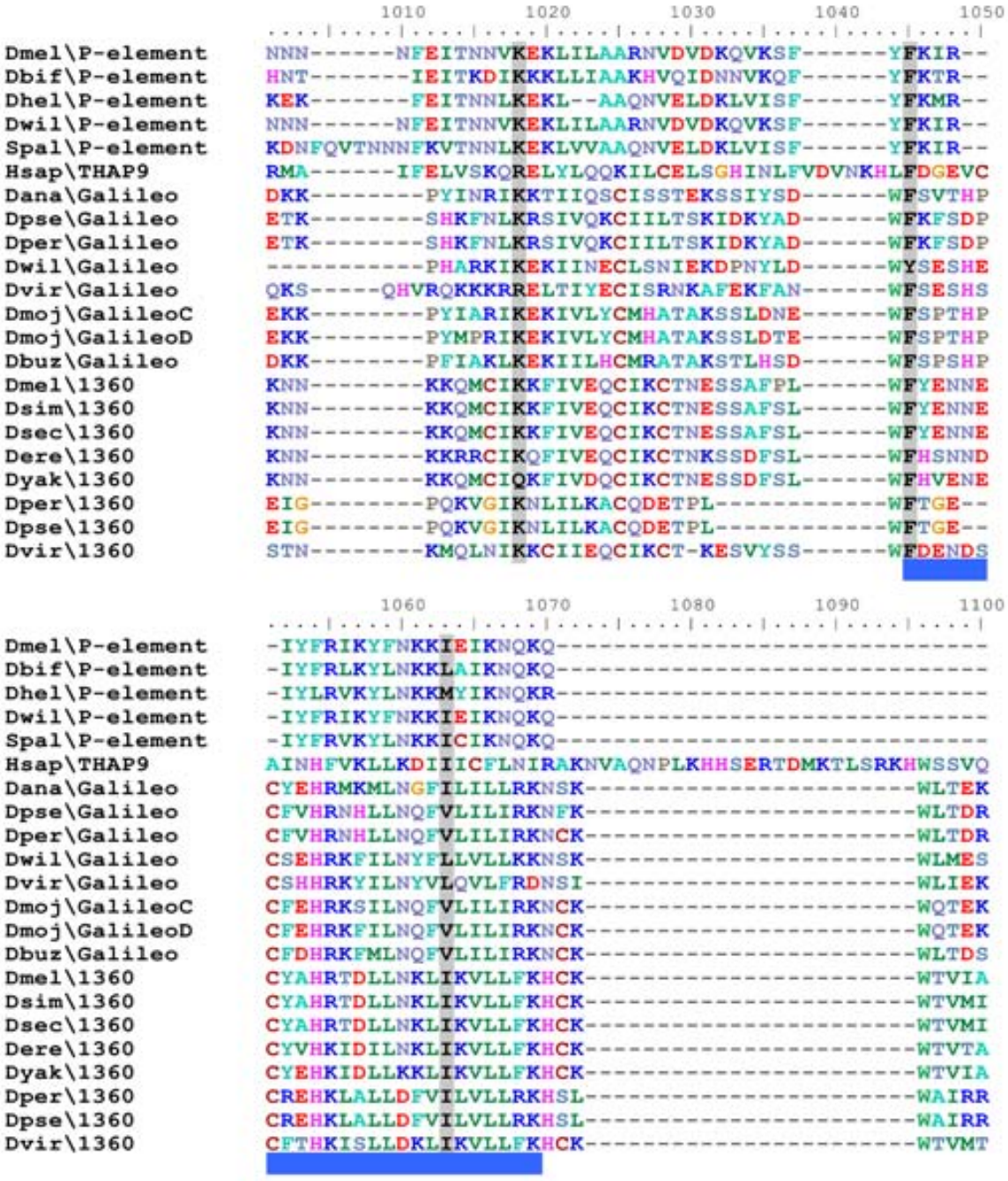

510 520 530 540 550
Dmel\P-element VVAIVSDLGTGNQKLWTELGISE----------SKTWF
Dbif\P-element VVSIVSDMGAGNQRLWRELGISE----------EKTWF
Dhel\P-element VVAIVSDLGSGNQKLWSELGVSE----------SKSWF
Dwil\P-element VVAIVSDLGTGNQKLWTELGISE----------SKTWS
Spal\P-element VVAIVSDLGSGNQRLWSELGVSECKFFTSIKIKNNLSLIFCNCFLAKIWF
Hsap\THAP9 VLAVTSDATAHSVQMAKALGIHI----------DGDDMKCTF
Dana\Galileo VRAIICDQGSNNRAAYKKWGVNI----------NKPSF
Dpse\Galileo VRAAICDQGSNNRAVFRRLGVDI----------KNPSF
Dper\Galileo VRAAICDQGSNNRAVFRRLGVDI----------KNPSF
Dwil\Galileo IRAVVCDQGPNNRGAFNKLGIVN----------EAPYF
Dvir\Galileo IKAIVCDQGSINRGAFTKYGVNK----------EVPYF
DmojGalileoC VKAVVCDQGSNNRAVFNRWGIDL----------NNHSF
DmojGalileoD VKAVVCDQGSNNRAVFNRWGIDF----------NNHSF
Dbuz\Galileo VKAVVCDQGSNNRAVFDRWGIDI----------NKPSF
Dmel\1360 VRGVVYDQGGNNRKCTSLLEVTN----------EKPYF
Dsim\1360 VRGVVYDQGGNNRKCTSLLEVTN----------EKPYF
Dsec\1360 VRGVVYDQGGNNRKCTSLLEVTN----------EKPYF
Dere\1360 VRGVVYDQGGNNRKCTSLLEVTN----------EKPYF
Dyak\1360 VRGVVYDQGGNNRKCTSLLEVTN----------EKPYF
Dper\1360 VCVLVXDQGSTNRKCFNNLGATI----------ENPFF
Dpse\1360 VKVLVXDQGSTNRKCFNNLGATI----------ENPFF
Dvir\1360 LRGVVYDQGGNNRKRTSLFKVTK----------EKPYF
560 570 580 590 600
Dmel\P-element SHPADDHLKIFVFSDTPHLIKLVRNHYVD-SGLTINGKKLTKKTIQEALH
Dbif\P-element GHPEDEDLKIFVFSDAPHLIKLVRNHYLA-TGLHINGQTLTKSTVEQTIT
Dhel\P-element SHPTDEHLKISVFPDTPHLIKLVRNHYVD-SGLTLYGKKLTKTTVQQTLN
Dwil\P-element SHPADDHLKIFVFSDTPHLIKLVRNHYVD-SGLTINGKKLTKKTIQEALH
Spal\P-element SHPTDENSKIFVFSDTPHLIKLVRNHYVD-SGFTLNGKKLTKTTVQQTLN
Hsap\THAP9 QHPSSSSQQIAYFFDSCHLLRLIRNAFQNFQSIQFINGIAHWQHLVELVA
Dana\Galileo NV---NDKEIFVIFDAPHLIKSLRNLLLK-NNLNTPDGEVSWDIIKKLYQ
Dpse\Galileo KV---QDKEIFAIYNVPHLIKSLRNIVRN-INLYTPDGVVSWKIVEELYE
Dper\Galileo KV---QDKEIFAIYDVPHLIKSLRNIVRN-RNLYTPDGVVSWKIVEELYE
Dwil\Galileo SL---DDQKIYGIYDVPHLTKSIRNILMR-DSIETPDGTVSWHVVVRLLE
Dvir\Galileo TI---DDKKIYGIYDDPHLFKSLRNILMR-NSLETPDVRVSWQILVKLFQ
DmojGalileoC EV---NGEKIFAIFDAPHLVKSLRNILLK-NNILAPEGTVSWGIIRRLYE
DmojGalileoD EV---NGEKIFAIFDAPHLVKSLRNILLK-KNISTPEGTVSWGIIRKLYE
Dbuz\Galileo HV---NDKEIFAVFDAPHLVKSLRNILLR-HNISTTQGTVSXNIIRKLYE
Dmel\1360 TL---NNKK-YMFYDIPHLFKSVRNNFLR-ANFETPDGLVDFDVIREVYE
Dsim\1360 TL---NNKKYYMFYDIPHLFKSIRNNFLK-ANFETPDGLVDFDVIRDITK
Dsec\1360 TL---NNKKYYMFYDIPHLFKSIRNNFLK-ANFETPDGLVDFDVIREVYE
Dere\1360 TF---NNKKYYMFYDIPHLFKSIRNNFLK-ANFETPDGLVDFDVIREVYE
Dyak\1360 TL---NSKKYYMFYDIPHLFKSIRNNFLK-ANFETPDGRVDFDVIREVYE
Dper\1360 EY---ENQKVFCMYDFPHLIKSLKNGLLT-CDLSSPDSIVSFKVVQELWE
Dpse\1360 XY---ENQKVFCMYDFPHLIKSLKNGLLT-CDLSSPDSIVSFKVVQELWE
Dvir\1360 YN---NNKRYYLFYDIPHIFKSIRNNLLK-AKFETPDGLVDFDVIREVYE

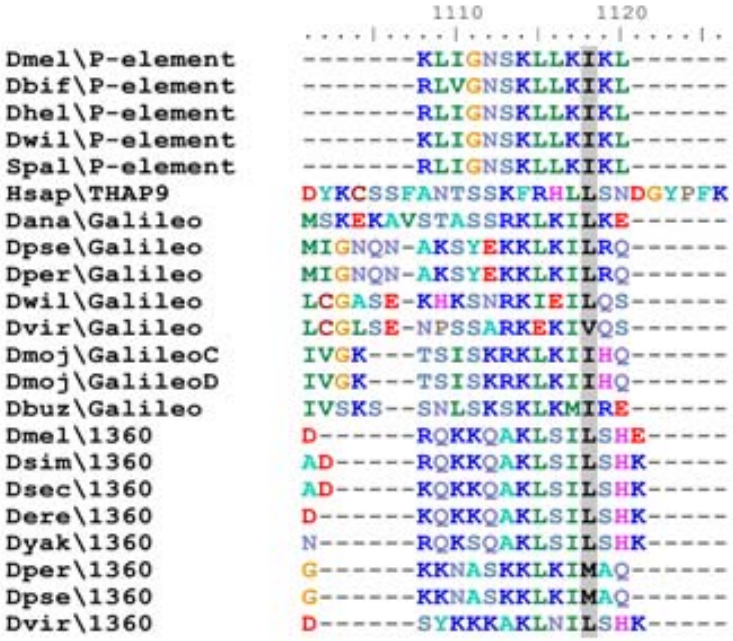

610 620 630 640 650
Dmel\P-element LCNKSDLS--ILFKINENHINVRSLAKQKVKLATQLFSNTTASSIRRCYS
Dbif\P-element HCCKTDVT--ILFKVNESHLNVRSFAKQKVKLATQLFSNTTASAIRRCYS
Dhel\P-element YCAKSDVS--ILFKISENHLNVRSLDKQKVNLATQLFSNTTASSIRRCYS
Dwil\P-element LCNKSDLS--ILFKINENHINVRSLAKQKVKLATQLFSNTTASSIRRCYS
Spal\P-element HCAKSDVS--ILYKISENHLNVRSLEKQKVKLATQLFSNTTASSIRRCYS
Hsap\THAP9 LEEQELSN---MERIPSTLANLKNH-VLKVNSATQLFSESVASALEYLLS
Dana\Galileo IESRNSTR--LCPKVTAKHINPNSFEKMKVKYATQIFSHTVAAAIRTVVD
Dpse\Galileo IDSRNSTR--LCPKLTTKHIYPNSFEKMKVKYATQVFSHSVAAALRTMIS
Dper\Galileo IDSRNSTR--LCPKLTTKHIYPNSFEKMKVKYATQVFSHSVAAALRTMIS
Dwil\Galileo IDTSNTSTR-MCPQLTRKHIFQNSFEKMKVKYATQVFSQTVSSAIKTLIQ
Dvir\Galileo IDTDITSTLFLCPKLSRKHIYPNYFKNMKVKYATQILSHAVASATKTLIQ
Dmoj\GalileoC TETKNLTR--LCPKLTLKHVSPNCFEKMKVKFATQIFSHSVAAAIRTVVE
Dmoj\GalileoD TETKNLTR--LCPKLTLKHVSPNCFEKMKVKLATQIFSHSVAAAIRTVVE
Dbuz\Galileo IESKNLTR--LCPKLTSKHVSPNCFEKMKVKYATQVFSHSVAAAIRTVID
Dmel\1360 LDHGSVTR---MTKLTRSHVNPTRFELMRVCLATQTLSHTVAAAIKTCNQ
Dsim\1360 LDHGSVTR---MTKLTRSHVNPTRFELMRVCLATQTLSHTVAGAIKTGNQ
Dsec\1360 LDHGSVTR---MTKLTRSHVNPTRFELMRVCLATQTLSHTVAAAIKTSNQ
Dere\1360 LDQRSVTR---MTKLTRSHVNPTRFELMRVCLATQTLSHTVAAAIKTCNQ
Dyak\1360 LDQGSVTR---MTKLTRSHVNPTRFELMRVCLATQTLSHTVAAAIKACNQ
Dper\1360 MEEHAGTK--MCPKLSRHHIYPNSFEKMRVKFATQIFSRTVQAAIKTVCE
Dpse\1360 MEEHAGTK--MCPKLSRHHIYPNSFEKMRVRFATQIFSRTVQAAITTVCE
Dvir\1360 LEQGSVTR---MTKLTKSHVNPTRFELMRVCLATQIFSHTVAAAIRTCNK
660 670 680 690 700
Dmel\P-element LG----YDIENATETADFFKLMNDWFDIFNSKLSTSNCIECSQPYGKQLD
Dbif\P-element LG----YQVENAVETSDLFKLLNDWFDVFNSKLSTSNCIETTQPYGKQLE
Dhel\P-element LG----YDVENACETSDLFKLLNDWFDLFNSKLSTANCIQSTQPYGKQLP
Dwil\P-element LG----YDIENATETADFFKLMNDWFDIFNSKLSTSNCIECSQPYGKQLD
Spal\P-element LG----YDVENACETSDLFKLLNDWFDVFNSKLSTANCIQSTQPYGKQLE
Hsap\THAP9 LDL---PPFQNCIGTIHFLRLINNLFDIFNSRN-----CYGKGLKGPLLP
Dana\Galileo SGGFV-DCRNSAEATANFIENVNKLFDCLNSHV-----LYEKNPDRCALQ
Dpse\Galileo SGGFL-KCKENAEATATFIEKMNRLFDCLNSHV-----LYDKNPFRSALQ
Dper\Galileo SGGFL-KCKENAEATATFIEKINRLFDCLNNHV-----LYDKNPFRSALQ
Dwil\Galileo HGKFI-DCEDVAIATSKFIEKVNRLFDCLNSSN-----IYDRNPNKSAIQ
Dvir\Galileo NGNFA-DCRDIALSTAKFIERVNKLLDCLKSNV-----LKDKNLFESALQ
Dmoj\GalileoC TGGFA-DCKDSAVATAIFIDKINNLFDCLNSHV-----LFDSNPYRCALR
Dmoj\GalileoD TGGFA-DCKDSAVATAIFIDKINNLFDCLNSHV-----LFDSNPYRCALR
Dbuz\Galileo SGGFS-DCKDSAVATAIFIEKINRLFDCLNSHV-----LFDSNPYRCALT
Dmel\1360 NKQLHRNSSEVAASTAAFVQKDNDYFDCLNSRV-----LTDKNPMKCALQ
Dsim\1360 NKQLHRNSSEVAASTAAFVQKVNDYFDCLNSRV-----LTEKNPMKCALQ
Dsec\1360 NKQLHRNSSEVAASTAAFVQKVNDYFDCLNSRV-----LTDKNPMKCALQ
Dere\1360 NKQFHRNSPEVAASTAAFVQKVNDYFDCLNSRV-----LTDKNPMKCALQ
Dyak\1360 NKQFNRNSPEVAASTAAFVQKVNDYFDCLNSRV-----LTDKNPMKCALQ
Dper\1360 TVGFKNSTYQVALSTAEFINKIDQIFDCMNSGS-----LYADNVYRSAIQ
Dpse\1360 TVGFKNSTYQVSLSTAEFINKVDQIFDCMNSGS-----LYADNVYRSAIQ
Dvir\1360 NKQLQRSSSEVADATATFVEKVNDYFDCLNSRV-----INDNNPMKCALQ


710
720
730
740
750
Dmel\P-element IQ-NDILNRMSEI----MRT------GILDKPKRLPFQKGIIVNNASLDG
Dbif\P-element LQ-RDILKQMSHI----MSN------RICGQTHRLPFQKGILINNASLDG
Dhel\P-element FQ-RDVLEKMSKI----MSS------EILGKSRKLPFQKGILVNNASLDG
Dwil\P-element IQ-NDILNRMSEI----MRT------GILDKPKRLPFQKGIIVNNASLDG
Spal\P-element FQ-RDVLEKMTQL----MCS------DILGRSQKLPFQKGIIVNNASLDG
Hsap\THAP9 ETYSKINHVLIEAKTIFVTLSDTSNNQIIKGKQKLGFL-GFLLNAESLKW
Dana\Galileo KN-NNVHNYLVEMRKYFAEF------KY---PQVVHCIDGMMLTISSVLA
Dpse\Galileo DK-NLVNETLSDMRKYFEEF------KY---PQEVYCIKGMILTINSILS
Dper\Galileo DK-NLVNETLSDMRKYFEEF------KY---PQEVYCIKGMILTINSILS
Dwil\Galileo KD-SDNEQYIIEMRDYFKKC------LY---RRKVYCLDGIVLSINAILM
Dvir\Galileo NN-NIKEKYITEMPNYFMKC------RY---LKTVYCINGLILTINSVLK
Dmoj\GalileoC EN-NNVHEYLQEMRDYFLNL------HY---PHKVYCIDGMLITISSVIA
Dmoj\GalileoD EK-NNIHEYLLEMRDYFKNL------HY---PHKVYCIDGMIITISSVIA
Dbuz\Galileo RN-NNVHEYLQEMRDYFHDL------QY---PQKVYCITGMIITISSVIA
Dmel\1360 VN-NGVWNKLKEMQEYLKSV------KY--HGNKIYCVDGLIQTTEAIFG
Dsim\1360 VN-NAVWNKQKEMQEYLKSV------KY--HGNKIYCVDGLIQTTEAIFG
Dsec\1360 VN-NAVWNKLKEMQEYLKSV------KY--HGNKIYCVDGLIQTTEAIFG
Dere\1360 FN-NTVWNKLKEMQEYLRNV------KY--HGNSIYCLDGLIQTTEAIFG
Dyak\1360 VN-NTVWHKLKEMQEYLRSV------KY--HGNNIYCLDGLIQTTEAIFG
Dper\1360 LN-NVPHKFIQFFLSYIKDV------NFVDSKKRVYFLDGIQITLKSLLL
Dpse\1360 LN-NVPHKFIQFFLSYIKDV------NFVDSKKRVYFLDGIQITLKSLLL
Dvir\1360 KE-NVVWKKLKEMQVYLRNI------RY--QGNLTYCIDGSLQTTEAIFG
760
770
780
790
800
Dmel\P-element LYK-YL-QENFSMQYILTSRLNQDIVEHFFGSMRSRGGQFDHPTPLQFKY
Dbif\P-element LHA-YC-NEKYGMEYILTSRLNQDIVENFFGAMRAKGGQHDHPSPLQFKY
Dhel\P-element LYI-YL-KDKYKMEYLLTSRLNQDIVDNFFGAMRSRGGQFDHPTPLQFKY
Dwil\P-element LYK-YL-QENFSMQYILTSRLNQDIVEHFFGSMRSRGGQFDHPTPLQFKY
Spal\P-element LFI-YL-KDKYNMEYLLTSRLNQDIVENFFGAMRSRGGQYDHPTPLQFKY
Hsap\THAP9 LYQNYVFPKVMPFPYLLTYKFSHDHLELFLKMLRQVLVTSSSPTCMAFQK
Dana\Galileo LSE-RVWS--SEIFYISTAKLNQDPLENLFYLIRARGATNNNPLMSEFNN
Dpse\Galileo LAQ-NVWSESAEVFYIATSKLNQDPLENLFYLIRSRGVTNNNPTMYEFNV
Dper\Galileo LAQ-NVWSESAEVFYIATSKLNQDPLENLFYLIRSRGVTNNNPTMYEFNV
Dwil\Galileo LTS-DIWNEGHGVFFLMLSRLNQDALEHVFYLIRSRGGTNNNPMLFEFNA
Dvir\Galileo LSQ-DIWREDSNVFFLILSRLNQDALENLFYLLRDRGITYSNPKLFEFNA
Dmoj\GalileoC LAE-DIWNNNTDIFFVATSKLNQDPLENLFYLIRCRGGTNSNPTVFEFNT
Dmoj\GalileoD LAE-NIWSDNNDLFFIATSKLNQDPLENLFYLIRSRGATNTNPTIFEFNS
Dbuz\Galileo LAE-NIWNDNNDLFFVATSKLNQDPLENLFYLIRSRGATNTNPTIFEFNS
Dmel\1360 LVE-DLFKDHTDHFFFLTSRVNQDPLENIFACVRAKGGNCRNPSVNEFNI
Dsim\1360 LVE-DLFKDHADHFFFLTSRVNQDPLENIFACVRAKGSNCRNPSVNEFNI
Dsec\1360 LVE-DLFKDHADHFFFLTSRVNQDPLENIFACVRAKGGNCRNPSVNEFNI
Dere\1360 LVK-DIFRDHTDHFFFLTSRVNQDPLENIFACVRAKGGKCRNPSVNEFNI
Dyak\1360 LVK-DIFKDHTDHFFFLTSRVNQDPLENIFACVRAKGGNCRNPSVNEFNV
Dper\1360 LAD-ELLT-NSDKIFIMTKSLNQDKLENTFAVVRQKGGNNTNPSVAEINN
Dpse\1360 LAD-ELLT-NSDKIFIMTKTLNQDKLENTFAVVRQKGGNNTNPSVAELKN
Dvir\1360 LVD-DLFKDHPDNFFFLTSRINQDPLENIFASVRAKGGNCRNPSVYEFNI


810 820 830 840 850
Dmel\P-element RLRKYII----------GMTNLKE-CV
Dbif\P-element RLRKYIV----------AKNTELLAG
Dhel\P-element RLKKYLI----------AKNTELLRN
Dwil\P-element RLRKYII----------ARNTEMLRN
Spal\P-element RLRKYLIGMSNLEELCGVLCVLSFMCFFQFYYLIIIFILYPAKNTELLRN
Hsap\THAP9 AYYNLET----------RYKFQDEVFLSKVS
Dana\Galileo IMSKMLS----------MKILTSKSV
Dpse\Galileo IISKMLS----------MKVLTSTTV
Dper\Galileo IISKMLS----------MKVLTSTTV
Dwil\Galileo IISKMLS----------MKLITSKTT
Dvir\Galileo IISKMLS----------MKIFTAKIS
DmojGalileoC IISKMLS----------MKMLTSASV
DmojGalileoD IISRMLS----------MKILTSASV
Dbuz\Galileo IISKMLS----------MKVLTSASI
Dmel\1360 IIAKLIS----------LHIFKF-SQ
Dsim\1360 IIAKLIS----------LHIFKF-SQ
Dsec\1360 IIAKLIS----------LHIFKF-SQ
Dere\1360 IIAKLIS----------LHIFKF-SQ
Dyak\1360 IIAKLIS----------LHIFKF-SQ
Dper\1360 IFARILN----------IKIVCS-SD
Dpse\1360 IFAKILN----------KLKIVRS-SD
Dvir\1360 IIAKLIS----------LHIFHF-TK
860 870 880 890 900
Dmel\P-element NKNVIPD-NSESWLNLDFSSKENENKSKDDEP--VDDEPVDEMLSNIDFT
Dbif\P-element NGNVDED-NCDSWLNLNITPNGNKE--------NEPDEGKWKGWSKEFEE
Dhel\P-element TGNVEED-NFDSWLNLDFSSK----------SLRNKPEDDEPEDDEQG
Dwil\P-element SGNIEED-NSESWLNLDFSSKENENKSKDDEP--VDDEPVDEMLSNIDFT
Spal\P-element TGNVAED-NCDSWLNLDFNSKSLEKKENKPED----VEPEDVEPEDEADE
Hsap\THAP9 IFDISIARRKDLALWTVQRQYGVSVTKTVFHEEGICQDWSHCSLSE----
Dana\Galileo SGNFGPD--DDTMLINVIQDCSTNKICNNLKT-DEEESTDFDMFSDEETE
Dpse\Galileo SGNCSPD--EDTMLINIIKDNVSSESASESKT----FDDDILISTNEDAE
Dper\Galileo SGNCSPD--EDTMLISIIKDNVSSESASESKT----FDDDILISTNENAE
Dwil\Galileo TGNCEPD--ED-MLINVIEETKHELAIENVNDQDQVYYEDFNIILDENMK
Dvir\Galileo SGNCQPN--GEFMLVNVIELANEKCKAFVLRT----KNICPITSSALNIS
DmojGalileoC SGNCIPD--EDLMLANIIKDSGSQLSVFHEQC-NSCHTPTEIEPLDDDLE
DmojGalileoD SGNCIPD--EDLMLANIIEDSGSQLSVFHDQC-NSRHTATDIEP-DADLE
Dbuz\Galileo SGNCILD--EDSMLANIIKDSGSTLSVFHSQC--EIHSSVYEEPSDPDFE
Dmel\1360 KSNCESD--DDVMLPIEFDSIIYQPFVEKKEI---QQQEYSVSFSKIVQD
Dsim\1360 KSNCESD--DDVMLPIEFDSIIYQPFIEKKKK-EIQQQEYSVSFSKMVQD
Dsec\1360 KSNCESD--DDVMLPIEFASIIYQPFIEKKEI---QQQEYSVSFSKILQD
Dere\1360 KSNCESN--DDVMLPIEFDSIIYQPCVEKKEI----QQEYSVSFSEIVQG
Dyak\1360 NSNCESD--DDVMLPIEFDSIIYQPCIEKKEI---QQQEYSVSFSEIVEG
Dper\1360 FGNCEADFEEGAAVQACIEGVFNESNNLKVEH-PNDHDELLSKLDFDGSG
Dpse\1360 FGNCEADFEEGIAVQACIEGVFNESNNLKDEH-PNHHDELLSKLDFDGSG
Dvir\1360 KSNCESD--DDVMLPVEFDSIIYEPYENNESK-VVPDNEFSVSLSQIVKG

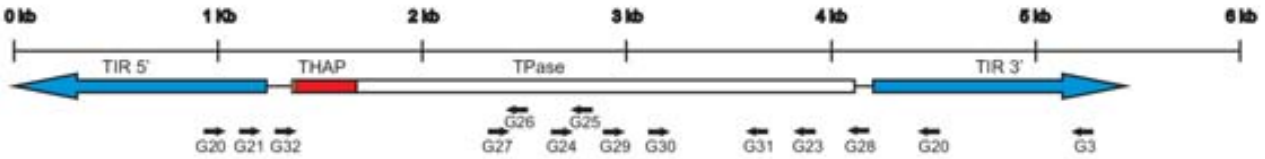

910 920 930 940 950
Dmel\P-element
Dbif\P-element
Dhel\P-element
Dwil\P-element
Spal\P-element
Hsap\THAP9
Dana\Galileo
Dpse\Galileo
Dper\Galileo
Dwil\Galileo
Dvir\Galileo
Dmoj\GalileoC
Dmoj\GalileoD
Dbuz\Galileo
Dmel\1360
Dsim\1360
Dsec\1360
Dere\1360
Dyak\1360
Dper\1360
Dpse\1360
Dvir\1360
960 970 980 990 1000
Dmel\P-element
Dbif\P-element
Dhel\P-element
Dwil\P-element
Spal\P-element
Hsap\THAP9
Dana\Galileo
Dpse\Galileo
Dper\Galileo
Dwil\Galileo
Dvir\Galileo
DmojGalileoC
DmojGalileoD
Dbuz\Galileo
Dmel\1360
Dsim\1360
Dsec\1360
Dere\1360
Dyak\1360
Dper\1360
Dpse\1360
Dvir\1360


1010 1020 1030 1040 1050
Dmel\P-element NNN------NFEITNNVKEKLILAARNVDVDKQVKSF------YFKIR--
Dbif\P-element HNT-------IEITKDIKKKLLIAAKHVQIDNNVKQF------YFKTR--
Dhel\P-element KEK-------FEITNNLKEKL--AAQNVELDKLVISF------YFKMR--
Dwil\P-element NNN------NFEITNNVKEKLILAARNVDVDKQVKSF------YFKIR--
Spal\P-element KDNFQVTNNNFKVTNNLKEKLVVAAQNVELDKLVISF------YFKIR--
Hsap\THAP9 RMA------IFELVSKQRELYLQQKILCELSGHINLFVDVNKHLFDGEVC
Dana\Galileo DKK--------PYINRIKKTIIQSCISSTEKSSIYSD------WFSVTHP
Dpse\Galileo ETK--------SHKFNLKRSIVQKCIILTSKIDKYAD------WFKFSDP
Dper\Galileo ETK--------SHKFNLKRSIVQKCIILTSKIDKYAD------WFKFSDP
Dwil\Galileo ----------PHARKIKEKIINECLSNIEKDPNYLD------WYSESHE
Dvir\Galileo QKS-----QHVRQKKKRRELTIYECISRNKAFEKFAN------WFSESHS
DmojGalileoC EKK--------PYIARIKEKIVLYCMHATAKSSLDNE------WFSPTHP
DmojGalileoD EKK--------PYMPRIKEKIVLYCMHATAKSSLDTE------WFSPTHP
Dbuz\Galileo DKK--------PFIAKLKEKIILHCMRATAKSTLHSD------WFSPSHP
Dmel\1360 KNN--------KKQMCIKKFIVEQCIKCTNESSAFPL------WFYENNE
Dsim\1360 KNN--------KKQMCIKKFIVEQCIKCTNESSAFSL------WFYENNE
Dsec\1360 KNN--------KKQMCIKKFIVEQCIKCTNESSAFSL------WFYENNE
Dere\1360 KNN--------KKRRCIKQFIVEQCIKCTNKSSDFSL------WFHSNND
Dyak\1360 KNN--------KKQMCIQKFIVDQCIKCTNESSDFSL------WFHVENE
Dper\1360 EIG--------PQKVGIKNLILKACQDETPL----------WFTGE--
Dpse\1360 EIG--------PQKVGIKNLILKACQDETPL----------WFTGE--
Dvir\1360 STN--------KMQLNIKKCIIEQCIKCT-KESVYSS------WFDENDS
1060 1070 1080 1090 1100
Dmel\P-element -IYFRIKYFNKKIEIKNQKQ----------
Dbif\P-element -IYFRLKYLNKKLAIKNQKQ----------
Dhel\P-element -IYLRVKYLNKKMYIKNQKR----------
Dwil\P-element -IYFRIKYFNKKIEIKNQKQ----------
Spal\P-element -IYFRVKYLNKKICIKNQKQ----------
Hsap\THAP9 AINHFVKLLKDIIICFLNIRAKNVAQNPLKHHSERTDMKTLSRKHWSSVQ
Dana\Galileo CYEHRMKMLNGFILILLRKNSK----------WLTEK
Dpse\Galileo CFVHRNHLLNQFVLILIRKNFK----------WLTDR
Dper\Galileo CFVHRNHLLNQFVLILIRKNCK----------WLTDR
Dwil\Galileo CSEHRKFILNYFLLVLLKKNSK----------WLMES
Dvir\Galileo CSHHRKYILNYVLQVLFRDNSI----------WLIEK
DmojGalileoC CFEHRKSILNQFVLILIRKNCK----------WQTEK
DmojGalileoD CFEHRKFILNQFVLILIRKNCK----------WQTEK
Dbuz\Galileo CFDHRKFMLNQFVLILIRKNCK----------WLTDS
Dmel\1360 CYAHRTDLLNKLIKVLLFKHCK----------WTVIA
Dsim\1360 CYAHRTDLLNKLIKVLLFKHCK----------WTVMI
Dsec\1360 CYAHRTDLLNKLIKVLLFKHCK----------WTVMI
Dere\1360 CYVHKIDILNKLIKVLLFKHCK----------WTVTA
Dyak\1360 CYEHKIDLLKKLIKVLLFKHCK----------WTVIA
Dper\1360 CREHKLALLDFVILVLLRKHSL----------WAIRR
Dpse\1360 CREHKLALLDFVILVLLRKHSL----------WAIRR
Dvir\1360 CFTHKISLLDKLIKVLLFKHCK----------WTVMT


1110
1120
Dmel\P-element -------KLIGNSKLLKIKL------
Dbif\P-element -------RLVGNSKLLKIKL------
Dhel\P-element -------RLIGNSKLLKIKL------
Dwil\P-element -------KLIGNSKLLKIKL------
Spal\P-element -------RLIGNSKLLKIKL------
Hsap\THAP9 DYKCSSFANTSSKFRHLLSNDGYPFK
Dana\Galileo MSKEKAVSTASSRKLKILKE------
Dpse\Galileo MIGNQN-AKSYEKKLKILRQ------
Dper\Galileo MIGNQN-AKSYEKKLKILRQ------
Dwil\Galileo LCGASE-KHKSNRKIEILQS------
Dvir\Galileo LCGLSE-NPSSARKEKIVQS------
Dmoj\GalileoC IVGK---TSISKRKLKIIHQ------
Dmoj\GalileoD IVGK---TSISKRKLKIIHQ------
Dbuz\Galileo IVSKS--SNLSKSKLKMIRE------
Dmel\1360 D------RQKKQAKLSILSHE-----
Dsim\1360 AD-----RQKKQAKLSILSHK-----
Dsec\1360 AD-----KQKKQAKLSILSHK-----
Dere\1360 D------KQKKQAKLSILSHK-----
Dyak\1360 N------RQKSQAKLSILSHK-----
Dper\1360 G------KKNASKKLKIMAQ------
Dpse\1360 G------KKNASKKLKIMAQ------
Dvir\1360 D------SYKKKAKLNILSHK-----

**SI Figure 7**. Isolation of *Galileo* in *D. buzzatii*. (*A*) Molecular structure of the putative full *Galileo* element from *D. buzzatii*. The big blue arrows are the TIRs and the *white* rectangle is the ORF coding for the TPase with the THAP domain shown in red. Primer location is indicated by small arrows underneath. (*B*) PCR amplification of the full *Dbuz\Galileo* copy. Four PCR reactions yielded relatively long products that were subsequently sequenced and assembled. The fact that each PCR product produced a single nucleotide sequence and that the overlapping portions between the four sequences were 99.99% identical (a single mismatch), suggests that they come from a single genomic *Galileo* copy. The sequence of the TIR ends was taken from the previously known *D. buzzatii Galileo*-3 sequence (accession no. AF368897). (*C*) Sequences of the primers used for amplification and sequencing of PCR products. Primers M13F and M13R are universal primers from the sequencing vector, bacteriophage M13.

A




B

| PCR product name | PCR product lenght | Sequencing primers |
|---|---|---|
| G24-G3 (G24 → G3) | 2697 bp | G3, G23, G24, G20 |
| G20-G25 (G20 → G25) | 1884 bp | G25, G26, G20, G21 |
| G27-G28 (G27 → G28) | 1890 bp | M13F, M13R, G29 |
| G20-G31 (G20 → G31) | 2758 bp | M13F, M13R, G21, G24, G26, G27, G30, G32 |

C

| Primer | Sequence (5'-3') |
|---|---|
| G3 | CAAAATCATTCCGCTTCGAG |
| G20 | CAATCGCCGAGTATCACTTGT |
| G21 | CGCAAATTCAGGTCACATATAAC |
| G23 | TGTAAGGTTGATTTAGCAGTAGC |
| G24 | ATATTGGACTATCGCAAGCAT |
| G25 | GCCCTGAGTTGTGGAGATGT |
| G26 | ACACAGACCTGCTGACCTAGAA |
| G27 | CTAATCCAGAGTTGCTAGAA |
| G28 | CATCCGTTATGTAGACATATT |
| G29 | GCATAACATCTCCACAACTCAG |
| G30 | ATGTGCATGAATACCTTCAGGA |
| G31 | CAGTTCTTTGCAACAGCTTGTG |
| G32 | TTGACACTCAACTTCCGA |
| M13F | GTAAAACGACGGCCAG |
| M13R | CAGGAAACAGCTATGAC |

**SI Table 1.1.** Number of significant hits produced in BLAST searches of the 12 *Drosophila* species genomes using different parts of *Galileo* and *1360* elements as queries. See Figure 1 for the longest copies of *Galileo* and *1360* in the six species and Materials and Methods for details.

| | | **Algorithms and queries used in the different searches** | | | | | |
|---|---|---|---|---|---|---|---|
| | | **TBLASTN** | **BLASTN** | **TBLASTN** | **BLASTN** | **BLASTN** | **BLASTN** |
| **Species** | **Database** | ***Dbuz\Galileo* TPase (912 aa)** | ***Dbuz\Galileo* TIR (40 bp)** | ***Dmel\1360* TPase (854 aa)** | ***Dmel\1360* TIR (31 bp)** | ***Galileo* longest copy** | ***1360* longest copy** |
| *D. simulans* | CAF1 Mosaic Chromosomes | 2 | 0 | 4 | 14 | — | 14 |
| *D. sechellia* | CAF1 Contigs | 18 | 0 | 84 | 575 | — | 690 |
| *D. melanogaster* | Release 4.2.1 Chromosomes | 2 | 0 | 6 | 7 | — | 7 |
| *D. yakuba* | CAF1 Contigs | 25 | 0 | 56 | 151 | — | 265 |
| *D. erecta* | CAF1 Contigs | 5 | 0 | 59 | 157 | — | 216 |
| *D. ananassae* | CAF1 Contigs | 11 | 4 | 8 | 0 | 216 | — |
| *D. pseudoobscura* | CAF1 Contigs (reconciled) | 23 | 0 | 12 | 0 | 109 | 121 |
| *D. persimilis* | CAF1 Contigs | 25 | 0 | 14 | 0 | 170 | 167 |
| *D. willistoni* | CAF1 Contigs | 90 | 17 | 49 | 0 | 495 | — |
| *D. mojavensis* | CAF1 Contigs | 64 | 25 | 35 | 0 | 367/287* | — |
| *D. virilis* | CAF1 Contigs | 7 | 78 | 9 | 0 | 134 | 295 |
| *D. grimshawi* | CAF1 Contigs | 0 | 0 | 0 | 0 | — | — |

* Two different searches have been performed in *D. mojavensis*, one with *Dmoj\GalileoC* (contig 10758) and another with *Dmoj\GalileoD* (contig 9930).

**SI Table 1.2.** Best hits recovered using TBLASTN and the amino acid sequence of the *Dbuz\Galileo* TPase as query. For *D. sechellia*, *D. erecta* and *D. ananassae*, two or three hits with similar E-values have been listed.

| Species | Best hit | Coordinates | Identity | Positives | BLAST score | E-value |
|---|---|---|---|---|---|---|
| *D. melanogaster* | Chr. 4 | 811865-809967 | 239/644 (37%) | 369/644 (57%) | 420 | 1e-123 |
| *D. simulans* | ChrU | 8187885-8186911 | 116/349 (33%) | 192/349 (54%) | 194 | 1e-083 |
| *D. sechellia* | Contig 5259 | 3350-4663 | 162/456 (35%) | 261/456 (56%) | 280 | 1e-114 |
| | Contig 9279 | 912-2552 | 203/557 (36%) | 323/557 (57%) | 367 | 1e-114 |
| | Contig 5902 | 3196-4743 | 195/534 (36%) | 308/534 (57%) | 340 | 1e-92 |
| *D. erecta* | Contig 7407 | 139694-138903 | 90/269 (33%) | 131/269 (48%) | 136 | 1e-154 |
| | Contig 7387 | 107457-106795 | 80/233 (34%) | 124/233 (52%) | 122 | 5e-55 |
| *D. yakuba* | Contig 0.40 | 345839-348313 | 296/892 (33%) | 466/892 (52%) | 459 | 1e-128 |
| *D. ananassae* | Contig 19410 | 9020-10615 | 322/555 (58%) | 407/555 (73%) | 633 | 0 |
| | Contig 15556 | 3697-5688 | 428/670 (63%) | 556/670 (82%) | 897 | 0 |
| *D. pseudoobscura* | Contig 3152 | 9685-8285 | 269/472 (56%) | 359/472 (75%) | 542 | 1e-164 |
| *D. persimilis* | Contig 7728 | 2761-3471 | 144/242 (59%) | 191/242 (78%) | 287 | 1e-128 |
| *D. willistoni* | Contig 10048 | 88626-85993 | 381/889 (42%) | 583/889 (64%) | 758 | 0 |
| *D. virilis* | Contig 16409 | 4917-6518 | 208/539 (38%) | 311/539 (57%) | 368 | 0 |
| *D. mojavensis* | Contig 11255 | 2735-5236 | 634/896 (70%) | 734/896 (81%) | 1289 | 0 |

**SI Table 1.3.** Best hits recovered using TBLASTN and the amino acid sequence of the *Dmel\1360* TPase as query. For *D. sechellia*, *D. erecta*, and *D. ananassae* two or three hits with similar E-values have been listed.

| Species | Best hit | Coordinates | Identity | Positives | BLAST score | E-value |
|---|---|---|---|---|---|---|
| *D. melanogaster* | Chr. 4 | 811868-809910 | 652/653 (99%) | 653/653 (99%) | 1333 | 0 |
| *D. simulans* | ChrU | 8187927-8186911 | 307/342 (89%) | 321/342 (93%) | 611 | 0 |
| *D. sechellia* | Contig 5902 | 5118-3148 | 565/672 (84%) | 589/672 (87%) | 1097 | 0 |
| | Contig 5259 | 3347-4723 | 437/468 (93%) | 447/468 (95%) | 882 | 0 |
| | Contig 9279 | 912-2606 | 533/565 (94%) | 548/565 (96%) | 1083 | 0 |
| *D. erecta* | Contig 7387 | 136850-137770 | 225/308 (73%) | 247/308 (80%) | 431 | 0 |
| | Contig 7407 | 139694-138894 | 192/272 (70%) | 204/272 (74%) | 353 | 0 |
| *D. yakuba* | Contig 0.40 | 345812-348370 | 755/855 (88%) | 805/855 (93%) | 1553 | 0 |
| *D. ananassae* | Contig 19410 | 9020-10603 | 176/536 (32%) | 282/536 (51%) | 281 | 1e-122 |
| | Contig 1556 | 3697-5718 | 259/691 (37%) | 393/691 (56%) | 445 | 1e-124 |
| *D. pseudoobscura* | Contig 784 | 23428-24918 | 179/542 (33%) | 276/542 (50%) | 265 | 1e-105 |
| *D. persimilis* | Contig 9857 | 65325-64489 | 106/294 (36%) | 153/294 (51%) | 154 | 1e-107 |
| *D. willistoni* | Contig 10048 | 88626-85951 | 296/908 (32%) | 469/908 (51%) | 468 | 1e-131 |
| *D. virilis* | Contig 17537 | 35605-33872 | 373/582 (64%) | 440/582 (75%) | 729 | 0 |
| *D. mojavensis* | Contig 11255 | 2777-5275 | 294/869 (33%) | 473/869 (53%) | 474 | 1e-133 |

**SI Table 1.4.** Complete and nearly-complete *Galileo* copies. TIRa is the TIR that is positioned at 5' end of the TPase. TSD sequences are given in the orientation they appear in the contig. When both TSD are exactly the same, only one sequence is given. Total lenght is expressed in bp and without the insertions of ther TEs that do not correspond to *Galileo* sequences. Insertions marked with TE are sequences inserted inside *Galileo* which have themselves TE structure (inverted repeats and/or TSD) and the ones marked as Ins are repetitive sequences dispersed in the genome but without any identifiable structure. The insertions with homology to known elements are indicated with the name of the corresponding TE.

| Contig | Coordinates | Orientation | Transposase coordinates | TIRa/ TIRb | TSD | Total lenght | Insertions | Observations |
|---|---|---|---|---|---|---|---|---|
| ***D. ananassae*** | | | | | | | | |
| 11169 | 1-2142 | Direct | 745-2142 | 198/- | — | 2142 | — | TIRa: end 457 bp and TSD missing; TIRb: missing<br>TPase: last 1178 bp missing |
| 15556 | 1821-6699 | Direct | 3049-5748 | 684/683 | ATATCAC | 4849 | — | — |
| 15979 | 71824-74395 | Direct | 73038-74395 | 661/- | GTAAAAT/- | 2572 | — | TIRb: missing; TPase: last 1154 bp missing |
| 16863 | 1564-5179 | Inverted | 4003-5179 | -/561 | ATTATAT/- | 2006 | 1897-3506 (Tc1-like) | TIRa: missing; TPase: first 1487 bp missing |
| 16864 | 6713-9433 | Inverted | 6713-8193 | 670/- | ATATAAT/- | 2721 | — | TIRb: missing<br>TPase: last 1666 bp missing |
| 17710 | 547-1780 | Inverted | 1107-1780 | -/314 | — | 1240 | — | TIRa: missing; TIRb: end 350 bp and TSD missing<br>TPase: first 1931 bp missing |
| 19410 | 7756-12565 | Direct | 8996-11618 | 671/682 | GCATAAC | 4810 | — | — |
| 19478 | 8730-10438 | Direct | 8730-9620 | -/602 | -/GGATGAC | 1709 | — | TIRa: missing<br>TPase: first 1488 bp missing, internal 293-bp deletion |
| 19479 | 11782-13967 | Inverted | 13266-13967 | -/621 | GTATAAT/- | 2186 | 1260-12706 (TE) | TIRa: missing; TIRb: insertion 647 bp<br>TPase: first 1675 bp missing, internal 309-bp deletion |

*(Continue on next page)*

| Contig | Coordinates | Orientation | Transposase coordinates | TIRa/ TIRb | TSD | Total lenght | Insertions | Observations |
|---|---|---|---|---|---|---|---|---|
| ***D. pseudoobscura*** | | | | | | | | |
| 3151 | 20609-23048 | Direct | 21387-22973 | 728/- | — | 2439 | — | TIRa: end 14 bp and TSD missing; TIRb: missing<br>TPase: first 1128 bp missing |
| 3152 | 8138-10463 | Inverted | 8131-9685 | 729/- | — | 2333 | — | TIRa: end 14 bp and TSD missing; TIRb: missing<br>TPase: first 1128 bp and last 25 bp missing |
| 3409 | 4744-6991 | Direct | 5918-6991 | 834/- | ATACAAC/- | 2208 | — | TIRa: 90-bp internal duplication; TIRb: missing<br>TPase: first 547 bp and last 1079 bp missing |
| 4007 | 52290-58223 | Direct | 55218-55959 | 543/474 | -/GTTGTAC | 2315 | 52702-54863 (Ins)<br>56378-57834 (ISY3) | TIRa: end 194 bp and TSD missing<br>TPase: first 1951 bp missing |
| 4025 | 5659-265 | Inverted | 782-265 | 658/- | -/GTAACTC | 1832 | 1527-5125 (TE) | TIRb: missing; TPase: last 2190 bp missing |
| ***D. persimilis*** | | | | | | | | |
| 2279 | 35952-37563 | Direct | 35952-37131 | -/235 | — | 1612 | — | TIRa: missing; TIRb: end 220 bp and TSD missing<br>TPase: first 1251 bp missing |
| 2979 | 64157-66292 | Inverted | 64159-65245 | 740/- | -/GTATAAC | 2136 | — | TIRb: missing; TPase: first 540 bp and last 1079 missing |
| 7728 | 1680-3474 | Direct | 2328-3504 | 579/- | CTTATTA/- | 1795 | — | TIRb: missing; TPase: first 1084 bp and last 292 missing |
| 7729 | 3501-6297 | Direct | 4139-5587 | 579/394 | CTTATTA/<br>TAATAAG | 2797 | — | TIRb: internal 202-bp deletion<br>TPase: first 1084 bp missing |
| 13439 | 141-2018 | Inverted | 141-740 | 689/- | — | 1878 | — | TIRa: end 145 bp and TSD missing<br>TPase: last 2085 bp missing |

*(Continue on next page)*

| Contig | Coordinates | Orientation | Transposase coordinates | TIRa/ TIRb | TSD | Total lenght | Insertions | Observations |
|---|---|---|---|---|---|---|---|---|
| ***D. willistoni*** | | | | | | | | |
| 6088 | 3693-5141 | Inverted | 4395-5141 | -/702 | ATTAAAG/- | 1449 | — | TIRa: missing; TPase: first 2000 bp missing |
| 8470 | 33713-37304 | Direct | 34270-35527 | 557/724 | CTGGAGC/ ACAATTT | 3095 | 36084-36580 (Ins) | TPase: first 1467 bp and last 16 bp missing |
| 9276 | 1-1020 | Direct | 747-1020 | 514/- | — | 1020 | — | TIRa: end 257 bp and TSD missing; TIRb: missing<br>TPase: first 51 bp and last 2409 bp missing |
| 9452 | 361705-364507 | Direct | 362664-363501 | 959/959 | CTACAAT | 2803 | — | TPase: first 1893 bp missing |
| 9601 | 20770-22344 | Inverted | 21779-2234 | -/411 | GTTCTAG/- | 1575 | — | TIRa: missing; TPase: first 2173 bp missing |
| 9602 | 1-1859 | Direct | 1-816 | -/412 | -/GTTCTAG | 1859 | — | TIRa: missing; TPase: first 1923 bp missing |
| 10048 | 85257-89642 | Inverted | 85942-88633 | 765/757 | ATTCTAG | 4386 | — | — |
| 12170 | 1-1320 | Inverted | 994-1320 | 749/- | — | 1320 | — | TIRa: end 16 bp and TSD missing; TIRb: missing<br>TPase: first 50 bp and last 2356 bp missing |
| ***D. virilis*** | | | | | | | | |
| 15993 | 12611-16444 | Direct | 12835-15635 | -/309 | — | 3197 | 13500-14136 (Ins) | TIRa: missing; TIRb: end 495 bp and TSD missing<br>TPase: 637 bp insertion |
| 15994 | 297-4088 | Direct | 525-3328 | -/326 | — | 3149 | 1190-1832 (Ins) | TIRa: missing; TIRb: end 495 bp and TSD missing<br>TPase: 543-bp insertion |
| 16409 | 3466-7941 | Direct | 4899-7707 | 767/232 | -/CTTCAAT | 4476 | — | TIRa: end 78 bp and TSD missing<br>TIRb: internal 553-bp deletion |

*(Continue on next page)*

| Contig | Coordinates | Orientation | Transposase coordinates | TIRa/ TIRb | TSD | Total lenght | Insertions | Observations |
|---|---|---|---|---|---|---|---|---|
| ***D. mojavensis*** | | | | | | | | |
| 7794 | 13889-19752 | Direct | 15733-19752 | 526/- | — | 3456 | 14470-15114 (TE4)<br>16251-16995 (ISBu)<br>17355-17921 (TE5) | TIRa: end 64 bp and TSD missing; TIRb: missing<br>TPase: deletion of first 71 bp and last 14 bp,<br>internal 36-bp duplication |
| 8435 | 1-4278 | Direct | 2326-4274 | 785/- | — | 3726 | 125-676 (TE2) | TIRa: end 29 bp and TSD missing; TIRb: missing |
| 9930 | 1467-8042 | Inverted | 2925-6622 | 574/576 | GTCCAAG/ ATTTAAG | 5675 | 2944-38449 (ISBu) | — |
| 10367 | 3528-7520 | Inverted | 4941-5542 | 1104/1104 | GTTGAGC | 3119 | 5846-6719 (ISBu) | TPase: first 2212 bp missing<br>TIRs: 391 bp of F2 included in both TIRs |
| 10369 | 31739-35574 | Direct | 33528-35574 | 694/- | CCTGAAC/- | 2898 | 32009-32946 (TE2) | TIRb: missing; TPase: first 767 bp missing |
| 10376 | 4316-10745 | Direct | 5737-8521 | 570/570 | TAATAAA | 5721 | 8815-9515 (TE1) | — |
| 10758.1 | 37586-44126 | Inverted | 38993-41776 | 813/713 | GTTACCG | 5989 | 43097-43648 (TE2) | — |
| 10765 | 52988-62433 | Inverted | 54923-58610 | 610/395 | CCACGAA/ CCACTAA | 4412 | 53253-54435 (TE3)<br>55104-56352 (ISBu x2)<br>59103-61704 (ISBu x4) | TPase: last 108 bp missing |
| 10770.1 | 9949-16490 | Direct | 11540-14367 | 583/709 | ATGGAGA/ TATTGAC | 5836 | 14413-15118 (ISBu) | — |

*(Continue on next page)*

| Contig | Coordinates | Orientation | Transposase coordinates | TIRa/ TIRb | TSD | Total lenght | Insertions | Observations |
|---|---|---|---|---|---|---|---|---|
| 10773 | 33627-39873 | Inverted | 35425-38494 | 570/376 | -/TTTATAT | 4130 | 34032-34586 (TE5)<br>35774-36615 (IN)<br>38495-39214 (ISBu) | TIRb: end 192 bp and TSD missing<br>TPase: first 15 bp missing |
| 10792 | 22486-27604 | Inverted | 23831-25781 | 815/645 | AATATAT | 5119 | — | — |
| 10918 | 2482-12359 | Direct | 8142-12359 | 694/- | ATACCAC/- | 3943 | 3762-8075 (Max_LTR)<br>8285-9121 (ISBu)<br>9800-10583 (ISBu) | TIRb: missing<br>TPase: last 79 bp missing |
| 10924 | 25932-30978 | Direct | 27530-30351 | 745/393 | -/CTTAAAT | 5047 | — | TIRa: end 41 bp and TSD missing |
| 10946 | 6739-13856 | Direct | 8917-12351 | 525/578 | CTGAATC/ CTAAATC | 5433 | 7431-8233 (ISBu)<br>8942-9823 (ISBu) | — |
| 11233 | 4001-11153 | Inverted | 5654-8461 | 788/959 | GTAGAAC/ GTATGGT | 6239 | 9305-10218 (ISBu) | TIRb: internal 216-bp duplication |
| 11255 | 1328-6787 | Direct | 2735-5284 | 557/556 | AATGTAT | 5460 | — | |

**SI Table 1.5.** Short non-autonomous copies of *Galileo* characterised in the genomes of six *Drosophila* species. TIRa is the first TIR that appears in the contig and TIRb is the other one. When both TSD are exactly the same, only one sequence is given. Total lenght is given in bp.

| Species | Contig | Beginning | Total length | TIRa/TIRb lenght | TSD | Observations |
|---|---|---|---|---|---|---|
| *D. ananassae* | 16072 | 49428 | 1145 | 317/317 | ATAGTAG | — |
| | 16215 | 13660 | 1226 | 329/329 | GTAATAC | — |
| | 16457 | 13174 | 1220 | 329/329 | GTTATAT | — |
| | 16780 | 48830 | 1194 | 331/331 | CATCAAC | — |
| | 16799 | 18339 | 1194 | 331/331 | GTAGCAG | — |
| | 17082 | 56734 | 1228 | 318/320 | GTTTCGT | — |
| | 18115 | 31181 | 1233 | 329/329 | ATTATAG | — |
| | 18348 | 28163 | 1225 | 329/329 | CTACGAG | — |
| | 18752 | 112063 | 1236 | 329/329 | GTATAAT | — |
| | 18811 | 266133 | 1244 | 329/328 | GATGAAC | — |
| | 18844 | 8374 | 1254 | 329/340 | CTTTTAT | — |
| | 18855 | 7980 | 1245 | 329/329 | GTATTAT | — |
| | 19356 | 469604 | 1219 | 329/329 | ACTGTAC | — |
| | 19465 | 81512 | 1158 | 329/317 | GTATAGT | — |
| | 19598 | 190010 | 1236 | 329/329 | CTTGTAC | — |
| | 19813 | 40154 | 1236 | 329/329 | GTTCAAC | — |
| | 19892 | 26883 | 1236 | 329/329 | GTATAAC | — |
| | 20116 | 42794 | 1227 | 331/330 | ACCAAAC | — |
| | 20508 | 21472 | 1194 | 331/330 | CTGCAAC | — |
| | 20509 | 13568 | 1208 | 329/304 | GTGCAAG | — |

*(Continue on next page)*

| Species | Contig | Beginning | Total length | TIRa/TIRb lenght | TSD | Observations |
|---|---|---|---|---|---|---|
| *D. pseudoobscura* | 1082 | 770 | 2899 | 462/464 | GTTCAAC | — |
| | 3611 | 65029 | 4538 | 545/541 | GTAATAT/GTTCTAT | — |
| | 3949 | 17417 | 1095 | 515/417 | ATAAATG | — |
| | 4181 | 261898 | 2106 | 664/630 | GTTCTAT | Worf insertion (263302-264678) |
| | 4197 | 39495 | 2439 | 629/630 | GTAGTAC | — |
| | 4227 | 206109 | 1982 | 552/552 | CTTCTGC | — |
| | 4314 | 60436 | 4458 | 543/545 | GTTCTAT | — |
| | 4350 | 119298 | 1975 | 514/514 | ATTCGAC | — |
| | 4355 | 300467 | 4085 | 528/493 | CTTTCAC/GTGAGAC | — |
| | 4360 | 196165 | 2036 | 630/629 | ATAAGAC | — |
| | 4568 | 59125 | 2036 | 630/629 | GCTAAAG | — |
| | 4925 | 27994 | 2024 | 623/623 | GTATACT | — |
| | 5286 | 303008 | 2059 | 778/778 | GTTTAAC | Insertion in TIRb (304447-304982) |
| | 5301 | 784883 | 2059 | 630/658 | CTACTAT | — |
| | 5346 | 39465 | 1954 | 561/624 | GTAATAT | — |
| | 5574 | 101925 | 3481 | 662/630 | ATAGCAT | — |
| | 5598 | 108632 | 2032 | 630/629 | GTACTAC | — |
| | 5611 | 88815 | 1610 | 536/535 | GTTGAGG | — |
| *D. persimilis* | 43 | 22757 | 1892 | 555/565 | GTTGAAT | — |
| | 497 | 41905 | 1893 | 558/549 | ATTATAC | — |
| | 661 | 13659 | 2012 | 737/733 | GTTGTAC | — |
| | 1311 | 29484 | 2087 | 550/555 | ATAGCAC | — |
| | 1330 | 56295 | 1957 | 515/515 | ATTGTAC | — |

*(Continue on next page)*

| Species | Contig | Beginning | Total length | TIRa/TIRb lenght | TSD | Observations |
|---|---|---|---|---|---|---|
| | 1462 | 9134 | 1908 | 559/559 | ATAAAAC | — |
| | 1531 | 30035 | 1949 | 725/719 | TTGATGG | — |
| | 2315 | 58360 | 1901 | 558/557 | ATATGAC | — |
| | 2841 | 31815 | 2005 | 555/553 | GTACAAC | — |
| | 3304 | 10212 | 1961 | 613/562 | GCAAGTC | — |
| | 4090 | 14313 | 1758 | 541/541 | ACCAACC | — |
| | 4609 | 1009 | 1902 | 559/569 | GTAATAC | — |
| | 5516 | 10726 | 2097 | 557/557 | GTTTTGC | — |
| | 6330 | 22364 | 1865 | 573/556 | GTAGTAG | — |
| | 9153 | 14702 | 1935 | 566/566 | CCGCAAA | — |
| | 9474 | 6659 | 1992 | 543/552 | CTTTCAT | — |
| | 10757 | 7184 | 2022 | 737/743 | GTATATA | — |
| | 11506 | 169 | 1946 | 727/725 | GTTGTGC | — |
| *D. willistoni* | 191 | 7706 | 1133 | 483/475 | ATATTAG | — |
| | 4171 | 163 | 1140 | 475/475 | GTATAAC | — |
| | 6280 | 219672 | 1126 | 517/517 | TGCAAAG | — |
| | 6423.1 | 286322 | 1452 | 624/620 | GTATTAG | — |
| | 6423.2 | 472923 | 1120 | 476/476 | ATTATAT | — |
| | 6463.1 | 759004 | 1141 | 475/483 | GTATTAG | — |
| | 6463.2 | 269200 | 1142 | 485/487 | ATCGTTT/ATCGTAT | 38-bp repeat at end of TIRb |
| | 6463.3 | 165185 | 1132 | 492/484 | ATATTAG | — |
| | 6661 | 512446 | 1269 | 522/528 | CTAAAAC | — |
| | 6834 | 123142 | 1105 | 517/521 | ATTCTGC | — |

*(Continue on next page)*

| Species | Contig | Beginning | Total length | TIRa/TIRb lenght | TSD | Observations |
|---|---|---|---|---|---|---|
| | 6840 | 36681 | 1130 | 519/519 | CTTGAAG | — |
| | 6847 | 108816 | 1114 | 518/517 | GTATTGA | — |
| | 6851 | 324389 | 1124 | 517/517 | CCTTTAC | — |
| | 7963 | 399697 | 1196 | 557/557 | CTACTGC | — |
| | 8445 | 45999 | 1140 | 518/518 | ATAGAAC | — |
| | 8628 | 277645 | 1133 | 475/483 | ATATTAC | — |
| | 9000 | 395296 | 1133 | 483/475 | GTCAAAG | — |
| | 9436 | 199308 | 1109 | 482/490 | CTTCTAC | — |
| | 9906 | 5225 | 1422 | 482/483 | GTATTAG | — |
| | 10422 | 27432 | 1402 | 476/475 | ATAACAG/CTCTAAC | — |
| *D. virilis* | 13546 | 3646 | 1991 | 852/990 | GTAATAC | Insertion within TIRa (3665-4506) |
| | 13964 | 226944 | 1602 | 709/696 | ACTGAAC/GCGATAG | — |
| | 14705 | 256157 | 1339 | 584/400 | ATATTAT | — |
| | 15758 | 236804 | 2232 | 785/785 | CTTAAAC | — |
| | 16069 | 102584 | 1480 | 608/614 | ACTTAAC | — |
| | 16071 | 203037 | 2343 | 1051/1003 | GTAACAG | — |
| | 16072 | 413871 | 1774 | 795/839 | GTATAAT/- | 5 terminal bp and TSD missing from TIRb |
| | 16403 | 21843 | 1635 | 725/730 | ATCGAGC/ATGCTGC | — |
| | 17557 | 39646 | 1882 | 824/832 | -/ATTACCA | 12 terminal bp and TSD missing from TIRa |
| | 17577 | 4833 | 1247 | 305/307 | GTAATAG | — |
| | 17588 | 42042 | 1616 | 699/687 | CAAGCAA | GTG is repeated between end of TIR and TSD |

*(Continue on next page)*

| Species | Contig | Beginning | Total length | TIRa/TIRb lenght | TSD | Observations |
|---|---|---|---|---|---|---|
| | 17658 | 62428 | 1947 | 710/723 | GTGATAC/CCATAAG | — |
| | 18052 | 6453 | 1597 | 698/722 | GTATTAC | — |
| *D. mojavensis* | 8189 | 12012 | 2382 | 894/894 | GTGCAGC | 124-bp internal direct repeats |
| | 8783 | 2229 | 2162 | 1039/990 | CTATAAC | Insertion in TIRa (2503-4931) |
| | 9647 | 4614 | 1903 | 578/578 | ATTTGAA | — |
| | 9832 | 28557 | 1600 | 577/565 | GTGATAT/AATACAC | — |
| | 10246 | 221912 | 2826 | 1116/1116 | GTATTTT | Two 220-bp repeats in each TIR; 127-bp internal direct repeats |
| | 10309 | 10744 | 2426 | 1147/1153 | GTACCGC | Two 220-bp repeats in each TIR |
| | 10727.1 | 44829 | 2206 | 1028/1030 | TCATTAC | — |
| | 10727.2 | 104767 | 3214 | 1260/1214 | GTATTAT | — |
| | 10741 | 55570 | 3363 | 1216/987 | ATATGTA/CTAATTG | TIRa has three copies of 220-bp repeat. TIRb has only two. Immediately upstream of this copy there is another TIR 813 bp long with two copies of 220-bp repeat flanked by 7-bp sequence CTATAAC. |
| | 10751 | 25070 | 2195 | 1021/1026 | TGTATAC | — |
| | 10758.2 | 55001 | 1660 | 486/508 | GTTATGC | 119-bp end duplication in TIRb |
| | 10764 | 29199 | 2369 | 715/715 | TTTATAT | Insertion in TIRb (35259-35358); ISBu insertion (30625-34928) |
| | 10770.2 | 92188 | 3199 | 1107/1107 | ATAGTAG/CTACTAT | — |
| | 10790.1 | 49101 | 1967 | 556/494 | ATACTAC | — |

*(Continue on next page)*

| Species | Contig | Beginning | Total length | TIRa/TIRb lenght | TSD | Observations |
|---|---|---|---|---|---|---|
| | 10790.2 | 84962 | 2199 | 1022/1028 | TCGAAAC | — |
| | 10940 | 39859 | 1547 | 458/458 | ATTGGGG | — |
| | 10945 | 2811 | 2179 | 1033/997 | GATACAC | Insertion in TIRb (4741-5878) |
| | 11229 | 48344 | 1736 | 769/766 | TTAATGC | — |
| | 11267 | 12789 | 2355 | 627/616 | GTATCAA | — |
| | 11679 | 95650 | 2174 | 1012/1014 | TTATGAG | ISBu insertion in TIRb (97089-97880) |

**SI Table 1.6**. General characteristics of non-autonomous *Galileo* copies found in the genomes of six *Drosophila* species. In D. mojavensis, 3 and 4 additional copies with nearly complete TPase were included for computation of average pairwise divergence in groups C and D, respectively. N indicates the number of characterised non-autonomous elements and Number of sites refers to the aligned TIR sites used for divergence analysis. Molecular clock was set assuming an average synonymous substitution rate of 0.016 substitutions/nucleotide/million years (19).

| Species | N | Average copy lenght (SD) | Average TIR length (SD) | Number of sites | Average pairwise divergence | | Divergence time (myr) |
|---|---|---|---|---|---|---|---|
| | | | | | Within copies | Between copies | |
| *D. ananassae* | 20 | 1217.9 (28.4) | 328.9 (4.4) | 286 | 0.0111 | 0.0282 | 1.76 |
| *D. pseudoobscura* | 18 | 2492.7 (989.7) | 596.0 (76.0) | 358 | 0.0053 | 0.0159 | 0.99 |
| *D. persimilis* | 18 | 1949.0 (81.1) | 598.2 (76.4) | 506 | 0.0044 | 0.0137 | 0.86 |
| *D. willistoni* | 20 | 1183.2 (110.5) | 506.5 (35.6) | 468 | 0.0161 | 0.0412 | 2.57 |
| *D. virilis* | 13 | 1745.0 (325.3) | 737.3 (184.9) | — | — | — | — |
| Group A | 6 | 2028.2 (216.6) | 870.0 (125.2) | 488 | 0.0245 | 0.0459 | 2.87 |
| Group B | 7 | 1502.3 (152.7) | 623.6 (150.5) | 248 | 0.0542 | 0.0571 | 3.57 |
| *D. mojavensis* | 20 | 2268.1 (525.2) | 885.2 (259.5) | — | — | — | — |
| Group C | 4 | 2284.5 (613.5) | 739.0 (255.2) | 408 | 0.0167 | 0.0242 | 1.51 |
| Group D | 2 | 1860.5 (443.4) | 736.0 (393.1) | 404 | 0.0120 | 0.0233 | 1.46 |
| Group E | 7 | 2147.4 (192.3) | 936.0 (167.7) | 555 | 0.0068 | 0.0221 | 1.38 |
| Group F | 6 | 2645.2 (626.0) | 1024.5 (283.1) | 460 | 0.0191 | 0.0887 | 5.54 |

**SI Table 1.7**. Complete and nearly-complete 1360 transposable elements found in different *Drosophila* species. TIRa is the TIR at 5' side of the transposase while TIRb designates the TIR at its 3' end. TSD are given in the same orientation they have in the contig.

| Species | Contig | Coordinates | Orientation | Transposase | TIRa/ TIRb | TSD | Total length | Insertions |
|---|---|---|---|---|---|---|---|---|
| *D. melanogaster*(*) | Chr. 4 | 809215-812470 | Inverted | 812470-809907 | -/31 | GCCATAC/- | 3615 | — |
| *D. simulans* | Chr. U | 8185141-8188796 | Inverted | 8188307-8185783 | -/31 | CCCGAAC/- | 3656 | — |
| *D. sechellia* | 5259 | 961-5382 | Direct | 2194-4726 | 31/31 | GTTCGAC/ ATTCCAC | 4422 | — |
| *D. erecta* | 7407 | 137172-141324 | Inverted | 140279-137852 | 31/31 | GTTTGAC | 4153 | — |
| *D. yakuba* | 0.40 | 344624-349004 | Direct | 345812-348373 | 31/31 | GTCAAAG | 4381 | — |
| *D. pseudoobscura* | 784 | 22941-26100 | Direct | 23339-25777 | 31/32 | ATCATAT | 3160 | — |
| *D. persimilis* | 9857 | 64151-67135 | Inverted | 66929-64479 | 32/31 | ATATGAT | 2985 | — |
| *D. virilis* | 17537 | 32378-37583 | Inverted | 33160-36198 | 32/31 | GTTTGAC | 4702 | LINE (33361-33864) |

(*) Type=chromosome; loc=4:1..1281640; ID=4; release=r4.2.1; species=dmel

**SI Table 1.8.** Comparison among the TPase proteins from different species. TPase length (L, first column), number of aligned amino acids omitting gaps (above diagonal) and percent identical amino acids (below diagonal) in pairwise comparisons between the sequences of the *Drosophila* 21 TPases (8 from *Galileo*, 8 from *1360* and 5 from *P*-element) and the *THAP containing 9* protein from *Homo sapiens*. The global alignment of the 22 protein sequences comprised 1126 positions.

| **Species** | **L** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** | **13** | **14** | **15** | **16** | **17** | **18** | **19** | **20** | **21** | **22** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1.** *Dana\Galileo* | 889 | | 885 | 879 | 881 | 880 | 886 | 885 | 883 | 834 | 838 | 836 | 834 | 833 | 794 | 799 | 836 | 722 | 726 | 718 | 723 | 735 | 807 |
| **2.** *Dpse\Galileo* | 908 | 56,2 | | 902 | 893 | 899 | 897 | 896 | 895 | 839 | 841 | 841 | 840 | 838 | 797 | 802 | 839 | 722 | 728 | 718 | 723 | 740 | 808 |
| **3.** *Dper\Galileo* | 902 | 55,6 | 97,2 | | 887 | 893 | 891 | 890 | 889 | 833 | 835 | 835 | 834 | 832 | 791 | 796 | 833 | 716 | 722 | 712 | 717 | 734 | 802 |
| **4.** *Dwil\Galileo* | 910 | 42,3 | 41,7 | 41,5 | | 888 | 902 | 901 | 898 | 836 | 840 | 838 | 836 | 835 | 795 | 800 | 838 | 718 | 726 | 715 | 719 | 735 | 806 |
| **5.** *Dvir\Galileo* | 938 | 41,6 | 39,9 | 39,4 | 52,3 | | 909 | 908 | 886 | 836 | 838 | 838 | 837 | 835 | 792 | 797 | 836 | 723 | 730 | 718 | 724 | 744 | 806 |
| **6.** *Dmoj\GalileoC* | 937 | 55,6 | 52,8 | 52,1 | 43,1 | 40,0 | | 936 | 905 | 840 | 844 | 842 | 840 | 839 | 800 | 805 | 842 | 724 | 731 | 721 | 725 | 740 | 809 |
| **7.** *Dmoj\GalileoD* | 936 | 57,2 | 55,0 | 54,3 | 43,3 | 41,0 | 87,3 | | 904 | 839 | 843 | 841 | 839 | 838 | 799 | 804 | 841 | 723 | 730 | 720 | 724 | 739 | 808 |
| **8.** *Dbuz\Galileo* | 912 | 58,3 | 56,2 | 55,6 | 41,8 | 40,9 | 72,4 | 74,0 | | 837 | 840 | 839 | 837 | 836 | 796 | 801 | 838 | 723 | 729 | 719 | 724 | 739 | 808 |
| **9.** *Dmel\1360* | 854 | 34,7 | 32,2 | 32,2 | 33,9 | 31,1 | 34,4 | 34,7 | 34,4 | | 854 | 854 | 853 | 852 | 797 | 802 | 853 | 721 | 727 | 716 | 721 | 738 | 798 |
| **10.** *Dsim\1360* | 858 | 34,6 | 31,6 | 31,6 | 33,1 | 30,9 | 33,2 | 33,7 | 33,1 | 93,8 | | 856 | 854 | 853 | 800 | 805 | 856 | 723 | 730 | 719 | 723 | 739 | 802 |
| **11.** *Dsec\1360* | 856 | 34,7 | 32,1 | 32,1 | 33,8 | 30,4 | 34,2 | 34,4 | 33,8 | 94,8 | 96,1 | | 854 | 853 | 798 | 803 | 854 | 722 | 728 | 717 | 722 | 739 | 800 |
| **12.** *Dere\1360* | 854 | 35,4 | 32,1 | 32,1 | 32,8 | 31,1 | 35,0 | 35,2 | 34,6 | 88,7 | 86,8 | 88,3 | | 852 | 797 | 802 | 853 | 721 | 727 | 716 | 721 | 739 | 798 |
| **13.** *Dyak\1360* | 853 | 35,4 | 31,5 | 31,4 | 32,9 | 30,9 | 34,7 | 34,8 | 34,6 | 88,8 | 87,0 | 88,2 | 91,7 | | 797 | 802 | 852 | 722 | 728 | 717 | 722 | 739 | 799 |
| **14.** *Dpse\1360* | 818 | 35,0 | 33,2 | 33,2 | 32,6 | 30,3 | 33,5 | 34,0 | 33,9 | 35,4 | 34,1 | 35,1 | 35,1 | 35,4 | | 811 | 804 | 695 | 703 | 691 | 695 | 710 | 764 |
| **15.** *Dper\1360* | 817 | 35,3 | 33,7 | 33,7 | 33,3 | 30,6 | 34,0 | 34,7 | 34,3 | 35,9 | 34,5 | 35,5 | 35,5 | 35,8 | 98,0 | | 799 | 691 | 698 | 686 | 690 | 706 | 760 |
| **16.** *Dvir\1360* | 856 | 34,7 | 32,4 | 32,4 | 32,7 | 29,5 | 33,8 | 33,9 | 33,9 | 71,7 | 70,4 | 71,5 | 72,5 | 72,5 | 34,8 | 34,5 | | 722 | 729 | 718 | 722 | 738 | 800 |
| **17.** *Dmel\P* | 751 | 23,1 | 24,2 | 24,0 | 22,3 | 21,2 | 23,8 | 24,6 | 25,3 | 23,2 | 22,3 | 23,0 | 22,7 | 22,7 | 23,3 | 22,9 | 23,3 | | 736 | 735 | 750 | 749 | 718 |
| **18.** *Dbif\P* | 757 | 23,1 | 23,4 | 23,5 | 22,7 | 21,1 | 24,6 | 24,7 | 24,8 | 23,4 | 23,0 | 22,9 | 22,1 | 22,9 | 23,6 | 23,2 | 23,0 | 66,7 | | 740 | 737 | 754 | 714 |
| **19.** *Dhel\P* | 746 | 22,7 | 22,8 | 22,9 | 22,1 | 20,1 | 23,9 | 24,0 | 24,1 | 23,2 | 22,7 | 22,6 | 21,9 | 22,3 | 22,3 | 22,0 | 22,6 | 75,0 | 66,5 | | 736 | 743 | 704 |
| **20.** *Dwil\P* | 751 | 22,5 | 23,7 | 23,4 | 22,3 | 21,1 | 23,0 | 23,9 | 24,9 | 23,2 | 22,3 | 23,0 | 22,7 | 22,6 | 23,3 | 22,9 | 23,4 | 98,1 | 66,9 | 75,7 | | 749 | 718 |
| **21.** *Spal\P* | 830 | 23,4 | 23,1 | 23,2 | 22,3 | 20,3 | 23,6 | 23,8 | 24,5 | 24,0 | 23,4 | 23,5 | 22,7 | 23,0 | 22,0 | 21,5 | 22,8 | 74,0 | 66,2 | 87,2 | 74,6 | | 731 |
| **22.** *Hsa\THAP9* | 903 | 20,9 | 20,9 | 21,1 | 19,1 | 19,1 | 20,5 | 20,3 | 20,4 | 18,8 | 18,7 | 18,6 | 18,9 | 19,9 | 18,1 | 17,6 | 19,1 | 21,0 | 19,2 | 20,6 | 21,0 | 20,1 | |

**SI Table 1.9.** Sequences used to construct the consensus transposases of *Galileo* and 1360 in the different species. Coordinates corresponding to the transposase sequence inside each contig are given following the transcriptional direction (from Methionine to STOP codon).

**A. *Galileo* sequences**

| Species | Contig | Coordinates |
|---|---|---|
| *D. ananassae* | 9736 | 951-1 |
| | 11169 | 745-2142 |
| | 15556 | 3049-5748 |
| | 15979 | 73038-74395 |
| | 16864 | 8193-6713 |
| | 19410 | 8996-11618 |
| *D. pseudoobscura* | 384 | 17-532 |
| | 521 | 5574-4963 |
| | 1362 | 13473-13218 |
| | 2192 | 3433-4322 |
| | 2193 | 3840-4137 |
| | 3151 | 21387-23048 |
| | 3152 | 9685-8131 |
| | 3311 | 5683-4590 |
| | 3409 | 5918-6991 |
| | 3514 | 4441-3863 |
| | 3688 | 28103-29511 |
| | 4007 | 55218-55959 |
| | 4025 | 782-265 |
| | 4842 | 7178-6832 |
| | 5255 | 6307-6857 |
| | 5529 | 5015-4446 |
| | 5668 | 514-1070 |
| *D. persimilis* | 2279 | 35952-37131 |
| | 2979 | 65246-64154 |
| | 7728 | 2360-3504 |
| | 7729 | 4139-5587 |
| | 7807 | 4183-4785 |
| | 9771 | 76484-77861 |
| | 11866 | 2506-3153 |
| | 12167 | 28-218 |
| | 12803 | 984-2401 |
| | 12806 | 4579-4847 |
| | 13439 | 740-141 |
| | 13644 | 5604-6247 |
| | 14651 | 3800-4468 |
| | 16801 | 936-669 |
| *D. willistoni* | 480 | 2760-1949 |
| | 1514 | 1430-255 |

*(Continue on next page)*

| Species | Contig | Coordinates |
|---|---|---|
| *D. willistoni* | 1633 | 1484-3134 |
| | 1765 | 758-1818 |
| | 3103 | 2956-1579 |
| | 3729 | 2677-1005 |
| | 4852 | 3775-2272 |
| | 5955 | 5147-3320 |
| | 5995 | 709-2234 |
| | 6043 | 7043-5519 |
| | 8665 | 22915-21026 |
| | 9276 | 484-1020 |
| | 9858 | 3576-3081 |
| | 10048 | 88633-85942 |
| | 12170 | 994-1320 |
| | 16933 | 1388-1 |
| *D. virilis* | 1717 | 1-1012 |
| | 15993 | 12835-13499, 14137-15635 |
| | 15994 | 525-1189, 1833-3328 |
| | 16046 | 32251-30912 |
| | 16409 | 4899-7707 |
| *D. mojavensis* | 7794 | 15733-16250, 16996-17354, 17922-19752 |
| | 8435 | 2326-4274 |
| | 9930 | 6622-3845, 2943-2925 |
| | 10367 | 5542-4941 |
| | 10369 | 33528-35574 |
| | 10376 | 5737-8521 |
| | 10758 | 41776-38993 |
| | 10765 | 58610-56353, 55103-54923 |
| | 10770 | 11540-14367 |
| | 10773 | 38494-36616, 35773-35425 |
| | 10792 | 25781-23831 |
| | 10918 | 8142-8284, 9122-9799, 10584-12359 |
| | 10924 | 27530-30351 |
| | 10946 | 8917-8941, 9824-12351 |
| | 11233 | 8461-5654 |
| | 11255 | 2735-5284 |

**B. *1360* sequences**

| Species | Contig | Coordinates |
|---|---|---|
| *D. melanogaster* | Chr 4 | 812470-809907 |
| | Chr 2L | 20145959-20144580 |
| | Hoppel-1 | Ref. 25 |
| | Hoppel-2 | Ref. 25 |
| | Hoppel Delta 5' | Ref. 25 |
| *D. simulans* | Chr 2L Random | 797960-799694 |
| | Chr 2L Random | 802845-804906 |
| | Chr 2R | 1199657-1199795 |
| | Chr 2R | 1200547-1201734 |
| | Chr 2R | 1208090-1206110 |
| | Chr 3L | 18143039-18143892 |
| | Chr U | 8188307-8185783 |
| *D. sechellia* | 3536 | 2255-2621 |
| | 9279 | 386-2615 |
| | 6826 | 2014-40 |
| | 11410 | 1640-11 |
| | 5259 | 2194-4726 |
| | 12180 | 2014-313 |
| | 5902 | 5125-2571 |
| | 3527 | 1-1564 |
| *D. erecta* | 7363 | 803868-801367 |
| | 6939 | 4861-5757 |
| | 7407 | 140279-137852 |
| | 7373 | 87864-86694 |
| | 7387 | 150284-149418 |
| | 7387 | 135352-137773 |
| | 7387 | 108989-106572 |
| | 6826 | 4906-5127 |
| *D. yakuba* | 260.3 | 23020-20458 |
| | 5.41 | 7616-10177 |
| | 0.40 | 345812-348373 |
| | 2.7 | 423893-421332 |
| *D. pseudoobscura* | 784 | 23339-25777 |
| | 1994 | 72290-72758 |
| | 4431 | 49805-50291 |
| | 520 | 17194-17712 |
| *D. persimilis* | 17644 | 428-816 |
| | 9857 | 66929-64479 |
| | 11446 | 2544-2069 |
| | 11871 | 3502-3042, 1903-1373 |

*(Continue on next page)*

| Species | Contig | Coordinates |
|---|---|---|
| *D. persimilis* | 14344 | 495-1631 |
| *D. virilis* | 17532 | 8869-8505 |
| | 13070 | 7536-6667 |
| | 15641 | 25683-27672, 28440-28823 |
| | 17537 | 36198-33865, 33361-33160 |
| | 4746 | 3288-4134 |

# 2.- DNA-binding properties of THAP-containing *Galileo* transposase.

Mar Marzo[1,2], Danxu Liu[2], Alfredo Ruiz[1], Ronald Chalmers[2]

[1] Departament de Genètica i Microbiologia, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain
[2] School of Biomedical Sciences, University of Nottingham, Queen's Medical Centre NG7 2UH, UK

E-mails: Mar Marzo – mar.marzo@uab.cat, Danxu Liu – danxu.liu@nottingham.ac.uk, Ronald Chalmers – chalmers@nottingham.ac.uk, Alfredo Ruiz – alfredo.ruiz@uab.cat

Corresponding author: alfredo.ruiz@uab.cat

Keywords: Transposable Element, *P-element*, *Galileo*, *Foldback*, THAP domain, TIR, DNA binding, transposase, reconstruction

## 2.1.- Abstract

Background: Transposable elements (TE) present huge variability in structure and transposition strategies. *Galileo* is a class II transposon involved in the generation of natural chromosomal inversions in Drosophila. It has been classified as a *P-element* superfamily thanks to the truncated transposase coding region found in the longest copies, although its long internally repetitive terminal inverted repeats (TIR) resemble the *foldback*-like type of TE. As repetitive sequences are a genomic instability source, the long *Galileo* TIR could affect the transposition reaction and/or have an active role in chromosomal rearrangements.

Results: In order to track possible effects of these long TIRs in the transposon mobilization, we tested the DNA binding activity, the first step of the transposition reaction. We inferred consensus and ancestor sequences for the DNA binding domain – THAP domain – of *Galileo* from three different species. We expressed these sequences and tested their binding activity showing specific DNA binding activity to the endmost part (150 bp) of the *Galileo* TIR. The DNA binding site was isolated and shared common traits with other THAP domains binding sequences. Furthermore, putative secondary binding sites were found in the tandem repeats of the TIR, which shed some light about why *Galileo* TIRs are so long. Finally an in vivo transposition experiment was carried out in *Drosophila* embryos where no transposition activity was detected.

Conclusions: *Galileo* THAP DNA binding domains were successfully reconstructed and expressed and showed specific binding activity. The length of the *Galileo* TIR seem to have tranposition role: provide secondary binding sites.

## 2.2.- Introduction

Transposable elements (TEs) are mobile genetic components of virtually all eukaryotic species (Feschotte & Pritham 2007; Wicker et al. 2007). These repetitive sequences make up a substantial proportion of most genomes and have a huge impact on the evolution of their hosts (Lander et al. 2001; Kidwell 2002; Kazazian 2004; Morgante 2006; Jurka et al. 2007). TEs are very diverse and employ many different mechanisms for mobilization. Two major groups are recognized depending on whether they use a retrotranscription step (retrotransposons or class I elements) or not (DNA transposons or class II elements) (Finnegan 1989). After this functional split TEs can be further grouped into subclasses, orders and superfamilies depending on their structure and sequence similarities (Feschotte & Pritham 2007; Jurka et al. 2007; Wicker et al. 2007). TIR transposons are recognized as an order of DNA transposons and characterised by their terminal inverted repeats (TIRs) of variable length. They encode a protein, called transposase (TPase), that catalyzes their mobilization by a "cut-and-paste" reaction. All TIR transposon families comprise autonomous and non-autonomous copies. Autonomous copies possess the capability of catalyzing their own transposition/movement. Non-autonomous copies contain internal deletions or point mutations in the transposase coding sequences that render them non-functional. These non-autonomous copies, which often outnumber their full-length counterparts, exploit the gene products of the autonomous copies (Feschotte & Pritham 2007).

The characterization of the different biochemical steps in the cut-and-paste reaction helps understanding how TIR transposons behave in the genome and make possible to recruit them as genetic tools. Since most of the transposon copies found in the genomes harbour mutations in the transposase coding region, rendering the encoded protein non functional, different strategies are used for inferring the possible functional sequences. Sometimes, a consensus sequence constructed from different genomic copies results in the restoring of the protein function (Ivics et al. 1997; Miskey et al. 2003; Sinzelle et al. 2008), but in other cases, because non functional sequences outnumbers the functional ones, the consensus results in a non functional sequence. For this reason, ancestor reconstruction is an alternative strategy that can be used for transposon recovery, where phylogenetic relationship among the sequences is taken in account for the putative

ancestral sequence deduction. This approach has successfully been used for the revival of different transposons, such as Hsmar1 (Miskey et al. 2007).

The *P-element* is one of the most intensively studied TEs. It was discovered in *Drosophila* melanogaster as the agent responsible for P-M hybrid dysgenesis (Rubin et al., 1982; Kidwell, 1985). It has since been studied in vivo and in vitro and is now widely used as a genetic engineering tool for genomic analysis of *D. melanogaster* (Rubin et al. 1985; Daniels et al. 1987; Spradling et al. 1995, 1999; Beall & Rio 1997; Rio 2002; Ryder & Russell 2003). The *P-element* defines a superfamily of TIR transposons, which includes 1360 and *Galileo* (see below). These elements harbour a transposase coding region surrounded by TIR, which are needed for the transposition reaction. The *P-element* transposase contains four functional domains: an N-terminal DNA binding domain, a coiled coil region involved in protein-protein interactions, a GTP binding domain and a catalytic domain with four acidic key residues (Rio 2002; Sabogal & Rio 2010). The *P-element* catalytic domain is thought to belong to the RNase H-like superfamily of polynucleotidyl transferases, although this remains uncertain because of the extreme divergence of its amino acid sequence (Rio 2002; Hickman et al. 2010; Sabogal & Rio 2010).

The cut-and-paste reaction of TIR transposons begins with the recognition and binding of the transposase to the transposon ends. The *P-element* transposase contains a THAP domain, which is responsible for site-specific DNA binding. The THAP domain is an evolutionary conserved motif shared by different animal proteins, including cell-cycle regulators, pro-apoptotic factors, transcriptional repressors and chromatin-associated proteins (Roussigne et al. 2003; Clouaire et al. 2005; Quesneville et al. 2005). The domain has a long zinc finger (~90 amino-acids) in which key residues are highly conserved (Roussigne et al. 2003). Recently, the THAP domain 3D-structure has been elucidated in two different proteins: the human THAP1 protein and the *D. melanogaster P-element* transposase (Campagne et al. 2010; Sabogal et al. 2010). The THAP domain interacts with its binding sequence in a bipartite manner, through the major and minor grooves of the DNA (Bessière et al. 2008; Campagne et al. 2010; Sabogal et al. 2010).

The *Galileo* transposon was discovered in *Drosophila buzzatii*, where it has recently caused three large chromosomal inversions (Cáceres et al. 1999; Casals et al. 2003; Delprat et al. 2009). Although originally considered a *Foldback*-like element, it was later included in the *P-element* superfamily of cut-and-paste transposons based on the sequence of the putative transposase (Marzo et al. 2008). *Galileo* is probably widespread within the *Drosophila* genus because it has been found in species of the two subgenera of *Sophophora* and *Drosophila* (Marzo et al. 2008). Many incomplete (non-autonomous) copies of *Galileo* have been detected in all species tested and in some cases two or more *Galileo* subfamilies have been found coexisting in the same genome (Figure 2.1). For instance, three subfamilies are present in *D. buzzatii* (G, K and N for *Galileo*, Kepler and Newton), while *D. mojavensis* harbours four subfamililes (C, D, E and F) (Marzo et al. 2008; Delprat et al. 2009). To date no potentially active copies of the transposon have been found because they all harbour premature stop codons and/or frameshifts. Nevertheless, consensus sequences present putative ORFs which harbours the main domains of the *P-element* transposase.

The most conspicuous features of *Galileo* are the 0.5 to 1.2 kb long TIRs which. This is considerably longer than other members of the *P-element* superfamily, in which the TIRs are 31 bp long. Indeed, it was the extreme length of *Galileo* TIRs that defined it as a 'foldback' family of transposons before they were recognized as members of the *P-element* superfamily. *Galileo* TIRs have another interesting property: namely, that the sequence conservation between elements in different species is restricted to the outer ~40 bp (Marzo et al. 2008). One obvious possibility is that these regions are functional transposition sequences, and would be the equivalent of the short TIRs of the *P-element*. If true, this leaves the function of the remaining 0.5 to 1.2 kb open to question. The fact that they are not conserved between elements in different species, and that they sometimes contain internal tandem repeats, suggests that secondary structure of the DNA may play a role in transposition. The mechanism of *Galileo* transposition may therefore prove to be of considerable interest, and may explain the frequency with which this element is able to generate chromosomal inversions in Drosophila. In the present work we have focused on the reconstruction of an active transposase and its binding to the inverted repeat. Although we have not succeeded in a full reconstitution of the transposition reaction, we have detected transposase binding to the extremities of

*Galileo* and putative secondary binding sites in the tandem repeats of the TIR. This represents the first steps in the characterization of *Galileo* recombination. Further characterization promises to reveal fascinating details of the interactions between this transposon and its host and perhaps even the reason it promotes chromosomal inversions so frequently.

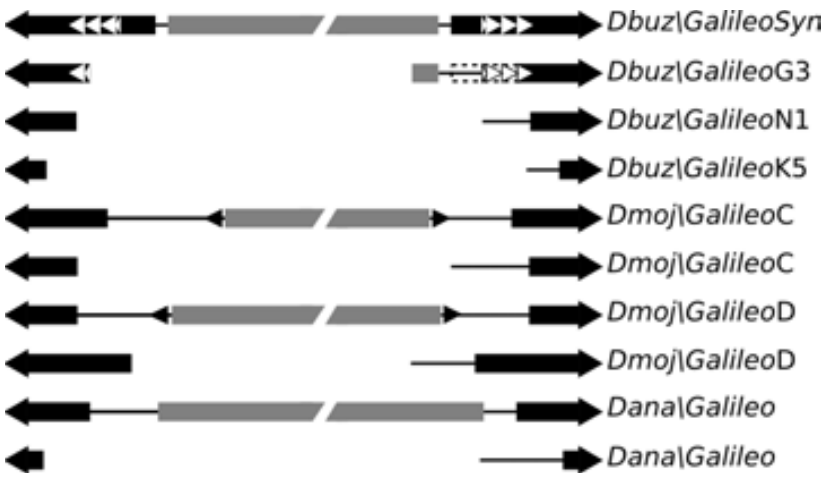



**Figure 2.1**. Structure of representative *Galileo* copies found in the species of *Drosophila* used in this work. Black arrows are the Terminal Inverted Repeats (TIR) of each element and white triangles are internal tandem repeats. Gray rectangles are the transposase coding regions and black arrowheads are internal inverted repeats found in some *D. mojavensis* copies. No copies harbour an intact ORF. Dbuz*Galileo*Syn (constructed copy) and *D. mojavensis* (contigs: 10758, 9847, 9930 and 11679) and *D. ananassae* (contigs 15556 and 16052) copies are from Marzo et al 2008. *Dbuz**Galileo*N1 and *Dbuz**Galileo*K5 are *Newton*1 and *Kepler*5 elements from Casals et al 2005..

## 2.3.- Results

*Galileo* sequence reconstruction. We generated four different consensus sequences: one using multi-strain PCR amplification sequences of the Dbuz*Galileo* whole transposase, and the other three using the THAP domains from genomic sequences of Dmoj*Galileo* subfamilies C and D and D. ananassae. These sequences showed a few differences when compared with previous studies (Marzo et al. 2008). Thus, the consensus sequence of the Dbuz*Galileo* transposase no longer contains premature stop codons, and presented two amino-acid changes. Likewise, the THAP domain sequence obtained for the Dmoj*Galileo*C was identical to the previously published, and the sequences for Dmoj*Galileo*D and Dana|*Galileo* had two and one amino-acid changes, respectively. Additionally, we also reconstructed the ancestral sequences of the THAP domains by maximum likelihood. When the inferred ancestor and consensus pair of sequences were compared, three, two and three differences were found in D. ananassae, the Dmoj*Galileo* subfamilies C and D, respectively (Figure 2.2). Although one of the amino-acid changes affected one of the key residues of the domain, it was a functionally similar amino-acid replacement (a Valine replaced by an Isoleucine). The comparison of the reconstructed sequences of the *Galileo* DNA binding domains with those of the *P-element* of *D. melanogaster* and the human THAP1 protein showed that the structural key residues of the THAP domain are conserved (Figure 2.2). However, the THAP domains of *Galileo* showed a longer and more variable N-terminus, along with a shorter and highly conserved loop 4 (L4).

Testing the first step of the transposition reaction: DNA binding activity. The reconstructed amino-acid sequences (ancestor and/or consensus) of the different *Galileo* THAP domains were E. coli codon-optimised, chemically synthesised and, finally, cloned into protein expression vectors. Seven THAP proteins were obtained *D. buzzatii* (two proteins of 90 and 150 amino acid length), the consensus of D. ananassae (ancestral sequence could not be purified), the consensus and the ancestral reconstruction of Dmoj*Galileo* C and D subfamilies(Figure 2.3A). Electrophoretic mobility shift assays (EMSA) were performed for each of the seven proteins with their cognate labelled TIR sequence (150 bp endmost portion). Different conditions for the assay were used: three different protein concentrations, presence/absence of ZnCl2 and

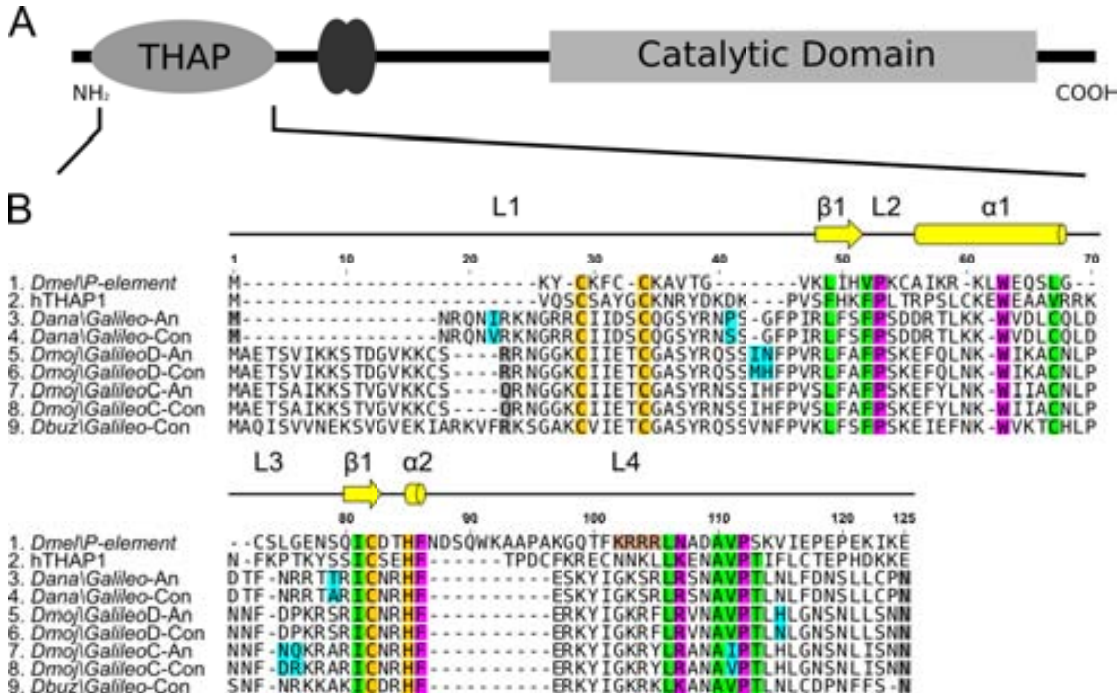



**Figure 2.2**. THAP domain protein sequences. A) Domain structure of the *Galileo* transposase: the THAP domain is the DNA binding domain, the coiled coil region is responsible of protein-protein interactions (represented as two overlapping circles) and the catalytic domain is located in C-terminal region. B) Alignment of the consensus and ancestral *Galileo* THAP domain sequences with the THAP domain of the *P-element* TPase (*D. melanogaster*) and THAP1 protein (Homo sapiens). The predicted secondary structures are shown above the alignment (adapted from (Bessière et al. 2008) and (Sabogal et al. 2010)): yellow arrows represent β sheets and yellow cylinders are α helix regions. Key residues are coloured: zinc coordination residues (C2CH) in yellow, conserved hydrophobic residues in green, invariant residues in pink, nuclear localization signal (NLS) in light brown. Segments cloned for protein expression are between grey shaded residues. Residues coloured in cyan are the amino-acid changes between ancestor and consensus sequences.

addition of unspecific DNA competitor (pBlueScript). Similar results were obtained for the seven proteins, but only the results for Dbuz*Galileo* THAP are shown (Figure 2.3B). These assays showed specific binding activity to the TIR independently of the addition of ZnCl2 or pBS to the reaction for all the tested THAP proteins. Furthermore, when an EMSA was performed with the same TIR and the respective ancestor and consensus proteins, no qualitative differences in binding activity were detected (Figure 3C). It is noteworthy that some extra shifted bands appeared with the highest protein concentrations (Figure 2.3 B and C). Thus, a fine titration was carried out with the Dbuz*Galileo*G-THAP-90 amino acid domain (Figure 2.3D). The results showed that the second and subsequent shifted bands are concentration-dependent, probably due to protein aggregation.

To test if a transposase would be able to bind or transpose different families or subfamilies of *Galileo* transposons that coexist in the same genome, we performed a cross-binding EMSA with Dbuz-THAP-protein with the 3 TIR sequences from this

genome (G, K and N)(Figure 2.4A). We observed that the Dbuz*Galileo*G-THAP domain binds both the Dbuz*Galileo*G TIR and the Dbuz*Galileo*K TIR, although binding is weaker in the last case. However, no trace of binding activity was found with the Dbuz*Galileo*N TIR. In this experiment, the size of the THAP domain (90 or 150 amino acid) did not show a qualitative effect on binding activity. Likewise, when we tested the 90 amino-acids protein of *D. buzzatii* against all the TIRs used in this work (*D. buzzatii* (G, K and N), *D. mojavensis* (C and D) and D. ananassae) a weak binding activity was observed in Dana*Galileo* TIR along with Dbuz*Galileo*G and Dbuz*Galileo*K binding (Figure 2.4B).

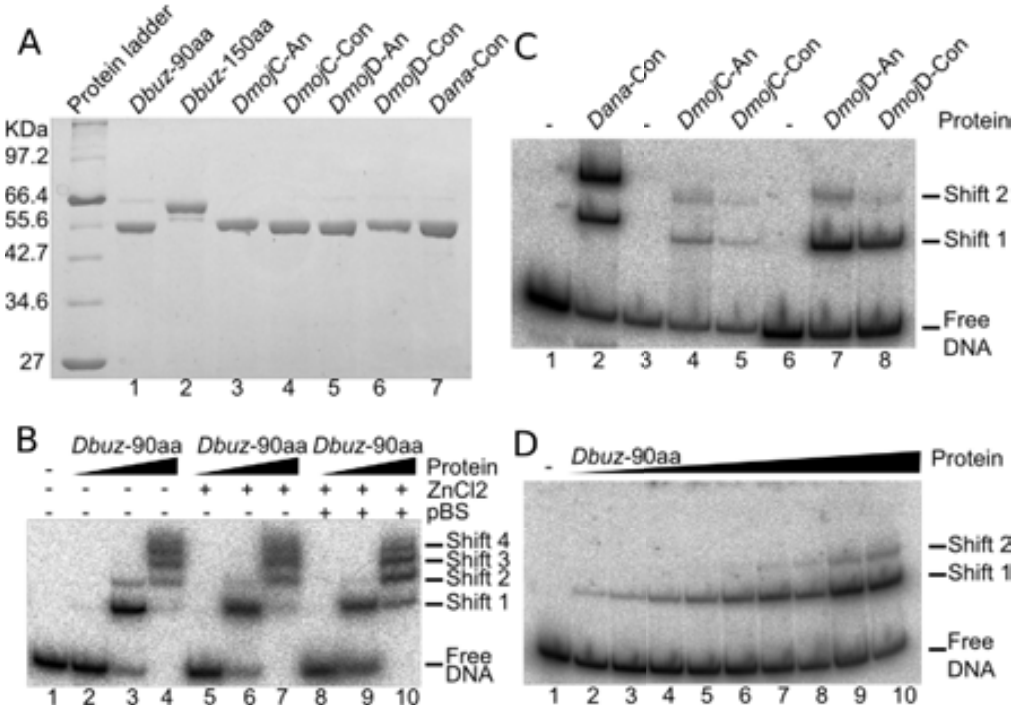



**Figure 2.3**. Protein assays. A) SDS-PAGE with the 7 expressed THAP domain proteins, ~5 µg protein/well. 1. Dbuz*Galileo*-THAP-90aa, 2. Dbuz*Galileo*-THAP-150aa, 3. Dmoj*Galileo*C-THAP-Ancestor, 4. Dmoj*Galileo*C-THAP-Consensus, 5. Dmoj*Galileo*D-THAP-Ancestor, 6. Dmoj*Galileo*D-THAP-Consensus and 7. Dana*Galileo*-THAP-Consensus. B) EMSA performed with Dbuz*Galileo*-THAP-90aa. Three different binding conditions were tested. First lane is Dbuz*Galileo*G labelled TIR (2.2 nM). Lanes 2, 3 and 4 are x100 increasing protein concentrations (470pM, 47nM and 4.7µM). Lanes 5, 6 and 7 are the same protein conditions as the previous lanes but 100µM ZnCl2 reaction condition was added to the binding reaction. Lanes 8, 9 and 10 are the same conditions as in the previous 3 lanes but 500ng of pBlueScript (Stratagene) plasmid was added as an unspecific DNA competitor. C) EMSA assay where Dana*Galileo*-THAP-Consensus (lane 2), Dmoj*Galileo*C-THAP-Ancestor (lane 4), Dmoj*Galileo*C-THAP-Consensus (lane 5), Dmoj*Galileo*D-THAP-Ancestor (lane 7), Dmoj*Galileo*D-THAP-Consensus (lane 9) have been tested to bind the consensus TIR of their *Galileo* subfamily. All the THAP domains bind their TIR DNA (final protein concentration: ~5.87 nM and TIR final concentration ~0.28nM). D) Fine titration EMSA of the Dbuz*Galileo*-THAP-90aa with its TIR (0.14nM). Protein concentrations (2 fold dilutions from 1/128 to 2X range): 0.184nM, 0.367 nM, 0.734 nM, 1.469 nM, 2.938 nM, 5.875 nM, 11.75 nM, 23.5 nM, 47 nM and 94 nM. A concentration dependence of the extra shifted bands can be appreciated.

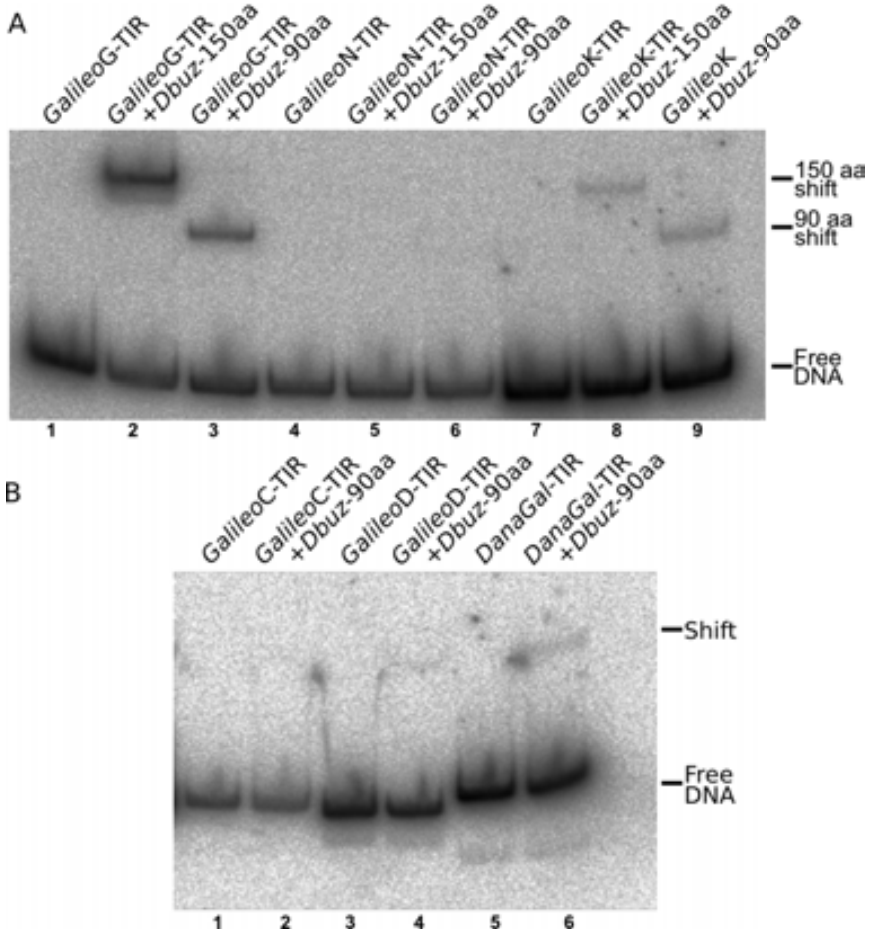



**Figure 2.4**. Cross binding EMSA experiments. A) Dbuz*Galileo*-THAP-90aa and Dbuz*Galileo*-THAP-150aa versus different *Galileo* TIRs from D. *buzzatii*. Lanes: 1. Dbuz*Galileo*G-TIR, 2. Dbuz*Galileo*G-TIR and Dbuz*Galileo*-THAP-150aa, 3. Dbuz*Galileo*G-TIR and Dbuz*Galileo*-THAP-90aa, 4. Dbuz*Galileo*N-TIR, 5. Dbuz*Galileo*N-TIR and Dbuz*Galileo*-THAP-150aa, 6. Dbuz*Galileo*N-TIR and Dbuz*Galileo*-THAP-90aa, 7. Dbuz*Galileo*K-TIR, 8. Dbuz*Galileo*K-TIR and Dbuz*Galileo*-THAP-150aa, 9. Dbuz*Galileo*K-TIR and Dbuz*Galileo*-THAP-90aa (final protein concentration: ~5.87 nM and TIR final concentration ~0.28nM). B) Dbuz*Galileo*-THAP-90aa against Dbuz*Galileo*G-TIR (lane 2), Dbuz*Galileo*N-TIR (lane 4), Dbuz*Galileo*K-TIR (lane 6), Dmoj*Galileo*C-TIR (lane 8), Dmoj*Galileo*D-TIR (lane 10), Dana*Galileo* TIR (lane 12).

DNA binding site of *Galileo.* We performed a DNase I footprinting analysis to determine the Dbuz*Galileo*G TIR binding site sequence (Figure 2.5). The protected region covers a continuous region of 18 bp from nucleotide +63 to +80 bp of the tested 150 bp sequence. The second shifted band seen in the EMSA was footprinted as well (Figure 2.5). There is no difference in the protection pattern, so the multiple shifted bands are due to protein aggregation in the same TIR location which is in agreement with the titration experiment.

The comparison of this 18 bp sequence with other THAP binding sites is shown in Figure 2.6. The *Dbuz\Galileo*G binding site is almost twice as long as the *P-element* and

THAP1 binding sites (18 bp versus 11bp). Nonetheless, based on similarities with the interaction sites of the *P-element* and THAP1, we propose that the putative major and minor groove sites are the GGGGT region and the upstream T, respectively (Figure 2.6). When we compared the binding sequence of Dbuz*Galileo*G with the homologous regions of the Dbuz*Galileo*K and Dbuz*Galileo*N TIRs, we observed that they are poorly conserved (not shown). This could explain the weak binding to *Galileo*K TIR and the absence of binding to *Galileo*N TIR.

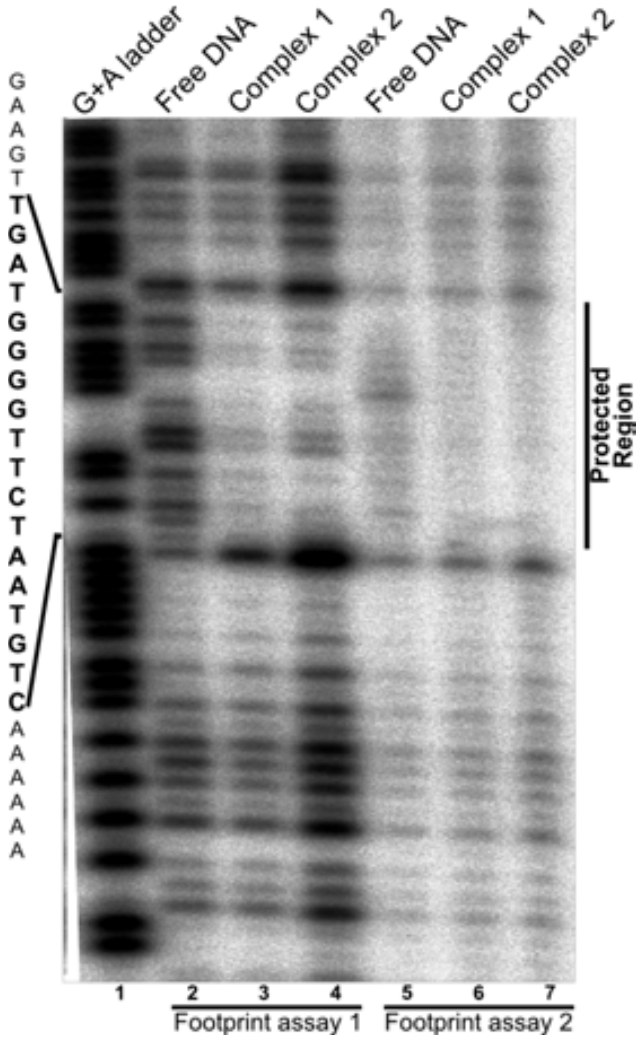



**Figure 2.5**. Sequence specific binding of THAP domain to *Galileo* element. The DNaseI footprints of the indicated complexes were performed as described in material and method. The footprints were resolved on a DNA sequencing gel and the radioactive signals were recorded on a phosphoimager. Lane 1, G+A ladder; Lane 2, Free DNA treated with DNaseI; Lane 3 and lane 5, footprints of complex 1; Lane 4 and Lane 6, footprints of complex 2; Lane 7, footprint of complex 3. The protected DNA sequence was shown on the left of the gel.

The *Dbuz**Galileo*G TIR is up to 1.2 kb long, partially due to the presence of three (and a partial fourth) internal tandem repeats. For that reason, we searched within the TIR for sequences similar to the identified binding site using Blast-2-sequences program (Altschul et al. 1997)), and found two significantly similar sequences located in the first two tandem repeats (E-values = 5x10-5 and 7x10-4, respectively). A comparison of the three binding sites located in the *Galileo*G TIR showed that the three sequences are very similar, particularly around the proposed GGGGT major groove region (Figure 2.6B). Although we did not test these internal sequences for binding activity with the THAP domain, the high sequence similarity with the identified binding site suggests that they might act as additional binding sites.

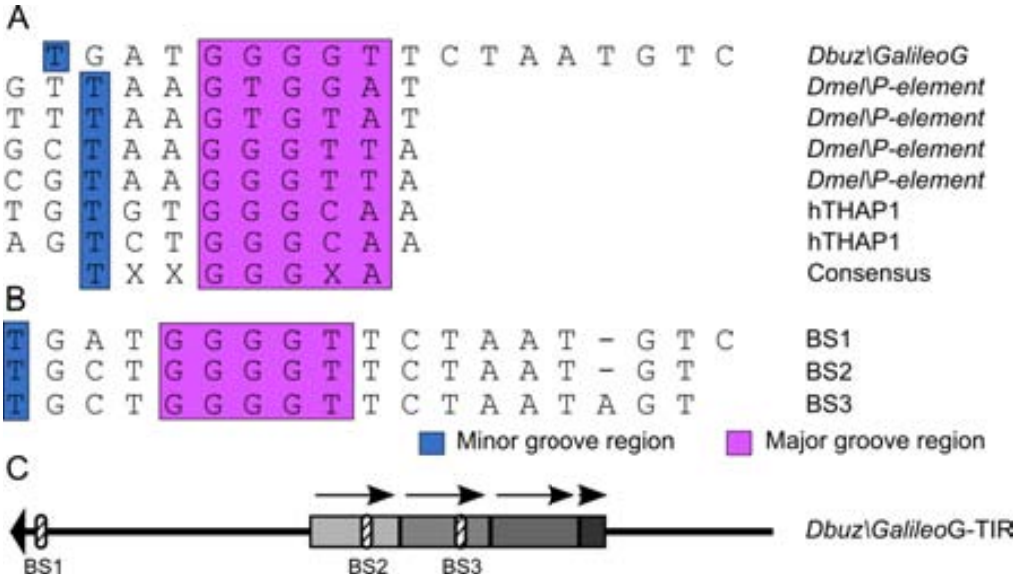



**Figure 2.6**. THAP domain binding sequence comparison. A) Dbuz*Galileo*G compared to Dmel*P-element* (Sabogal et al. 2010) and hTHAP1(Bessière et al. 2008; Campagne et al. 2010) binding sites. The major and minor groove interacting regions are coloured. A putative consensus THAP binding sequences, including Dbuz*Galileo*G sequences has been proposed. This consensus is in agreement with the previously proposed by (Sabogal et al. 2010). B) Alignment of the Dbuz*Galileo*G binding site with other putative binding sites found downstream in the Dbuz*Galileo*G-TIR. C) Structure of the Dbuz*Galileo*G-TIR where the tandem repeats are drawn as grey rectangles and the binding sites are drawn as *white* stripped rectangles (BS1, BS2 and BS3).

*Galileo* in vivo transposition. We performed an in vivo experiment to test whether the consensus whole transposase from *D. buzzatii* was fully functional. To this end, we adapted the *Drosophila P-element*-based general transformation vectors to test for *Galileo* activity in *Drosophila* melanogaster *white* strain. These vectors consisted in a helper plasmid where the transposase was cloned after a Hsp70 promoter, and a donor plasmid where a reporter gene (mini-*white* gene in this case) was cloned surrounded by the transposon TIRs. If the transposase is active, when these two plasmids are injected

into *white* (w-) *Drosophila* embryos, the enzyme will insert the mini-*white* gene in the precursors of the germinal cell line. Then, the crossing of injected individuals with non-injected w- adultss enables the detection of the transposition activity by screening the F1 generation for red eyes.

In our experiment we performed three different injections: i) one using the general *P-element* transformation vectors as a positive control, ii) a second one using these *P-element* vectors with the original transposon sequences replaced by *Galileo* sequences (the whole *D. buzzatii* consensus sequence of *Galileo* transposase in the helper plasmid and 150 bp of *Galileo* TIR in the donor plasmid), and iii) a third injection with the *Galileo* donor plasmid but without the *Galileo* helper plasmid as a transposition negative control. The injection-surviving adults were crossed with *D. melanogaster white* (w-) individuals. The offspring of these crosses was screened for transformed flies by observing the eyes pigmentation. In the positive control, transposition events were detected in 19 of 91 of the crosses (384 flies with red eyes of 26637 F1 screened flies). As expected, the negative control did not show any transformant (96 crosses, 31201 F1 screened flies), discarding the spontaneous insertion of the mini*white* gene. Finally, when the offspring from *Galileo* sequences injection was screened, no transgenic individuals were found (99 crosses, 32537 F1 screened flies).

## 2.4.- Discussion

TIR transposons encode a transposase that is required for their mobilisation by a cut-and-paste reaction. However, most of the transposon copies found in the genomes harbour mutations in the coding regions that render non-functional proteins. The revival of these proteins allows studying how the transposition processes take place in real time. Different strategies can be used for inferring the original functional sequences of these transposons. Probably, the simplest approach is the construction of a consensus sequence using different transposon copies from the genome. Alternatively, a more sophisticated method that can be used consists in the reconstruction of ancestral sequences under a model of evolution by maximum-likelihood methods. These two approaches have successfully been used for the revival of several different transposons, such as Sleeping Beauty, Frog Prince, Hsmar1 and Harbinger (Ivics et al. 1997; Miskey et al. 2003, 2007; Sinzelle et al. 2008).

The transposon *Galileo* has been recently active in the genome of *D. buzzatii* (Delprat et al. 2009) and perhaps other species (Marzo et al. 2008). However, all *Galileo* copies found so far are not functional and we used both approaches to reconstruct the DNA binding domain. The ancestrally reconstructed and consensus sequences showed few differences which did not involve the domain key residues responsible of stabilising the hydrophobic core of the protein (Sabogal et al. 2010). When we compared these sequences with the homologues of other THAP domains, we found that the most divergent regions were the N-terminus and the Loop 4. The N-terminus was longer and more variable in *Galileo*, with a length ranging from 12 to 28 residues instead of the 2 to 5 residues found in other THAP domains. However, the Loop 4 was very conserved in all *Galileo* copies. This differentiation is in agreement with the binding-specificity role proposed for these two regions in *P-element* and hTHAP1 after the analysis of their tridimensional structure by X-Ray diffraction and NMR (Campagne et al. 2010; Sabogal et al. 2010).

We detected similar strength and specificity in the binding activity for sequences inferred by both strategies, at least qualitatively. Moreover, we detected some cross-binding where a *Galileo* THAP domain have been able to recognise and bind some TIR from different transposon subfamilies. This would be in agreement with the fact that, in

some cases, elements that do not own a transposase take advantage from functional transposons and use their transposition machinery. This is a general behavior found in different TE groups, for example SINEs parasitise LINEs and MITEs parasitise some class II elements, (Jurka et al. 2007; Wicker et al. 2007; Yang et al. 2009). If a transposition reaction would be set up, it could be tested that *Galileo* elements also suffer from its own parasites (Gonzalez & Petrov 2009). In addition, although multiple shifted bands were observed in the EMSA, we ruled out the possibility of the existence of multiple binding sites in the 150 bp tested TIR region by means of a titration experiment and a footprint assay, leaving the aggregation of proteins as the only plausible explanation for our observations.

The isolated binding site of *Galileo* is almost twice as long as other THAP target sequences. This might be explained by the larger size of the protein due to the existence of an insertion of 16 amino-acids after the initial methionine, which seems important for the interaction with the binding site (Sabogal et al. 2010). However, we cannot discard that this length could be an experimental artefact due to steric hindrance between the large protein-expression tag MBP and the DNase I enzyme used in the assay. Despite this noticeable difference in length, the *Galileo* binding site does present regions homologous to the major and minor grooves interacting zones of DNA that have been found to be essential for the recognition by the THAP domains of other proteins (Campagne et al. 2010; Sabogal et al. 2010).

The location of the binding sites is strikingly similar in *Galileo* and the *P-element*, This way, 61 and 50 bp from each transposon end in the *P-element*, and at 63 bp from both transposon ends in *Galileo*. In contrast with the *P-element*, the binding sites of *Galileo* are located within its long TIRs. When we extended the comparisons to the whole TIRs of *P-element* and *Galileo*, we found profound differences in length and structure. Thus, whereas *P-element* TIR is a non-repetitive region of 31bp length, the TIR of *Galileo* comprises up to 1.2 kb and harbours several internal tandem repeats. It is peculiar that although the part of the TIRs of *Galileo* involved in the binding recognition did not show any conservation, we found that the endmost region is highly conserved across different species. This suggests that this region may have a role in the catalytic step of the transposition reaction, in a similar way to the short TIR of the *P-element* (Rio 2002).

The existence of secondary binding sites or transposition enhancers has been reported in different transposons and these sequences can be part of the TIR or not. For example, *P-element* has subterminal transposition enhancers located outside the short TIR (Rio 2002), whereas the secondary binding sites of Sleeping Beauty and Bari-like elements lie within the long TIRs in the form of tandem repeats (Ivics et al. 1997; Moschetti et al. 2008). A similar structure has been found in Tnr8 and Phantom elements, although if their tandem repeats act as binding sites remains untested (Cheng et al. 2000; Marquez & Pritham 2010). Although evolutionary unrelated, *Galileo* is structurally more similar to these elements, where their secondary binding sites are found as tandem repeats. All these TIR elements have a considerable size, which is a trait negatively correlated with the efficiency of the transposition reaction (Atkinson & Chalmers 2010). Therefore, the presence of multiple binding sites may constitute an evolutionary convergent strategy to overcome length limitation by successfully recruiting the transposase and enhancing the transposition process. In fact, this strategy has been already applied to artificially improve transposition reactions (Zayed et al. 2004).

Finally, we carried out an in vivo transposition experiment to test if consensus Dbuz*Galileo* transposase was functional. After screening for transformants, we were not able to detect transposition activity. As we do not know the *Galileo* transposition frequency, this result could be due to a very low transposition rate that would need a bigger sampling for transformants (e.g. at least ~106.individuals must have been screened for a 10-6 transposition rate). But, if we assume that *Galileo* transposition rate could be similar to the *P-element*, our positive control in the experiment, some *Galileo* transformants must have been found. So, there may be other reasons responsible for the negative result, such as: the lack of secondary binding sites in the donor construct, the consensus transposase might not be functional or might be toxic for the flies, or, as the tested transposon comes from *D. buzzatii*, there may be missing specific cellular factor or unknown incompatibilities that do not allow *Galileo* to mobilize in *D. melanogaster*. These two flies are distantly related as they belong to two different lineages that split 40-60 million years ago (Russo et al. 1995; Tamura et al. 2004). Further studies could shed some light in this issue.

## 2.5.- Conclusions

This work constitutes the first step in the characterization of the transposition reaction of *Galileo*. Since *Galileo* copies are non-functional in the genomes of *Drosophila* species, we had to reconstruct functional sequences. Although we were not able to detect a whole transposition reaction with these revived candidates in an in vivo experiment, we confirmed that they can recognise and interact with DNA in vitro. Furthermore, we found that even though the isolated *Galileo* binding sequence is longer than in any other THAP domains, the recognised binding sites are homologous to those of other proteins. We also detected the presence of putative secondary binding sites in the TIR internal tandem repeats. The confirmation of these regions as functional binding sites would provide the first evidence of the convergent evolution of this mechanism to overcome the drawbacks caused by increased TIR length.

## 2.6.- Materials and methods

Amplification of *D. buzzatii Galileo* transposase coding sequence by PCR. Three overlapping regions, that spanning the whole transposase coding sequences were PCR amplified in eight *D. buzzatii* strains (st-1, Maz-4, j-9, jq7-4, jz3-2, jq7-1, Sar-9 and j-4). These PCRs were carried out in a total volume of 25 µl including 100-200 ng of genomic DNA, 20 pmol of each primer, 200 µM dNTPs, 1.5 mM MgCl2 and 1-1.5 units of Taq DNA polymerase. The products were gel-purified and sequenced.

Generation of THAP domain sequences. A consensus sequence of the Dbuz*Galileo* transposase segment was generated with the PCR products using the majority rule (Geneious assembly algorithm in Geneious (Drummond et al. 2010)). This consensus sequence differs from the reported Dbuz*Galileo* sequence (Marzo et al. 2008) by 5 nucleotides and can be translated into a fully functional protein. The THAP domain region of the consensus sequence is located in the N-terminal 450 bp portion.

Consensus sequences were also generated for D. ananassae and *D. mojavensis* transposase sequences. The sequences found in these genomes in previous work (SI Table 2.1) were aligned with MUSCLE 4.8.4 algorithm (Edgar 2004) implemented in Geneious software (Drummond et al. 2010) and a majority rule consensus of the THAP domain was generated (450 bp). As described in our previous work, there are four different *Galileo* subfamilies (C-F) in *D. mojavensis* (Marzo et al. 2008). Here, we generated transposase consensus sequences for the *Galileo*C and *Galileo*D subfamilies.

Finally, a reconstruction of the 450 bp ancestral THAP domains was carried out for *D. ananassae* and *D. mojavensis* (C and D subfamilies). MUSCLE 4.8.4 (Edgar 2004) alignments were used for generating the best trees by maximum likelihood using RAxML phylogenetic software and GTR+gamma evolution model (Stamatakis 2006). The trees were rooted with an appropriate outgroup using FigTree 1.3.1 software (Rambaut 2006) and after rooting, the outgroup was removed from the tree manually. These rooted phylogenetic trees and the alignments were used for inferring the ancestral sequence by maximum likelihood using the CODEML module in PAML software (Yang, 1997) (parameters: seqtype= 1 (codons); codonfreq=2; NSsites = 0 1; rateancestor=1; fix_blength= 1).

TIR cloning. In order to test the DNA binding ability of the *Galileo* THAP domains, 150 bp TIR consensus sequence was generated for *Galileo* elements in *D. buzzatii* (*Galileo*G, *Galileo*N and *Galileo*K subfamilies), *D. mojavensis* (*Galileo*C and *Galileo*D subfamilies) and D. ananassae. These consensus sequences were generated using the majority rule, as above. A genetic construct (pRC1525) was created concatenating the inferred sequences plus *Galileo* representative target site duplications. Unique restriction sites were located in between each TIR for releasing them individually from the vector and allowing radioactive dCTP labelling using an exo- Klenow polymerase.

THAP Protein expression. The inferred ancestral and consensus 450 bp sequences were codon optimized and synthesized (Bio S and T Inc., Canada). From these sequences a 270 bp (90 amino acid) predicted core THAP domain was PCR amplified (Phusion enzyme) and cloned in pOPINM (N-ter MBP-tag vector from The Oxford Protein Production Facility, UK) using the In-Fusion® cloning technology (Clontech Inc.). In the *D. buzzatii* case, as no ancestral sequence was reconstructed the 450 bp THAP sequence (150 amino acid) was cloned in pOPINM expression vector as well. The effect of the THAP domain length could be tested this way. The expression vectors with the THAP domains were sequenced for verifying the ORF and were transformed in BL21 (DE3) E. coli expression cell line. The protein expression was induced in DO680 =0.5 LB cultures with 100 ug/ml ampicillin cultures, 1mM of IPTG and 100uM of ZnCl2 at 16ºC over night. The cells were harvested by centrifugation and resuspended in HSG buffer (50mM HEPES pH 7.5, 200mM NaCl, 2mM dithiothreitol (DTT), 5mM EDTA and 10% glycerol). The cells were lysed in a French press and centrifuged at 25000g for 30 min. The supernatant was loaded onto an amylose resin column (New England Biolabs). The column was washed several times with HSG buffer and the protein eluted with HGS buffer plus 10mM maltose. The fractions containing MBP transposase were pooled and aliquots were stored at -80ºC.

Electrophoresis mobility shift assay (EMSA). Purified recombinant THAP domains were incubated for 2 hours at room temperature with the labelled TIR in 20 ul reaction of binding buffer (20 mM Tris-HCl, pH 7.5, 100 mM KCl, 100 g/ml bovine serum albumin, 2.5 mM DTT, 5% glycerol). Different conditions were tested: different protein concentration (1, 1:100, 1:10000 from the stock protein solution (5ug/µL or 94 µM), addition of ZnCl2 (100 µM final concentration) and addition of unspecific competitor

DNA (pBlueScript, ~500ng/reaction). The reactions were loaded in a 4% TAE-polyacrilamide gel and run for 2 hours at 300V at 4ºC.

Footprint assay. A sample of the EMSA reaction was digested by 0.05U of DNase I for 1 minute at room temperature. The enzyme was diluted to 1U/μL with dilution buffer (5 mM MgCl2, 0.5 mM CaCl2). The reaction was stopped using 1 μL of 500 mM EDTA. DNA was purified by phenol-chloroform extraction and ethanol precipitation. The cleavage pattern was analysed by electrophoresis on a 5% polyacrylamide sequencing gel. DMS/piperidin reactions were performed following standard procedures to reveal G positions and were used to localize the DNase I protected regions.

*In vivo Galileo* transposition experiment. Plasmids generation. Helper plasmid: pTURBO-*Galileo* (pRC1510). The inferred Dbuz*Galileo* consensus transposase ORF (see above) was generated by directed mutagenesis PCR (see primers in SI Table 2.2). The different PCR fragments were assembled thanks to the addition of unique silent restriction sites at each end. This consensus ORF was cloned in the pTURBO (pUChsΔ2-3, FlyBase recombinant construct *FB*mc0000938, (pRC1501)) plasmid replacing the *P-element* transposase. For this purpose, a PCR of whole pTURBO sequence except the *P-element* ORF was performed and two unique restriction sites (MluI and EagI) were added for cloning the *Galileo* transposase. After cloning the ORF was sequenced to check that the coding sequence was the proper one.

Donor plasmid. pCASPER-*Galileo* (pRC1517). The plasmid pCaSpeR-4 (FlyBase recombinant construct *FB*mc0000178, (pRC1502)) was used as donor plasmid. Two PCRs were performed for amplifying and ligating all the plasmid without the *P-element* sequences. In this step 4 unique restriction sites were added (PstI, NotI, NsiI and BamHI) surrounding the mini*white* gene. These 4 unique restriction sites were used for cloning 150-pb *Galileo* TIR in the proper orientation and TSD, surrounding the mini*white* gene (TIR1: PstI and NotI, TIR2: NsiI and BamHI). The mini*white* ORF and the TIR were sequenced for checking the sequence. The PCRs carried out in this section were performed with Phusion polymerase (Finnzymes).

*Drosophila* injections. 3 different injections were performed in *Drosophila* melanogaster *white* embryos (strain w1118, Genetic Services Inc. USA): one with the *P-element* plasmids without any change as a positive control (pRC1501 -helper- and

pRC1502 -donor-), another one with the two *Galileo* generated plasmids (pRC1510 -helper- and pRC1517 -donor-) and the last one with pRC1517 alone as a negative control. Each injected fly (91 positive controls, 99 *Galileo* transposition elements and 96 negative control) was crossed with three virgin females or three males depending on their gender. The tubes of the crosses with *Drosophila* media were changed every two days (in the case of one injected male with 3 virgin females) or every 4 days (in the case of one injected female with 3 males) during 12 to maximise the number of offspring. Finally the F1 of each cross was counted and non-*white* eyes were screened (from light orange to deep red eyes) as a marker of transposition activity.

## 2.7.- Supplementary material

Supporting tables list:

SI Table 2.1

SI Table 2.2

SI Table **2.1**. Sequences used for inferring the THAP domain sequences. CAF1 assemblies.

| Species/Group | | Coordinates |
|---|---|---|
| *D. mojavensis* C | scaffold_6262 | 13889-19752 |
| | scaffold_6541 | 1141978-1149130 |
| | scaffold_6500 | 31288762-312953303 |
| | scaffold_6358 | 1-5345 |
| | scaffold_6500 | 31981325-31980812 |
| *D. mojavensis* D | scaffold_6500 | 31458921-31464785 |
| | scaffold_6482 | 614003-617184 |
| | scaffold_6482 | 617185-618411 |
| | scaffold_6485 | 39163-45738 |
| | scaffold_6540 | 1175880-1182997 |
| *D. ananassae* | contig_15979 | 71824-74395 |
| | contig_11169 | 1-2142 |
| | contig_19410 | 7756-12565 |
| | scaffold_13082 | 2449985-2467038 |

SI Table **2.2**. Primers used in this work

| Name | Sequence (5'-3') | Template |
|---|---|---|
| TIR1_PstI-TSD-F | GACAGTCTGCAGGTGATAGCACTAACCATACAACACATAGACTG | pGPE_Dbuz*Galileo* TIR |
| TIR1_NotI_150bp-R | GTGACTGAGCGGCCGCCGGAATGATTTTGTCATCA | pGPE_Dbuz*Galileo* TIR |
| TIR2_BamHI-TSD-R | CTATGTGGATCCGTGATAGCACTAACCATACAACACATAGACTG | pGPE_Dbuz*Galileo* TIR |
| TIR2_NsiI_150bp-F | GTGACTGAATGCATCGGAATGATTTTGTCATCA | pGPE_Dbuz*Galileo* TIR |
| Dbuz_TPase_EagI_Met-F | GATCTACGGCCGAAAATGGCGCAAATAAGTGTTGTG | Consensus TPase |
| Dbuz_TPase_end-MluI-R | GACGAAACGCGTTATTTTTATTCACGAATCATTTTCAGTTTACTTTTAC | Consensus TPase |
| pTURBO-EagI-R | CTAGATCGGCCGTTTATTCCACGTAAGGGTTAATG | pTURBO |
| pTURBO-MluI-F | CTTCGTACGCGTGAGTTAATTCAAACCCCACG | pTURBO |
| Bb_pCAS-BamHI-F | TCTGATGGATCCGCAAGGAGTAGCCGACATATATC | pCASPER |
| Bb_pCAS-PstI-R | TAAGCATCTGCAGCGGAGAAGTTAAGCGTCTC | pCASPER |
| *White*-NsiI-BamHI-R | CATGCTAGGATCCATAGCTAGTTGAGATGCATCTACACAAGGAAC | pCASPER |
| *White*-PstI-(NotI)-F | CATGCTCTGCAGACTAGTGGCCTATGCGGC | pCASPER |
| GalBspEI-R | TTCATCCGGAATACAATTTCCAGATATTGAAG | Previous TPase sequence |
| GalBspEI-F | GTATTCCGGATGAAGATTCAATGCTAG | Previous TPase sequence |
| GalBsaI-F | TCTACGGGTCTCATGAGGTTAAAATTAAGAAAGGTCTTC | Previous TPase sequence |
| GalMet-F | ATGGCGCAAATAAGTGTTGTGAACG | Previous TPase sequence |
| GalBsaI-R | TCTTAAGGTCTCCCTCATCGAAAACTAATACTGCATAC | Previous TPase sequence |
| GalStop-R | TTATTCACGAATCATTTTCAGTTTACTTTTAC | Previous TPase sequence |
| An_pM-R | ATGGTCTAGAAAGCTTTAGTTCGGACACAGCAGGGAGT | BioS&T plasmid |
| AnA-pM-F | AAGTTCTGTTTCAGGGCCCGATGAATCGCCAGAACATCCG | BioS&T plasmid |
| AnC-pM-F | AAGTTCTGTTTCAGGGCCCGATGAATCGCCAGAACGTTCG | BioS&T plasmid |
| Bu-150pM-F | AAGTTCTGTTTCAGGGCCCGATGGCTCAGATCAGCGTGG | BioS&T plasmid |
| Bu-150-pM-R | ATGGTCTAGAAAGCTTTAAAAAATCAGCAGGTTTTCAATCAG | BioS&T plasmid |
| Bu_pM-F | AAGTTCTGTTTCAGGGCCCGCGTAAATCCGGTGCGAAATG | BioS&T plasmid |
| Bu_pM-R | ATGGTCTAGAAAGCTTTAATTGGAAAAGAAGTTCGGATCG | BioS&T plasmid |
| MoC_pM-F | AAGTTCTGTTTCAGGGCCCGCAGCGTAATGGCGGTAAGTG | BioS&T plasmid |
| MoC_pM-R | ATGGTCTAGAAAGCTTTAGTTGTTAGAAATCAGGTTGGAGTTACC | BioS&T plasmid |
| MoD-pM-F | AAGTTCTGTTTCAGGGCCCGCGTCGTAACGGTGGTAAATGC | BioS&T plasmid |
| MoD-pM-R | ATGGTCTAGAAAGCTTTAATTGTTGGACAGCAGGTTGC | BioS&T plasmid |

List of abreviations

bp: base pair
BS: binding site
EMSA: electrophoretic mobility shift assay
kb: kilobase
MBP-tag: maltose binding protein tag
ORF: open reading frame
TIR: terminal inverted repeat

The authors declare that they have no competing interests.

Authors' contributions

MM constructed the ancestral and consensus sequences, cloned and expressed the tested proteins, performed the EMSA, constructed the *Drosophila* vectors, performed the in vivo transposition experiment and drafted the paper. DL carried out the footprint assay. RC and AR supervised the research, provided funding for the research and finalized the manuscript. All authors read and approved the final manuscript.

Description of additional data files

The following additional data are available with the on line version of this paper. Additional data file 1 is a table listing the genomic *Galileo* sequences used for inferring the consensus and ancestral THAP sequences of *D. mojavensis* and D. ananassae. Additional data file 2 is a table where the primers used in this work are shown.

Acknowledgements

We would like to thank Azeem Siddique and Corentin Claeys Bouuaert for helping with experimental design, Ray Owens for structural THAP protein suggestions and Martí Badal for *Drosophila* plasmids. Montse Sales, Raquel Ferraz, Alejandra Delprat, Núria Rius, Andrea Acurio and Víctor Soria helped with fly counting. This work was supported by a Formación de Personal Investigador doctoral fellowship (to M.M.) and grant BFU2008-04988 (awarded to A.R. from the Ministerio de Ciencia e Innovación, Spain).

# 3.- Striking structural dynamism and nucleotide sequence variation of the *Galileo* transposon in the genome of *Drosophila mojavensis*

Mar Marzo[1,2], Xabier Bello[3], Marta Puig[1,4], Xulio Maside[3] and Alfredo Ruiz[1].

[1] Departament de Genètica i de Microbiologia, Universitat Autònoma de Barcelona, Spain
[2] School of Biomedical Sciences, Queen's Medical Centre, Univesity of Nottingham, UK
[2] Grupo de Medicina Xenómica, Departamento de Anatomía Patolóxica e Ciencias Forenses, Universidade de Santiago de Compostela , Spain
[3] Functional and Comparative Genomics group, Institut de Biotecnologia i Biomedicina, Univeristat Autònoma de Barcelona, Spain

E-mails: Mar Marzo – mar.marzo@uab.cat, Xabier Bello – xbello@gmail.com, Marta Puig – marta.puig@uab.cat, Xulio Maside – xulio.maside@usc.es , Alfredo Ruiz – alfredo.ruiz@uab.cat

Keywords: transposable element, TIR, length, recombination, subfamily, evolution, *Drosophila*, *foldback*, transposase

## 3.1.- Abstract

*Galileo* is a transposable element responsible for the generation of three chromosomal inversions in natural populations of *Drosophila buzzatii*. Although the most characteristic feature of *Galileo* is the long-internally repetitive Terminal Inverted Repeats (TIR) which resemble the *Drosophila Foldback* element, its transposase-coding sequence presents significant similarity to the *P-element* transposase. This has led to its classification as a member of the *P-element* superfamily (Class II, subclass 1, TIR order). Furthermore, *Galileo* was detected in six of the 12 *Drosophila* sequenced genomes, suggesting a wide distribution in the *Drosophila* genus. *D. mojavensis* is among the six species, the closest to *D. buzzatii*, and the *Galileo* sequences found in this sequenced genome presented the highest diversity in sequence and structure.

In the present work, we carried out a thorough search and annotation of all the *Galileo* copies present in the *D. mojavensis* sequenced genome. Our set of 170 *Galileo* copies present a huge variability in length and structure, ranging from nearly-complete copies to copies with only two TIR or even solo-TIR elements. In addition, the sequence diversity showed the existence of five subfamilies (C, D, E, F, and X), four of them harbouring transposase-coding sequence and a fifth one which presents a putative chimeric origin. Our analysis suggests that *Galileo* is currently active or has been active until very recently. Finally, we have explored the structure and length variation of the *Galileo* copies which points out to relatively frequent rearrangements within and between *Galileo* elements. Different mechanisms responsible of these rearrangements are discussed.

## 3.2.- Introduction

Transposable elements (TE) are genetic entities capable of changing their location in the genome (Kidwell & Lisch 2002). Because of their disperse and repetitive nature, they are considered part of the middle repetitive DNA portion and they make up significant fractions of different genomes, such as 14% in *Arabidopsis thaliana*, ~15% in *D. melanogaster*, ~45% in humans or ~80% in some crops (Lander et al. 2001; Kidwell 2002; Wicker et al. 2007; Hua-Van et al. 2005). They have been found in virtually all the studied species, showing what could be considered a great success in their strategy or the ancientness of their existence (Feschotte & Pritham 2007). Since their new insertion sites are usually random, they are considered as mutational agents, which allowed them to be firstly considered as junk DNA (Doolittle & Sapienza 1980; Orgel & Crick 1980). Nevertheless, they can be taken as powerful facilitators of evolution, since they generate variability, the row material for evolution, along with some adaptive TE insertions which have been reported (Oliver & Greene 2009, 2011).

Since TEs present huge variability in length, structure and transposition strategies, a classification system is needed to understand and handle all the information about this type of DNA. Although classification criteria have not reached a complete consensus, there is a general agreement about the first split in the classification: the existence of a retrotranscription step (Finnegan 1989). Structural and homology criteria are used to further classify the different elements in subclasses, orders, superfamilies and families (Feschotte & Pritham 2007; Jurka et al. 2007; Wicker et al. 2007). TIR DNA transpososns (Class II, subclass I) comprise those elements without the retrotrascription step and with Terminal Inverted Repeats (TIR) (Wicker et al. 2007). These elements are mobilised by a transposase protein encoded by autonomous or canonical copies of the element usding a cut-and-paste mechanisms.

Apart from transcription-active (canonical) copies of a transposon family, most genomes also harbour defective copies which are unable to encode a functional protein and thus non-autonomous. These copies appear due to mutations in the canonical-structured elements, along with genomic deletion and unequal exchange after non-allelic homologous recombination (NAHR) and the transposon activity, generate deletion derivatives copies (Petrov & Hartl 1998; Rio 2002). These defective copies

usually present a gradient of random deletions and there are from almost-complete copies to copies that are only made up of TIRs and a spacing region (Brunet et al. 2002; Rio 2002; Feschotte & Pritham 2007). Furthermore, there is a special kind of defective elements that are called MITEs (Miniature Inverted repeat Transposable Element), which seems to have acquired non-related sequences and only present homology to the canonical copies in the TIRs or the very ends of the TIRs. These MITEs use or parasite the transposition machinery coded in the complete copies and have been proposed as the ultimate parasites (Gonzalez & Petrov 2009; Yang et al. 2009).

*Galileo* is a transposable element discovered in *D. buzzatii* where it has been responsible for the generation of three natural chromosomal inversions (Cáceres et al. 1999; Casals et al. 2003; Delprat et al. 2009). Because the first copies of *Galileo* were only made up of long TIR sequences, it was tentatively classified as *Foldback*-like element (Cáceres et al. 2001; Casals et al. 2005). However, when the *Galileo* transposase sequence was discovered, it was definitely classified as a member of the *P-element* superfamily of DNA transposons (class II, subclass I and TIR elements order), being the longest TIR element (from ~300 bp to 1.2 kb TIR length) of its superfamily (Marzo et al. 2008). Despite the first studies pointed out that *Galileo* distribution was limited to the closest species to *D. buzzatii* (Casals et al. 2005), the bioinformatic analysis of the 12 sequenced *Drosophila* genomes uncovered a broader distribution, because six of the 12 species harboured it (Marzo et al. 2008). In this initial bioinformatic analysis, one of these species, *D. mojavensis* showed a remarkable diversification of *Galileo* sequences, with four phylogenetically differentiated groups, and huge structural variability among the copies. Both *D. mojavensis* and *D. buzzatii* are members of the *repleta* group of the *Drosophila* subgenus.

In the present work, we carried out a more detailed search and analysis of the transposon *Galileo* in the *D. mojavensis* genome. 170 *Galileo* copies were identified using different automated searching strategies coupled with a detailed manual annotation in each of them. A huge variability in length and structure were found, thus sequences from nearly-complete copies to only two TIR elements were found. In addition, the sequence diversity found allowed the description of five *Galileo* groups/subfamilies, one more than the previous work; four of them harbour defective transposase sequences and one of them could have a chimeric origin. The activity of

these *Galileo* copies has been explored through bayesian analysis, which suggests that it has been active until recently or maybe it could be still active. Finally, the structural dynamics, which comprise the TIR extension, has been analysed in detail and mechanisms for this dynamism are discussed.

## 3.3.- Methods

Bioinformatic searches of *Galileo* copies in the *Drosophila mojavensis* genome. Consensus TIR sequences of previously described Dmoj*Galileo* subfamilies plus 50 bp overall consensus TIR end, were used as query sequences against the CAF1 scaffold assembly of *D. mojavensis* genome (Clark et al. 2007). The searches were carried out using an automated process based on wuBlast (http://blast.wustl.edu) and the Chao algorithm (Chao & Miller 1995) for the handling of the sequence discontinuities in the blast searches. The hits were selected using a 80-80 criteria with the query TIR (80% identity and 80% of the length, (Wicker et al. 2007)) and were considered as part of the same *Galileo* copy if arranged in the proper orientation at a distance < 10 Kb. If one TIR did not meet all the mentioned criteria the 3 kb flanking region where the other TIR would be expected to be found was further explored by blast. More *Galileo* copies were found in this way. When no partner was found for a given TIR in the surrounding area, it was considered as a solo-TIR copy for further analysis.

All hits from each search were manually curated and thoroughly analysed to discard wrong automated identifications. Decisions on the acceptance of a search hit were based on the comparison with previously characterised copies and the identification of characteristic structures by careful annotation. This way, we identified the different regions in each *Galileo* copy: the Terminal Inverted Repeats (TIR), the transposase-coding region, and the spacing sequences upstream and downstream of the transposase-coding region (those we have named F1 and F2 respectively). Only sequences showing a clear sign of some of these structures were selected for further analysis.

Annotation of *Galileo* copies. All selected sequences were manually analysed and annotated using several tools found in Geneious 5.1.7 software package (Drummond et al. 2010). The closest annotated sequence for each new copy was detected by a search with blastn (Altschul et al. 1997) and used as reference for the detailed annotation of the new copy. When a region of a new copy was not located in the chosen reference copy, this region was used as blast query against different *Galileo* sequences and other *Drosophila* TEs in order to detect regions in common with other *Galileo* copies or TE insertions. TIR span was determined by aligning each copy with the corresponding reverse complement sequence. All copies were classified by structure in one of the

following five categories: i) nearly-complete (NC), when two TIR and more than 2 kb of transposase-coding sequence were found; ii) deletion derivatives (DD), when either two TIR and less than 2 kb of transposase-coding sequence were found, or a complete or partial transposase-coding sequence was found, but only one TIR was identified; iii) two TIR elements (2T), when two TIR separated by a short middle region (usually not coding for transposase) were found; iv) two extended or recombinant TIR (2RT), when two TIR were found and they were either longer than the NC copies or presented duplicated sequences (there has been extra sequence recruited in a longer TIR); and v) solo-TIR (ST), when only one TIR was found. Detailed information of the genome location and annotation of each *Galileo* copy is provided in Supplemental Table 2.

TIR phylogeny. The phylogenetic relationship between *Galileo* copies was inferred from the analysis of a 630 bp sequence from the 5' end of the representative consensus TIR. Shorter than 450 bp selected sequences (due to partial deletions) were excluded from the analysis to improve the alingment. These TIR regions were aligned with MAFFT using the following parameters: E-ins-I; --op 1.53; --maxiterate 1000; --genafpair; --ep 0; --inputorder; --kimura 200, as it is set in Geneious software (Katoh et al. 2002; Drummond et al. 2010). The alignment was filtered with Gblocks 0.91b to remove regions too divergent and poorly aligned (Castresana 2000; Talavera & Castresana 2007). Gblocks was set up with relaxed parameter values (Minimum Number Of Sequences For A Conserved Position: 120; Minimum Number Of Sequences For A Flanking Position: 120; Maximum Number Of Contiguous Nonconserved Positions: 10; Minimum Length Of A Block: 5; Allowed Gap Positions: With Half) selecting 53% of the original alignment (547 bp of the 1018 original positions). JModeltest 1.0 (Posada 2008) was used to find the substitution model that best fits the data by means of the Akaike Information Criterion (AIC), which resulted to be HKY+G (Hasegawa, Kishino and Yano plus gamma (Hasegawa et al. 1985)). Maximum likelihood (ML) search was performed with PhyML 3.0 (20110304) (Guindon & Gascuel 2003; Guindon et al. 2010) using the Subtree Pruning and Regrafting (SPR) algorithm. The parameters of the substitution model were estimated by the program, using four categories to estimate the gamma distribution and support was calculated with 100 bootstrap replicates. Bayesian inference (BI) was carried out with BEAST 1.6.1 (Drummond & Rambaut 2007), using an uncorrelated lognormal

relaxed clock (UCLN (Drummond et al. 2006)) and the substitution model from jModeltest. We used a birth-death process as a tree prior setting a uniform (0, 1000) distribution for growth and death rates. All others priors were left with default values. Two MCMC chains of 50 million generations were run and combined with the LogCombiner program included in BEAST package. In both cases, the chains were sampled every 1,000 steps, and the first 10% of the samples was removed as burnin. Convergence was ensured checking that ESS values for all parameters were over 200. We obtained the maximum clade credibility summary tree with median node heights using TreeAnnotator (also included in BEAST package).

Recent transpositional activity. A BEAST phylogenetic inference was carried out with the aim of displaying the relative age of each *Galileo* copy. For this purpose only one TIR region (of at least 450bp long) was picked up from each copy and chimeric elements were excluded. The BEAST priors were set up as mentioned above with the same evolutionary model (HKY+G). Absolute time estimation was performed using the 0.011 changes/base/myr proposed as neutral mutation rate in *Drosophila* (Tamura et al. 2004). After that, a lineage through time plot was generated which depicts copy accumulation through time (Barraclough & Nee 2001). We performed statistical test to find out the best fitting model to a sample of 9000 trees from the BEAST inference. The diversification models tested were: pure-birth (constant rate), birth-and-death (constant rate), DDX (variable rate), DDL (progressive change with saturation) and Yule-two-rates (abrupt change of the rate in one point). These models were adjusted by ML and the best one was chosen using an Akaike Information Criteria (AIC). Furthermore, simulations to test if the best fitting model was due to incomplete sampling or data variability were carried out.

Transposase-coding region phylogeny. Transposase-coding sequences found in the different groups longer than 2 kb (12 elements: 6498-22531F, 6500-31458D, 6541-16442D, 6540-11758D, 6540-23860D, 6485-39163D, 6540-41449X, 6262-30856C, 6541-11419F/C, 6500-31288C, 6482-60893F) were aligned with MAFFT (same parameters as above), and jModelTest was run to find the best evolutionary model for the transposase-coding sequences. ML and BEAST tree were inferred for these sequences (evolutionary model JC+G+I). The cognate TIR of each copy with a transposase-coding segment >2 kb were aligned with MAFFT and new phylogenies

with PhyML and BEAST were obtained. The topologies of the transposase-coding sequences and TIR phylogenies were compared and the differences were evaluated with an Approximated Unbiased test (AU test) performed with CONSEL program (Shimodaira & Hasegawa 2001; Shimodaira 2002).

Chromosomal distribution of *Galileo* copies and relation to protein-coding and RNA genes. The genomic and cytological location of *Galileo* copies was inferred from the scaffold coordinates and the correspondence of scaffolds with polytene chromosomes (Schaeffer et al. 2008). In order to analyse the intrachromosomal distribution of *Galileo* copies, each chromosome was divided in three regions: telomeric, central and centromeric, conainting 10%, 80% and 10% of the sequence, respectively (Casals et al. 2005, 2006). This was only possible for chromosomes 2, 3, and 4, each of them represented by a single major scaffold (Schaeffer et al. 2008). Statistical analyses of chromosomal distribution were carried out with JMP 8.0.2 (SAS Institute Inc. 2009). The *D. mojavensis* gene annotations were downloaded from Flybase.org (ftp://ftp.flybase.net/releases/*FB*2011_04/). The coordinates of protein-coding and RNA genes were compared with those of *Galileo* copies using ad hoc perl scripts. All *Galileo* copies were classified as located in scaffolds without genes, in intergenic regions or in intronic regions. Statistical tests to compare the total length and TIR length with genes distances were performed with JMP 8.0.2 (SAS Institute Inc. 2009). Information about the gene function was extracted from FlyBase.

## 3.4.- Results

Different bioinformatic search strategies were used to maximise the probability of finding *Galileo* copies (see Methods). A total of 170 *Galileo* copies were identified and manually annotated (a 370% sample increase over the 36 previously described copies (Marzo et al. 2008)). These copies were classified according to subfamily, structure and chromosomal distribution (see Table 3.1 for a summary and SI Table 3.1 and SI Table 3.2 for detailed information). Subfamily classification was based on the phylogenetic analysis of TIR sequences and resulted in five well-supported groups (C, D, E, F and X). Twelve copies were found to contain sequences belonging to different subfamilies and were considered as chimeric (Table 3.1). Structural classification produced five groups: nearly-complete (NC), deletion derivatives (DD), two TIR elements (2T), two extended or recombinant TIR elements (2RT) and solo-TIR (Table 3.1). Some representative copies of these structural groups are depicted in Figure 3.1.

**Table 3.1.** Summary of the *Galileo* copies studied in this work. The different subfamilies and structures are indicated.

| Structural type | Subfamily | | | | | | Total |
|---|---|---|---|---|---|---|---|
| | C | D | E | F | X | Chimeric | |
| Nearly complete (>2 kb Tpase) | 2 | 5 | 0 | 1 | 1 | 1 | 10 |
| Nearly complete deletion derivatives | 4 | 2 | 0 | 1 | 2 | 0 | 9 |
| 2 TIR | 5 | 0 | 7 | 28 | 3 | 6 | 49 |
| 2 TIR longer | 2 | 2 | 22 | 3 | 4 | 5 | 38 |
| solo TIR | 6 | 10 | 19 | 26 | 3 | 0 | 64 |
| Total | 19 | 19 | 48 | 59 | 13 | 12 | 170 |

*Galileo* subfamilies in the *D. mojavensis* genome. A phylogenetic tree was built using the homologous TIR region of all the copies (Figure 3.2A). The tree shows five groups with significant statistical support, four of them (C, D, E and F) agree with the previously described *Dmoj\Galileo* subfamilies (Marzo et al. 2008), whereas the fifth, that we have named X, is a novel group (Figure 3.2A). The general relationship among the groups is similar to that found in the previous work, with two main lineages, one comprises the D, E and X group, and the other the C and F groups. Furthermore, the phylogeny also detected 12 chimeric copies (not shown in Figure 3.2A) with the two TIR belonging to different phylogenetic groups. In addition, these copies are flanked by non-matching 7-bp sequences instead of identical direct target site duplications (TSD) as most other copies.

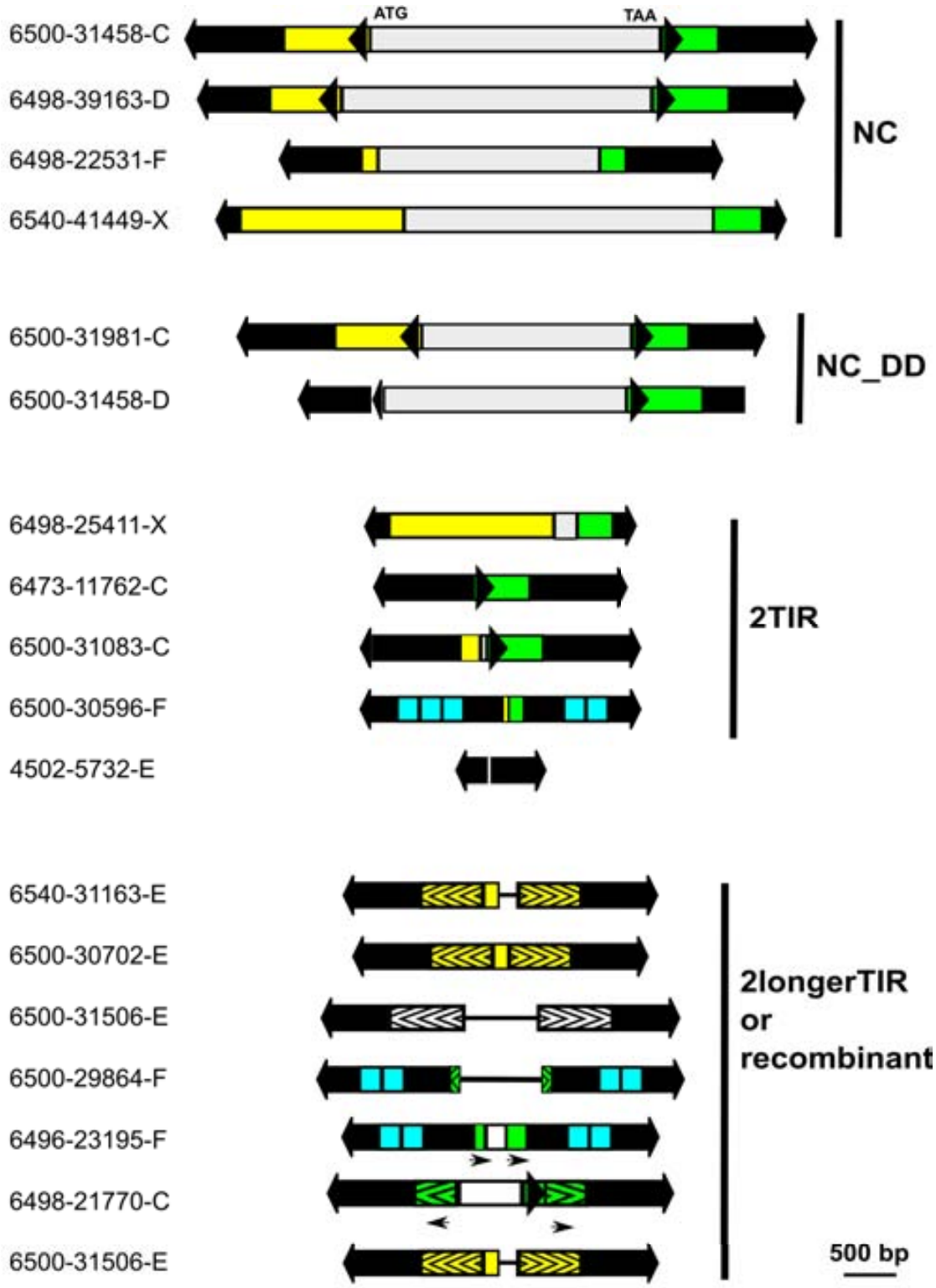



**Figure 3.1**. Structures of representative *Galileo* copies found in the *D. mojavensis* genome. The black arrows are the TIR, the grey middle region is the transposase sequence, the yellow region is the F1 (spacing sequence between the TIR 1 and transposase coding segment), the green region is the F2 (spacing sequence after the transposase-coding segment and the TIR-2). The blue squares are tandem repeats found in the F group. The region with bracketed pattern (>>>) is the extra TIR region recruited in the extended TIR copies. The black arrowheads are internal short inverted repeats found in C and D groups. NC copies are nearly-complete, NC_DD are deletion derivatives of the nearly-complete ones.

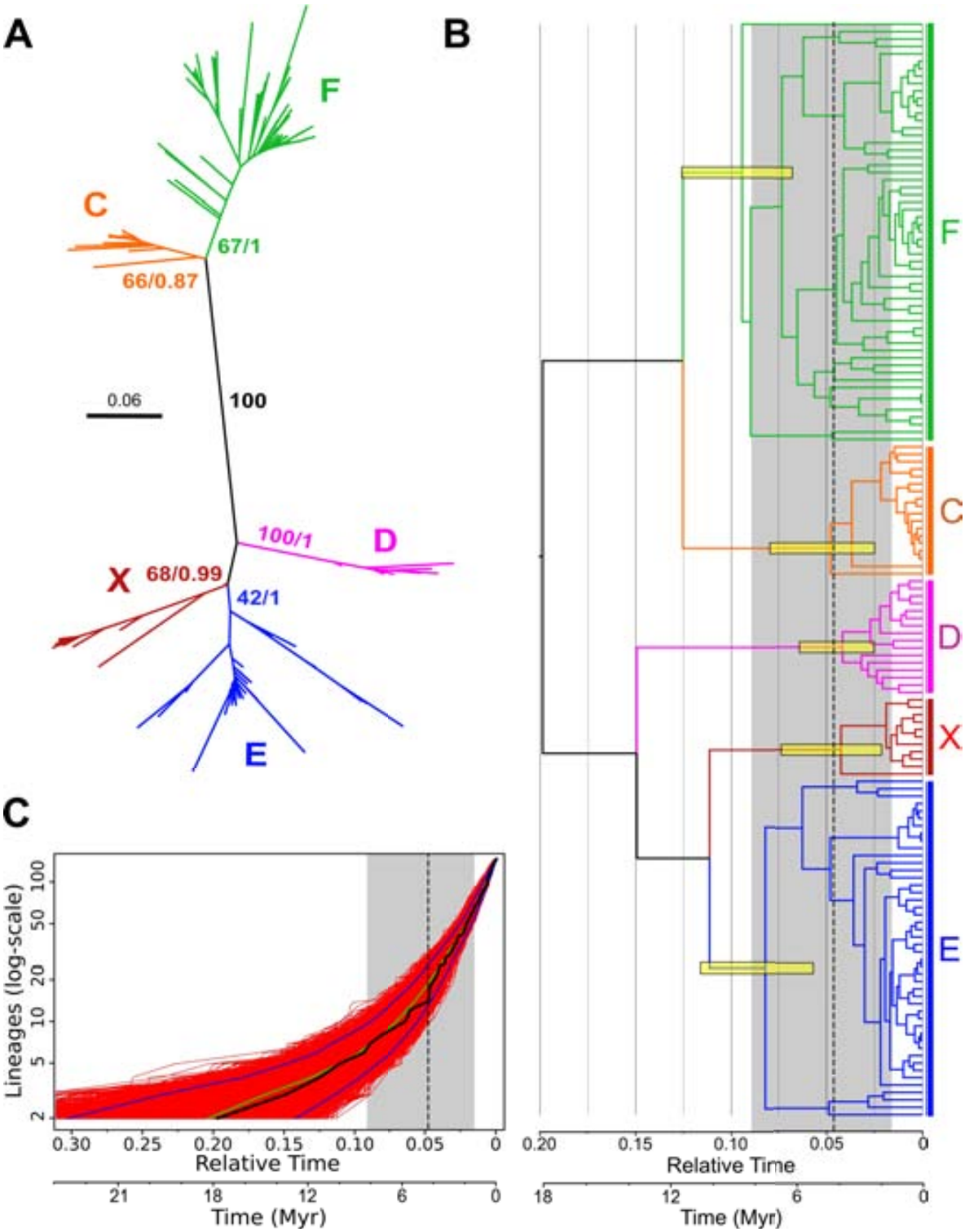



**Figure 3.2**. *Galileo* phylogenetic analyses. A) Unrooted tree inferred using 241 TIR sequences of *Galileo*. Phylogenetic reconstructions were carried out by means of ML (PhyML) and BI (BEAST) methods using a HKY+G evolutionary model. Numbers on nodes indicate the support of each group as bootstrap and Bayesian posterior probability, respectively. The five groups show strong support. B) BEAST ultrametric summary tree inferred using 148 TIR sequences of *Galileo* (only one TIR of each *Galileo* copy was used and chimeric copies were excluded). The yellow bars correspond to the 95% Highest Posterior Density intervals for node ages. The ML best-fit model of diversification was a yule-2-rate in which a constant duplication rate changes to another constant rate at a certain time, and the discontinuous vertical line indicates the shift in the duplication rate (0.048 substitutions/position, ~4.36 myr) and the grey area represents the 95% confidence interval obtained using 10,000 trees sampled from the Bayesian analysis. C) Lineages Through Time (LTT) plots representing the accumulation of cladogenesis events. The black line shows the LTT plot of the summary Bayesian tree. Red and blue lines represent the mean and the 2.5% and 97.5% percentiles of the 10,000 sampled trees LTT plots, respectively.

In order to explore the evolutionary dynamics of *Galileo* copies through time, an ultrametric tree was generated using a relaxed molecular clock (Figure 2B). In this case, only one TIR sequence per *Galileo* copy was included (usually TIR1, and in some cases TIR2 when TIR1 was not present or was too short) and chimeric copies were omitted. In this tree we included an estimation of absolute time, which provides ages for each node. If we take into account the common ancestral node for each one of the *Galileo* subfamilies, different ages are found. For example, the last common ancestral node for all the F copies is ~8.6 myr, which means this group would be the first one diversifying in this genome. It would be followed by E (~7.45 myr), C (~ 4.35 myr), D and X (these last two less than 4 myr). Most of the copies (~ 60%), regardless the phylogenetic group, seem to be quite recent as they appeared in the last million year. In addition, the cumulative graphic of Lineages Through Time (LTT plot) showed an exponential growth of the number of *Galileo* sequences without any apparent deceleration in the curve (Figure 3.2C). Thus, *Galileo* has not stopped its transposition activity in the time depicted in the graphic. Furthermore, we have performed a diversification rate test and, at least, one shift has been detected which is located in 0.048 relative time units (substitutions/position) (~4.36 myr vertical discontinuous line in the tree, Figure 3.2 B and C) where the rate of *Galileo* proliferation changes from 16.28 sequences/relative time units to 48.66 sequences/relative time units (95% confidence interval for each rate: 5.87-30.31 and 39.77-58.24 lineages/time). These observations indicate that *Galileo* is still active or has been active until very recently.

Twenty *Galileo* copies were found to contain variable portions of the transposase-coding region (Table 3.1, SI Table 3.1), yet none of them harbours an intact ORF that can be translated into a functional protein (i.e. all of them contain chain termination mutations and/or deletions and frame-shift mutations). These copies belong to subfamilies C, D, F and X, whereas no copies of the E subfamily contain any trace of the transposase-coding region. A phylogenetic tree was built with transposase-coding sequences longer than 2 kb found in the different subfamilies (12 *Galileo* copies in total, see methods). For comparison, the TIR region of these 12 copies was used to generate a new tree with the same methods. Both phylogenetic trees were similar and recovered the same groups (Figure 3.3, Table S3). However, the relationship among the subfamilies seem somewhat discordant: in the transposase-coding region tree groups F and D belong

to one of the main lineages, and groups X and C belong to the other, whereas the TIR tree shows the same relationship between groups found previously in the global TIR tree (Figure 3.2 A and B). Differences in topology can be due to different evolutionary histories, but also to phylogenetic uncertainty. In fact, the grouping of F and D in the transposase-coding tree has a low bootstrap support (41%). Moreover, an AU test was performed (CONSEL program) to test if any of the two topologies could be significantly rejected using the information in both alignments. This way, neither of the two topologies could be rejected in the case of the transposase alignment (TIR topology: P = 0.39, transposase-coding topology: P = 0.61), indicating that information in the alignment does not allow discriminating between both phylogenetic hypotheses. However, when the TIR alignment was used, we found that the transposase-coding topology was significantly rejected (TIR topology P = 1; transposase-coding topology P = 7e-11). These results suggest that the position of the F subfamily in the transposase coding segment tree might be biased, as a consequence of the reduced number of sequences used, phylogenetic noise in this *Galileo* region or recombination.

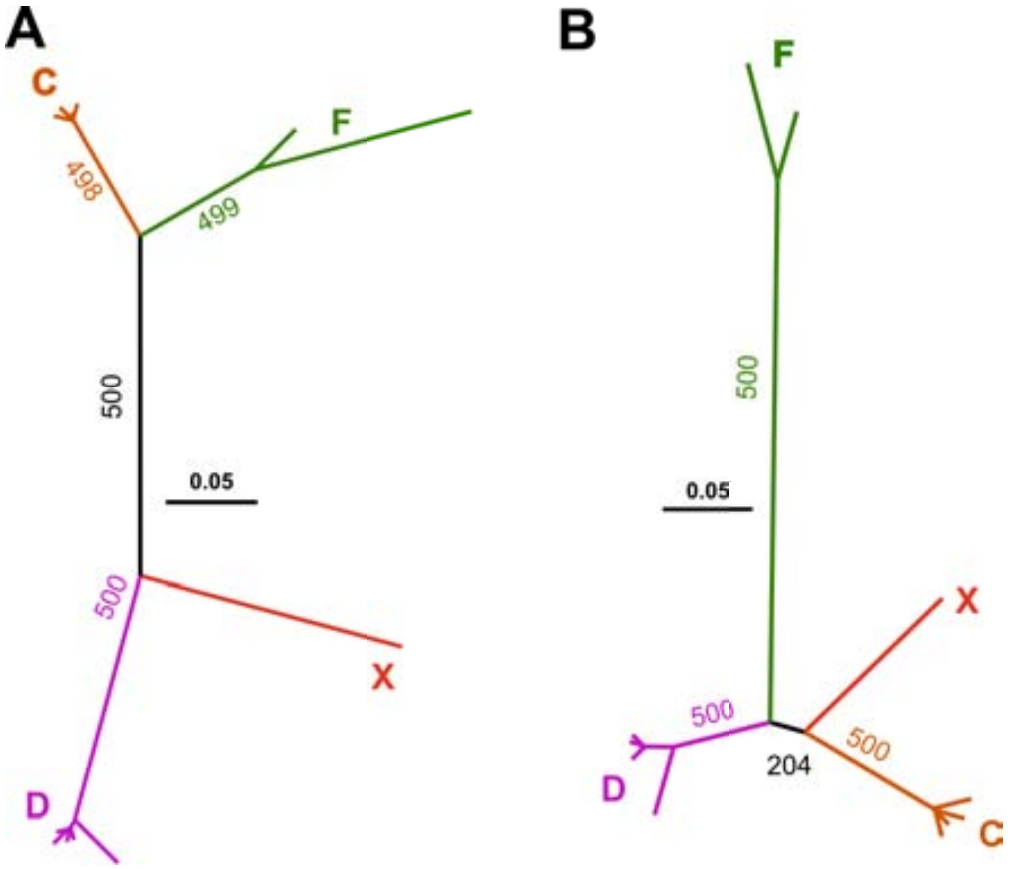



**Figure 3.3**. TIR and transposase coding region phylogenies. 12 *Galileo* elements were used for these analyses. A) TIR phylogeny. B) Transposase phylogeny, PhyML analysis with JC+G+I evolutionary model. The AU test was performed to compare the two tree topologies.

*Galileo* structural variation. *Galileo* copies exhibit a striking amount of structural variation (Figure 1). For the purpose of description and analysis, we have grouped all copies into five structural groups: NC, DD, 2T, 2RT and solo-TIR (see methods). All phylogenetic groups except D and E contained copies of the five different structures described (see Table 3.1). The D subfamily lacked 2T elements, whereas the E subfamily did not contain any copy with transposase sequence (neither NC nor DD).

The *Galileo* TIR, defined as the terminal sequence inverted and repeated in each end, is the most variable region among the copies of the element, not only in nucleotide sequence as phylogeny shows but also in length. TIR length varies from 18 bp to 1250 bp with a total average of 668 bp. The variation of TIR length is found in all the subfamilies (see SI Table 3.1 where means and standard deviation are found), but when the five subfamilies means are compared, the only pairs of comparisons that present statistical differences are between the X and E subfamily and X and F subfamily (Tukey-Kramer means comparison test, $P<0.05$). The X subfamily possesses the shortest TIR, and subfamilies E and F the longest TIRs. When the TIR length is compared among the different structural types, the only significant length different is found between the 2T and the 2RT type, which is in agreement with the classification criterion (Tukey-Kramer means comparison test, $P<<0.05$). We have explored the sequences comprising the TIRs. Generally, the shortest TIRs are due to the lack of TIR sequence in one of the *Galileo* ends. Thus, although one transposon end still posses a whole TIR, the repeated span gets shorter because of the sequence missing in the other end (it is not repeated any more). This is how some very short TIRs are found in copies like F subfamily 6680-244202 or X subfamily 6498-95069, E subfamily 4198-1393 or C subfamily 6540-613211 (see copy 4502-5732E in Figure 3.1).

On the other hand, when the longest TIR are explored, we have observed differences among the subfamilies. For example, in the F subfamily, the presence of direct tandem repeats inside the TIR (located in ~264-467 bp from the TIR end) seems to account for part of the variation in the TIR length. There are TIRs with no internal repeats and TIRs with two or three copies of the internal tandem repeat. Since the tandem repeat region is ~210 bp long, when three copies of this sequence are present, TIR length increases by ~420 bp. This fact was found in the TIR1 of 6500-30596F and 6500-31107F which are 1264 and 1263 bp long because they harbour three internal tandem repeats. In contrast,

copies 6540-32286F or 6540-57500F harbour 892-bp TIRs due to the lack of internal tandem repeats. It is noteworthy that the tandem repeat expansion and contraction was only found in the F group and was located always in the same region of the TIR, except in copy 6500-30494F which harboured two tandem repeats located in 196-101 bp from the TIR2 end.

In the other groups, although the tandem repeat structure in the TIR was not found, some copies showed also longer TIR, when compared to the NC copies. In these cases, the detailed exploration of the TIR sequences uncovered the recruitment of non-TIR *Galileo* sequences (usually the region found immediately after the TIR in the NC *Galileo* element) to generate a longer TIR. For example, part of the sequence of the F1 area (the sequences after TIR1 but upstream the transposase coding segment) appeared repeated in inverted orientation immediately before the beginning of the TIR2 extending the repetitive span inside the *Galileo* element. This way, an originally non-duplicated neither repetitive *Galileo* sequence made up a longer TIR. We observed that the extra region of TIRs can come both from the F1 or the F2 region, however, the F2 region appeared duplicated only in the groups C (2 copies) and F (once as direct repeat, another time as inverted repeat and it is found in a chimeric copy, as well) whereas F1 region appeared repeated in the C, D (2 copies), E (22 copies) and X (4 copies plus 2 chimeric) groups.

The *Galileo* copy with the longest TIRs showed a combination of the two expansive traits: tandem repeat expansion (two times the tandem repeat in each TIR) along with the recruitment of 121 bp of F2 sequence in the TIR. This copy is 6500-29864F (see Table S2), and has TIR lengths of 1260 bp and 1241 bp (TIR1 and TIR2, respectively with a 95.2% of nucleotide identity). The second and third longest TIR copies belonged to the C group, where two 2RT copies recruited F2 region for the TIR reaching 1107 bp long. The next longest copy was found in the E group, followed by copies in the D and X groups (SI Table 3.2). It is noteworthy that the copies with the longest TIRs were never the nearly-complete ones but the non-autonomous without the transposase-coding ORF, i.e. 2T and 2RT copies (SI Table 3.1 and SI Table 3.2). All *Galileo* subfamilies present substantial TIR length variation, because in all the groups there are copies with very short and very long TIR.

Chimeric copies. Twelve *Galileo* copies were composed of two TIR with an unusually high nucleotide divergence and were bounded by different 7-bp sequences instead of identical TSD (see SI Table 3.2). The TIR phylogeny confirmed that these *Galileo* copies were chimeric (not shown). Structurally, one of these copies was NC and all the others are 2T. Regarding the subfamily, there are 4 F/C (including the NC), 1 F/D, 2 E/F, 1 E/C and 4 F/X. The contribution of each subfamily to the chimeric copies is in agreement with its abundance (Chi square test, $P > 0.05$). The fact that F TIR were more frequent in the chimeric copies would be due to the larger number of F copies in the genome. On the other hand, we have tested if the different subfamilies are randomly combined or whether there are subfamily preferences when the chimeric copies are generated. We have not detected any significant departure from randomness ($P>>0.05$).

We have detected the presence of another kind of chimeric copies, with the two TIR from the same phylogenetic subfamily, but the internal region from another one. Furthermore, the central region of all these copies seems to have the same origin, the central region of 6680-240698D, one of the 2RT copies of the D subfamily. The central region of this copy presents 441bp of F1 duplicated and inverted expanding the TIR length. When the E subfamily was explored, the central region of its copies presents high identity to this internal region of the 6680-240698D copy (98% of identity), while the 570 bp of the end of each TIR presents 77% of identity and, as the phylogenies show, belong to different subfamilies. Likewise, we have found this same central region in two 2T copies classified in the X group (copies 6498-29033 and 6500-29395, classified as X group, ~1640 bp total length). Thus, the same central region was found accompanied by TIRs from three different subfamilies, D E and X.

*Galileo* chromosomal distribution and relationship with genes. We have analysed the interchromosomal and intrachromosomal distribution of the *Galileo* copies (SI Table 3.3 and SI Table 3.4). 138 of the 170 *Galileo* copies are located in scaffolds assigned to the *D. mojavensis* chromosomes (Schaeffer et al. 2008). The remaining 32 copies are located in scaffolds that are likely to contain pericentromeric heterochromatin and have not been assigned to any chromosomes yet. The distribution of the 138 copies was 29, 26, 43, 14, 3 and 23 for *D. mojavensis* chromosomes X, 2, 3, 4, 5, and 6 (dot), respectively. This interchromosomal distribution shows a significant departure from a random distribution (taking into account the size of each chromosome, chi square test

P<<0.05). There is an excess of *Galileo* copies in the dot chromosome, whereas fewer than expected copies are found in the chromosome 5.

In addition, we have explored, the intrachromosomal distribution of *Galileo* copies. In the *D. mojavensis* there are three chromosomes (2, 3 and 4) represented each by a single major scaffold (6540, 6500, 6680, respectively) (Schaeffer et al. 2008)). We have subdivided these scaffolds in distal (10% of the sequence), central (80% of the sequence) and proximal (or centromeric, 10% of the sequence) segments in relation to the position of the centromere, and tested if *Galileo* copies present a uniform distribution in these regions. We observed a very significant departure from what was expected by chance, since *Galileo* copies tend to accumulate in the proximal region near to the centromere (P<< 0.01, in the three cases, SI Table 3.4).

Furthermore, coordinates of *Galileo* copies have been compared to those of the predicted genes in *D. mojavensis* genome (including protein-coding and RNA-coding genes). The 170 *Galileo* copies were classified as follows: 23 are located in scaffolds without genes, 23 are located inside genes (all of them inside introns) and 124 are located in intergenic regions (see SI Table 3.5 and SI Table 3.6). The distances to the closest gene of the intergenic *Galileo* copies ranged from 29 to 110537 bp (average 11439bp, median 5253bp). No correlation was observed between copy length and distance to the nearest gene (Spearman's rho P>>0.05), or between copy length and intergenic region length (Spearman's rho P>>0.05). There was no differential distribution regarding the 5' or 3' gene regions (chi-square test P>>0.05) neither when the different subfamilies (P>>0.05, from 1 to 0.36) or the structural *Galileo* type (P>>0.05, from 0.22 to 1) were taken into account.

A set of 17 *Galileo* copies are located very close to genes (less than 500 bp, SI Table 3.5). The function of these genes have been explored and they are involved in different cellular processes, such as tRNAs, methyl transferases, helicases, DNA binding proteins and 14 of them possess a *D. melanogaster* ortholog. Another group of copies (23 *Galileo*) have been found inside genes. In all the cases the *Galileo* elements were located inside16 different introns (in some introns there were more than one *Galileo* element). The length of these introns ranged from 1478 to 172415 bp, and 10 of the 16 genes whose introns harboured *Galileo* copies, have been assigned an orthologous gene

in *D. melanogaster*. (SI Table 3.6). There was no correlation between *Galileo* length and intron length, neither type nor subfamily is over-represented inside the genes (P>>0.05).

## 3.5.- Discussion

In a previous work, we uncovered the presence of *Galileo* elements in six of the 12 sequenced *Drosophila* genomes (Marzo et al. 2008). Among them, *D. mojavensis* genome showed the highest variability in *Galileo* sequence and structure. A small sample of 16 nearly-complete copies that contained transposase-coding sequences and 20 non-autonomous copies was analysed. Analysis of the TIR sequence variation showed that the copies clustered in four different groups or subfamilies (that were named C, D, E and F). Two of these subfamilies, C and D, harboured truncated transposase coding region, while the other two groups were only composed by non-autonomous copies (mainly 2 TIR structure). The existence of different groups in the same genome suggested different amplification bursts in the past. Furthermore, a high variability in TIR length was detected. Since the TIR length is the most characteristic feature of *Galileo* elements, the *D. mojavensis* genome offered the opportunity to study this trait in detail.

Here, we carried out a thorough analysis of *Galileo* variation and distribution in the *D. mojavensis* genome sequence. In the present work we have uncovered the existence of at least 5 subfamilies of *Galileo* elements. Four of them contain nearly complete copies with transposase-coding segments, what implies the putative co-existence of four fully functional subgroups. The co-existence of different subgroups or subfamilies has previously been reported for *D. melanogaster P-element* and other transposons (Hartl et al. 1997; Quesneville et al. 2006; Miskey et al. 2007; Moschetti et al. 2008). There are two main hypotheses which would explain the co-existence of different subfamilies in the same genome: horizontal transfer and genomic diversification. On the one hand, in case of horizontal transfer events, the *Galileo* element could have arrived to *D. mojavensis* via some close spatio-temporal species, such as mites or other intimate parasites (Houck et al. 1991; Silva et al. 2004; Le Rouzic & Capy 2005; Loreto et al. 2008). If the five subfamilies (C, D, E, F and X) had arrived through this mechanism, this would imply at least 5 independent events of successful horizontal transfer and invasion of *D. mojavensis* genome. If our estimation of each subfamily age is taken into account, these horizontal transfer events would have happen in a ~5 myr period, which would mean an average of one horizontal transfer event per million year. When the

variability of the age nodes is taken into account, this time range reaches ~9.5 myr (from 0.125 to 0.02 changes/time, 11.36 and 1.81 myr, respectively), which would mean ~0.53 horizontal transfers per myr. This would imply something like a "*Galileo* bombing" against *D. mojavensis* genome in the past. This HT rate is higher than the 0.04 HT/myr/family obtained by Bartolomé et al. (2009), even if we divide our estimation among the number of *Galileo* subfamilies, we still get a higher rate of 0.1 HT/myr/subfamily. This massive horizontal transfer seems unlikely.

On the other hand, the different *Galileo* subfamilies could have diverged vertically from an ancestral resident in the genome. This putative ancestor sequence would have existed ~18 myr ago (0.20 units/relative time, considering 0.011 changes/position/myr (Tamura et al. 2004), as it is seen in our Beast ultrametric tree (Figure 2B). Such functional differentiation would have to be driven by specific selective pressures to form several subfamilies producing distinct *Galileo* transposases to overcome the cell transposition repression. When a new transposase appears along with high-affinity sequences, a tranposition burst would happen. After that, truncated copies of the successfully transposed ones would appear, rendering deletion derivatives, 2T, 2RT and solo_TIR copies. In each subfamily, all these structural types would appear independently and could spread while they conserve the affinity for the enzymes encoded elsewhere in the genome by an autonomous copy (Le Rouzic & Capy 2006; Gonzalez & Petrov 2009; Yang et al. 2009). This is the landscape *Galileo* presents in *D. mojavensis* genome.

Furthermore, another factor that would influence the *Galileo* diversification would be the genetic drift, which is very sensitive to the host population structure. *D. mojavensis* is a species with very divergent populations which are even considered as races. It could be possible that in each population a different *Galileo* subfamily evolved and secondary contacts with these populations mixed the different groups. However, our time estimation of each subfamily it is not in agreement with the putative ages of the different *D. mojavensis* races, which would have probably less than one myr (Machado et al. 2007; Reed et al. 2007). Thus, population structure seems not to explain the existence of *Galileo* subfamilies in *D. mojavensis*.

Nevertheless, the two mechanisms, horizontal transfer and genetic diversification are not mutually exclusive, thus, a combination of the two phenomena could have happened. However, it seems more parsimonious the vertical diversification of *Galileo*. Our estimations depicted that *D. mojavensis Galileo* subfamilies have a common ancestor ~18 myr ago. This is showing us that *Galileo* has an old history in *D. mojavensis*, which is in agreement with the *Galileo* ancient origin in the genus (Marzo et al. 2008). Likewise, recent data from the *repleta Drosophila* species group have uncovered the existence of *Galileo* elements in almost all the species of the complex (Andrea Acurio, Deodoro Oliveira and Alfredo Ruiz, in preparation). However, although the *Galileo* last common ancestor in the genus could be as old as the origin of the *Drosophila* genus, the subfamilies found in *D. mojavensis* diversified quite recently (4-9 myr ago). Consequently, only closely related species to *D. mojavensis* are expected to harbour these very same subfamilies, and other different subfamilies probably exist in more distantly related species.

The genomic dynamics of transposons seems to be similar for the different subfamilies. The natural cycle of a transposon would begin with the invasion of a new genome of a fully functional transposon, for example through horizontal transfer (Silva et al. 2004; Le Rouzic & Capy 2006; Loreto et al. 2008). After that, since class II transposition depends entirely on the cell replication and repairing machineries of the double strand breaks, the truncated copies start to appear due to errors in the repair process. Likewise, the truncated copies that would maintain the sequences recognised by the transposase, would be able to spread better than the complete copies, probably due to the overcome of the putative length penalty some transposons suffer (Atkinson & Chalmers 2010). Moreover, even shorter copies would appear, the so-called MITEs and, eventually, the transposon would end inactivated and disappear (Silva et al. 2004; Feschotte & Pritham 2007).

*Galileo* element structures clearly show this dynamics. The nearly-complete copies are 5.2 kb average length and a gradient of shorter copies with different deletions appeared. This way, a bunch of copies where no transposase sequence is found appears, which is composed almost entirely of TIR. Maybe, these copies could be considered as *Galileo* MITEs, but there are some drawbacks for this definition. First of all, the main trait of MITE is its length, usually less than 600bp (Feschotte et al. 2002; Feschotte &

Pritham 2007; Wicker et al. 2007). *Galileo* 2-TIR elements are 1.7- 2.2 kb average length, mainly due to the TIR length *per se*. Secondly, MITEs usually posses sequences which are not found in the complete copies, a fact that made very difficult to find the parental elements of the first MITEs (Feschotte et al. 2003). In *Galileo*, the changes from the most complete copies to the 2TIR elements are traceable virtually all copies. Finally, although the 2TIR copies outnumber the nearly-complete ones, the number of copies is not as many as the MITEs thousand copies reached in some genomes (Feschotte & Pritham 2007). Thus, we propose 2TIR element tag for this kind of *Galileo* copies.

Regarding the *Galileo* TIR dynamics, we have observed length expansion and contraction. On the one hand, for the contraction, the genomic deletion rate in TEs has been studied and would explain how this would happen (Petrov & Hartl 1998). On the other hand, the expansion of the TIR would be a bit more complex than deletion. The expansion of the TIR in the F groups is mainly due to the expansion and contraction of the direct tandem repeats which are located inside the TIR. We have observed different number of tandem repeats in each of the TIR of a *Galileo*-F copy, rendering independent TIR dynamism. This would be in agreement with the statement that any region generated by duplication can thereafter be duplicated (Newman & Trask 2003; Fiston-Lavier et al. 2007). Furthermore, the tandem repeats in the TIR or in subterminal regions of transposons have been proposed to be secondary binding sites for the transposase (Cheng et al. 2000; Cui et al. 2002; Moschetti et al. 2008; Marquez & Pritham 2010). In our case, *Galileo* elements contain these tandem repeats as well, and they have been found independently in two different subfamilies: *D. mojavensis* F(Dmoj*Galileo*F) and *D. buzzatii* G (Dbuz*Galileo*G) (Casals et al. 2005; Marzo et al. 2008, 2011). The multiple binding sites seems to be a convergent trait that appears in different transposable element superfamilies and could be positively selected for an improved transposition reaction, thanks to a higher affinity for the transposition machinery.

Besides the tandem repeat expansion, we have detected another source of TIR extension: the recruitment of internal sequences to extend the TIR. This could be due to the structure of the *Galileo* sequences, where two close inverted repeats of least ~600 bp long might attract recombination, whether due to the DSB after transposon excision,

the structural instability or ectopic recombination as a result of being a genomic dispersed repetition. We could suggest that *Galileo* would have a behaviour similar to the segmental duplications besides its transpositional nature. Segmental duplications are repetitive regions of the genome that are able to recombine, exchange and convert sequences (Bailey & Eichler 2006). For example, if a *Galileo* copy suffers a DSB in the TIR2 (due to a problem during replication step, for example) it could be repaired through non-allelic homologous recombination (NAHR). If for repairing this TIR2 it is used as template the TIR1 of a copy of the same subfamily (the two TIR present 98-100% nucleotide identity between the TIRs of the same *Galileo* copy) it is possible that it would be copied more sequence than the strictly TIR. In that case, since the TIR1 is being copied where the TIR2 is located, the region that was downstream of the TIR1 would appear upstream of the TIR2 as well, becoming a repetitive sequence in inverted orientation and extending the TIR span. The result is TIR1-F1-F1-TIR2. The expansion of inverted repeat sequences have been reported for segmental duplications, and *Polintons* inverted repeats (TE), thus, the dynamics of inverted repeats seems a general genomic dynamic trait (Cáceres et al. 2007; Fiston-Lavier et al. 2007; Jurka et al. 2007)

Thus, we can imagine ectopic recombination and genomic conversion would be acting among all *Galileo* copies and different products may appear, among them the chimeric elements. In these cases, if one of the exchange breakpoints (of the conversion tract) is located inside the element, it would generate a chimeric element with two well-defined segments from two different subfamilies. These chimeric copies resemble the *Galileo* copies found in the breakpoints of polymorphic inversions in *D. buzzatii*, what is in agreement with the *Galileo* inversion generations due to ectopic recombination attraction (Cáceres et al. 1999; Casals et al. 2003; Delprat et al. 2009). Furthermore, if the two exchange breakpoints are located inside the element, this would render, for example, the X-E-X copies and, probably, this could be the origin of the whole E subfamily as well.

We would like to propose that long TIR, although they imply a handicap for the transposition reaction (Atkinson & Chalmers 2010), they could be useful for the survival of the transposon: the more recombination rate among these sequences due to the length of the TIRs, the more chance to appear a new *Galileo* subfamily. There would be more raw material where the transposase could choose from and a new

transposition burst would be triggered. The TIR length dynamics, along with the chimeric origin observed among *Galileo* copies is in agreement with an important dynamic DNA exchange of sequences and recombination (Bailey & Eichler 2006; Cáceres et al. 2007; Fiston-Lavier et al. 2007). Thus, this would explain why different non-related class II transposon present subfamilies with long TIR and why TIR length is not a reliable feature for transposon classification (Ivics et al. 1997; Cheng et al. 2000; Moschetti et al. 2008; Marquez & Pritham 2010).

Generally, the mutations or inactivation of the transposase sequence drives the death of a transposon, because without the transposition reaction there is no duplication of the sequences. The fact that we have not found any *Galileo* functional transposase, points out that *Galileo* may be an inactive element. However, our *Galileo* sequences lineages through time (LTT) plot, where the accumulation of nodes in the tree is depicted, did no show any decrease or stationary rate of *Galileo* sequences duplication. Thus, if *Galileo* is not still active, it has stopped working quite recently. In this regard, it is worth to mention that in genome sequencing projects, there are heterochromatic regions that have not been sequenced. Furthermore, there is a lot of variability among the individuals of a species which it is not represented by only one genome sequence. Then, we cannot discard the existence of *Galileo* active sequences in other individuals or other genomic regions of *D. mojavensis*.

## 3.6.- Supplementary material

Supporting tables list

SI Table 3.1. Summary table of the copies found (groups, structures and TIR length) and statistical tests.

SI Table 3.2. Detailed data of the *Galileo* copies included in this study.

SI Table 3.3. Interchromosome distribution of *Galileo* elements.

SI Table 3.4. Intrachromosome distribution of *Galileo* elements and statistical tests.

SI Table 3.5. Nearest genes to *Galileo* copies.

SI Table 3.6. Intronic *Galileo* copies.

SI Table **3.1**. Copy total element length and TIR length in the different *Galileo* subfamilies and subgroups.

**C**

| | Total length | | | TIR length | | |
|---|---|---|---|---|---|---|
| | N | Mean | Std. Dev. | N | Mean | Std. Dev. |
| Nearly complete (>2 kb TPase) | 2 | 5912.5 | 108.19 | 2 | 704.5 | 82.73 |
| Nearly complete deletion derivatives | 4 | 4070 | 1185.41 | 3 | 730.67 | 34.00 |
| 2 TIR | 5 | 1383.8 | 530.50 | 5 | 318.3 | 242.31 |
| 2 TIR longer | 2 | 3119 | 0 | 2 | 1107 | 0 |
| solo TIR | 6 | 772.5 | 171.47 | - | - | - |
| Total | 19 | 2504.5 | | | 617.208 | |

**D**

| | Total length | | | TIR length | | |
|---|---|---|---|---|---|---|
| | N | Mean | Std. Dev. | N | Mean | Std. Dev. |
| Nearly complete (>2 kb TPase) | 5 | 5283.8 | 657.41 | 5 | 545.2 | 41.482 |
| Nearly complete deletion derivatives | 2 | 3286 | 147.08 | 0 | 0 | 0 |
| 2 TIR | | 0 | 0 | 0 | 0 | 0 |
| 2 TIR longer | 2 | 1860.5 | 443.36 | 2 | 735.5 | 392.44 |
| solo TIR | 10 | 552.2 | 146.68 | - | - | - |
| Total | 19 | 2222.67 | | | 599.57 | |

**E**

| | Total length | | | TIR length | | |
|---|---|---|---|---|---|---|
| | N | Mean | Std. Dev. | N | Mean | Std. Dev. |
| Nearly complete (>2 kb TPase) | 0 | 0 | 0 | 0 | 0 | 0 |
| Nearly complete deletion derivatives | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 TIR | 7 | 1424.86 | 695.49 | 7 | 289.07 | 225.93 |
| 2 TIR longer | 22 | 2114.045 | 369.76 | 22 | 907.21 | 210.37 |
| solo TIR | 19 | 778.90 | 285.43 | - | - | - |
| Total | 48 | 1469.29 | | | 758 | |

**F**

| | Total length | | | TIR length | | |
|---|---|---|---|---|---|---|
| | N | Mean | Std. Dev. | N | Mean | Std. Dev. |
| Nearly complete (>2 kb TPase) | 1 | 0 | 0 | 1 | 733 | 0 |
| Nearly complete deletion derivatives | 1 | 0 | 0 | 0 | 0 | 0 |
| 2 TIR | 28 | 1424.86 | 695.49 | 28 | 709.88 | 308.85 |
| 2 TIR longer | 3 | 2114.046 | 369.76 | 3 | 1086.83 | 180.03 |
| solo TIR | 26 | 778.90 | 285.43 | - | - | - |
| Total | 59 | 1528.42 | | | 776 | |

**X**

| | Total length | | | TIR length | | |
|---|---|---|---|---|---|---|
| | N | Mean | Std. Dev. | N | Mean | Std. Dev. |
| Nearly complete (>2 kb TPase) | 1 | 5047 | 0 | 1 | 147.5 | 0 |
| Nearly complete deletion derivatives | 2 | 2249.5 | 245.37 | 2 | 168 | 0 |
| 2 TIR | 3 | 1262.33 | 666.27 | 3 | 311.67 | 192.84 |
| 2 TIR longer | 4 | 1723.25 | 77.66 | 4 | 581.75 | 28.01 |
| solo TIR | 3 | 517 | 209.45 | - | - | - |
| Total | 13 | 1675.15 | | | 374.55 | |

**Chimeric**

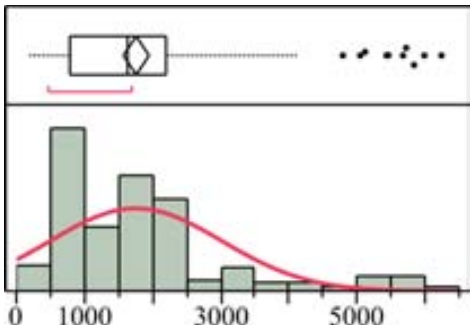
| | Total length | | | TIR length | | |
|---|---|---|---|---|---|---|
| | N | Mean | Std. Dev. | N | Mean | Std. Dev. |
| Nearly complete (>2 kb TPase) | 1 | 6239 | 0 | 1 | 873.5 | 0 |
| Nearly complete deletion derivatives | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 TIR | 6 | 1769.17 | 389.7474 | 6 | 599.67 | 196.51 |
| 2 TIR longer | 5 | 1903.6 | 576.99 | 5 | 491.3 | 252.22 |
| solo TIR | - | - | - | - | - | - |
| Total | 12 | 2197.67 | | | 528.65 | |

**Total**

| | Total length | | | TIR length | | |
|---|---|---|---|---|---|---|
| | N | Mean | Std. Dev. | N | Mean | Std. Dev. |
| Nearly complete (>2 kb TPase) | 10 | 5356.6 | 745.61 | 10 | 588.9 | 196.24 |
| Nearly complete deletion derivatives | 9 | 3436.11 | 1047.91 | 5 | 505.6 | 309.12 |
| 2 TIR | 49 | 1738.88 | 562.31 | 49 | 571.93 | 322.95 |
| 2 TIR longer | 38 | 2139.5 | 497.62 | 38 | 833.88 | 271.38 |
| solo TIR | 64 | 741.47 | 234.82 | 0 | 0 | 0 |
| | 170 | 1755.59 | 1259.58 | 102 | 667.93 | 317.045 |

Statistical Tests

1. Total *Galileo* length.




*Galileo* length distribution

Fitted Normal

Parameter Estimates

| Type | Parameter | Estimate | Lower 95% | Upper 95% |
|---|---|---|---|---|
| Location | μ | 1755.5941 | 1564.8851 | 1946.3031 |
| Dispersion | σ | 1259.5819 | 1138.4166 | 1409.8387 |

-2log(Likelihood) = 2908.54105324984

**Goodness-of-Fit Test: Shapiro-Wilk W Test**

| W | Prob<W |
|---|---|
| 0.834216 | <.0001* |

Ho = The data is from the Normal distribution. Small p-values reject Ho.

### *Galileo* length by *Galileo* subfamily

Means Comparisons: Comparisons for all pairs using Tukey-Kramer HSD

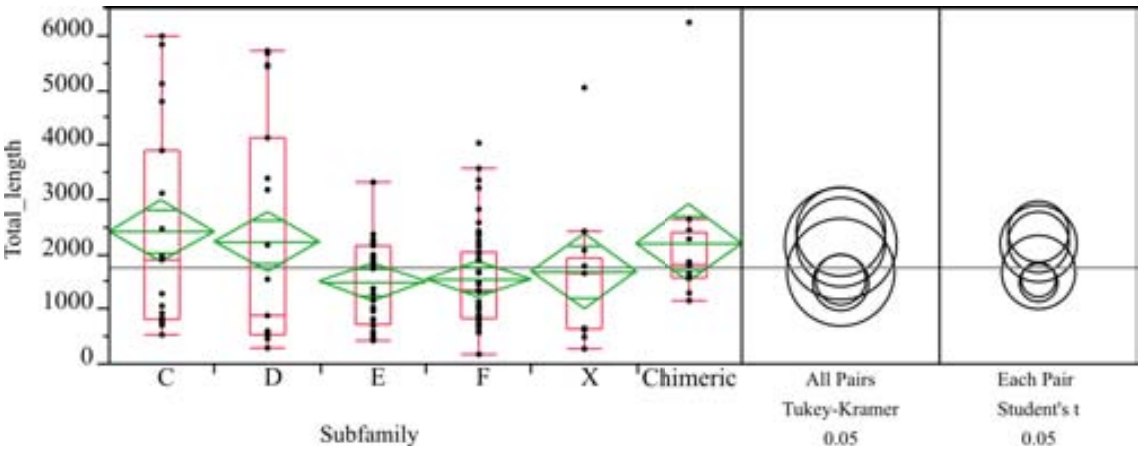
Abs(Dif)-LSD

| | C | D | Chimeric | X | F | E |
|---|---|---|---|---|---|---|
| C | -1148.26 | -955.473 | -1087.05 | -533.405 | -58.4314 | -28.685 |
| D | -955.473 | -1148.26 | -1279.84 | -726.195 | -251.221 | -221.474 |
| Chimeric | -1087.05 | -1279.84 | -1444.87 | -894.295 | -463.593 | -429.642 |
| X | -533.405 | -726.195 | -894.295 | -1388.18 | -949.693 | -916.45 |
| F | -58.4314 | -251.221 | -463.593 | -949.693 | -651.617 | -632.487 |
| E | -28.685 | -221.474 | -429.642 | -916.45 | -632.487 | -722.433 |

Positive values show pairs of means that are significantly different.

| Level | | Mean |
|---|---|---|
| C | A | 2415.6316 |
| D | A | 2222.8421 |
| Z.Chimeric | A | 2197.6667 |
| X | A | 1675.1538 |
| F | A | 1540.4915 |
| E | A | 1485.0417 |

Levels not connected by same letter are significantly different.




Total length distribution in the different *Galileo* subfamilies of *D. mojavensis*.

### *Galileo* length by *Galileo* structural type

Means Comparisons: Comparisons for all pairs using Tukey-Kramer HSD

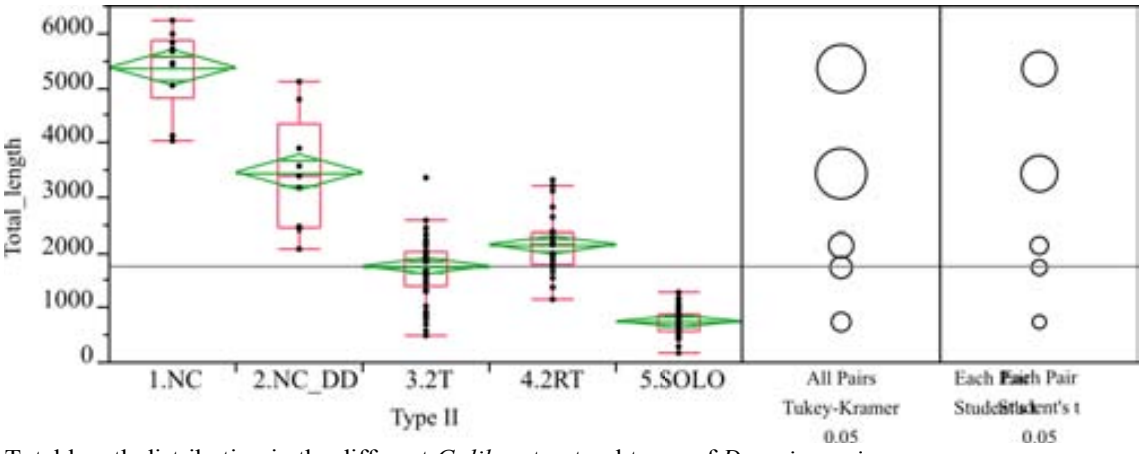
Abs(Dif)-LSD

| | 1.NC | 2.NC_DD | 4.Longer_2TIR | 3.2TIR | 5.SOLO |
|---|---|---|---|---|---|
| 1.NC | -619.371 | 1284.146 | 2724.874 | 3137.145 | 4144.195 |
| 2.NC_DD | 1284.146 | -652.874 | 783.1924 | 1194.971 | 2201.598 |
| 4.Longer_2TIR | 2724.874 | 783.1924 | -317.731 | 101.2543 | 1114.4 |
| 3.2TIR | 3137.145 | 1194.971 | 101.2543 | -279.803 | 734.511 |
| 5.SOLO | 4144.195 | 2201.598 | 1114.4 | 734.511 | -244.828 |

Positive values show pairs of means that are significantly different.

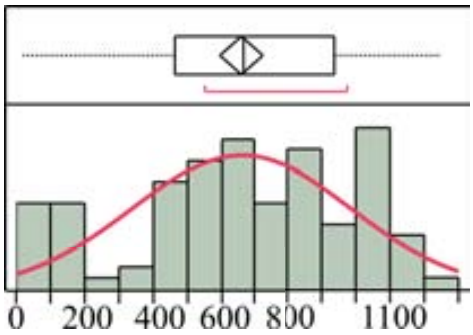
| Level | | Mean |
|---|---|---|
| 1.NC | A | 5356.6 |
| 2.NC_DD | B | 3436.1111 |
| 4.2RT | C | 2139.5 |
| 3.2T | D | 1738.8776 |
| 5.SOLO | E | 741.4688 |

Levels not connected by same letter are significantly different.




Total length distribution in the different *Galileo* structural types of *D. mojavensis*.

1. *Galileo* TIR length.




TIR length distribution

Fitted Normal

Parameter Estimates

| Type | Parameter | Estimate | Lower 95% | Upper 95% |
|---|---|---|---|---|
| Location | μ | 661.95146 | 599.16181 | 724.74111 |
| Dispersion | σ | 321.27395 | 282.58952 | 372.32654 |

-2log(Likelihood) = 2908.54105324984

**Goodness-of-Fit Test: Shapiro-Wilk W Test**

| W | Prob<W |
|---|---|
| 0.954506 | 0.0014* |

Ho = The data is from the Normal distribution. Small p-values reject Ho.

**TIR length by *Galileo* subfamily**

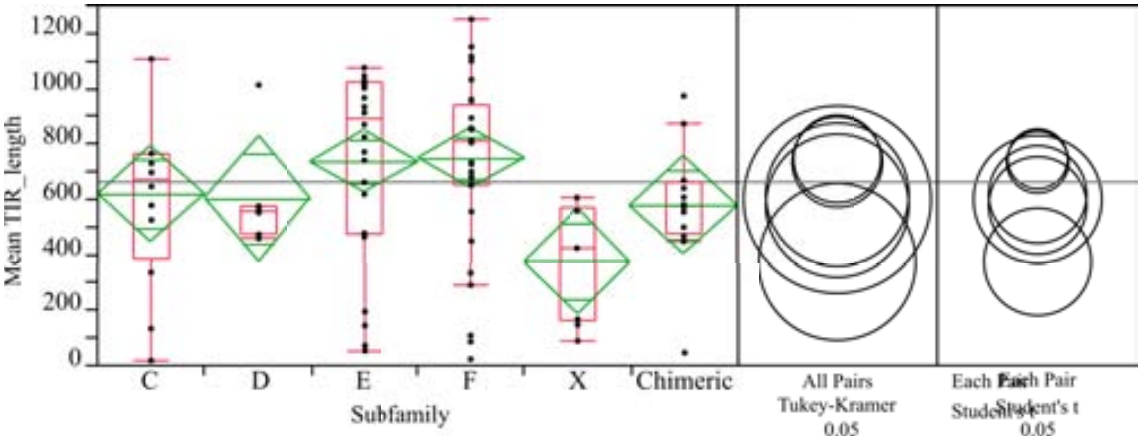
Means Comparisons: Comparisons for all pairs using Tukey-Kramer HSD

Abs(Dif)-LSD

| | F | E | C | D | Z.Chimeric | X |
|---|---|---|---|---|---|---|
| F | -223.629 | -215.854 | -174.065 | -226.88 | -134.19 | 47.31895 |
| E | -215.854 | -230.963 | -188.276 | -240.578 | -148.401 | 33.28627 |
| C | -174.065 | -188.276 | -365.184 | -407.789 | -325.309 | -140.35 |
| D | -226.88 | -240.578 | -407.789 | -478.138 | -403.188 | -215.8 |
| Z.Chimeric | -134.19 | -148.401 | -325.309 | -403.188 | -365.184 | -180.225 |
| X | 47.31895 | 33.28627 | -140.35 | -215.8 | -180.225 | -400.039 |

| Level | | Mean |
|---|---|---|
| F | A | 745.9375 |
| E | A | 734.46667 |
| C | AB | 617.20833 |
| D | AB | 599.57143 |
| Z.Chimeric | AB | 577.33333 |
| X | B | 374.55 |

Levels not connected by same letter are significantly different.




**TIR length by Structural Type**

Means Comparisons: Comparisons for all pairs using Tukey-Kramer HSD

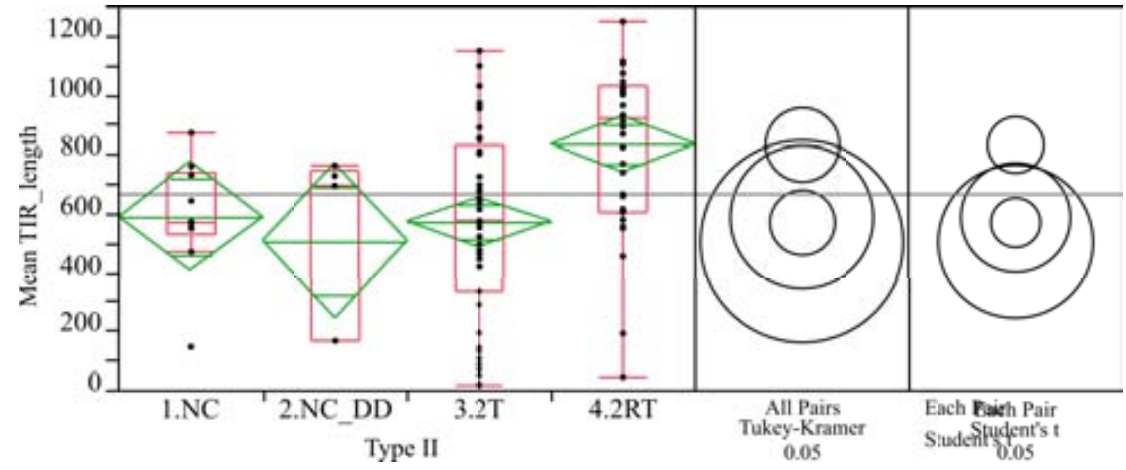
Abs(Dif)-LSD

| | 4.2RT | 1.NC | 3.2T | 2.NC_DD |
|---|---|---|---|---|
| 4.2RT | -176.177 | -27.9503 | 95.9579 | -37.0464 |
| 1.NC | -27.9503 | -343.432 | -249.502 | -337.316 |
| 3.2T | 95.9579 | -249.502 | -155.147 | -294.2 |
| 2.NC_DD | -37.0464 | -337.316 | -294.2 | -485.686 |

Positive values show pairs of means that are significantly different.

| Level | | Mean |
|---|---|---|
| 4.2RT | A | 833.88158 |
| 1.NC | AB | 588.9 |
| 3.2T | B | 571.92857 |
| 2.NC_DD | AB | 505.6 |

Levels not connected by same letter are significantly different.



SI Table **3.2**. *Galileo* copies

I. Nearly Complete copies

| TAG | Scaffold | Start | End | Subfamily | TSD1 | TSD2 | Contig | Start | End | Total_length | TIR1 length | TIR2 length | TIR identity | TPase |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 6500-30856C | 6500 | 30856283 | 30862823 | C | GTTACCG | GTTACCG | 10758 | 37586 | 44126 | 5989 | 813 | 713 | 98.7 | 2784 |
| 6500-31288C | 6500 | 31288762 | 31295303 | C | ATGGAGA | TATTGAC | 10770 | 9949 | 16490 | 5836 | 583 | 709 | 61.9 | 2828 |
| 6541-11419q | 6541 | 1141978 | 1149130 | Chimeric F/C | GTAGAAC | GTATGGT | 11233 | 4001 | 11153 | 6239 | 788 | 959 | 80.1 | 2808 |
| 6485-39163D | 6485 | 39163 | 45738 | D | GTCCAAG | ATTTAAG | 9930 | 1467 | 8042 | 5675 | 574 | 576 | 99.3 | 2814 |
| 6498-23860D | 6498 | 2386095 | 2392524 | D | TAATAAA | TAATAAA | 10376 | 4316 | 10745 | 5721 | 570 | 570 | 100 | 2785 |
| 6500-31458D | 6500 | 31458921 | 31465167 | D | - | TTTATAT | 10773 | 33627 | 39873 | 4130 | 570 | 376 | 94.1 | 2228 |
| 6540-11758D | 6540 | 1175880 | 1182997 | D | CTGAATC | CTAAATC | 10946 | 6739 | 13856 | 5433 | 525 | 578 | 89.3 | 2553 |
| 6541-16442D | 6541 | 1644296 | 1649755 | D | AATGTAT | AATGTAT | 11255 | 1328 | 6768 | 5460 | 557 | 556 | 99.1 | 2550 |
| 6498-22531F | 6498 | 2253149 | 2269701 | F | CCTGAAC | GTAGCAG | 10369 | 31739 | 35574 | 4036 | 693 | 773 | 95.4 | 2047 |
| 6540-41449X | 6540 | 414493 | 419539 | X | - | CTTAAAT | 10924 | 25932 | 30978 | 5047 | 127 | 168 | 98.4 | 2822 |

II. Nearly Complete Deletion Derivatives copies

| TAG | Scaffold | Start | End | Subfamily | TSD1 | TSD2 | Contig_1 | Start | End | Total_length | TIR1 length | TIR2 length | TIR identity | Tpase |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 6262-13889C | 6262 | 13889 | 19752 | C | no | no | 7794 | 13889 | 19752 | 3899 | 526 | 0 | 0 | 2698 |
| 6358-1C | 6358 | 1 | 5345 | C | no | GTACAAT | 8435<br>8436 | 1<br>1 | 4274<br>736 | 4793 | 732 | 662 | 99.4 | 1949 |
| 6500-31981C | 6500 | 31981325 | 31986443 | C | AATATAT | AATATAT | 10792 | 22486 | 27604 | 5119 | 815 | 645 | 91.9 | 1951 |
| 6482-61400D | 6482 | 614003 | 617184 | D | no | no | 9847 | 20748 | 23929 | 3182 | 254 | 0 | 0 | 1739 |
| 6482-61718D | 6482 | 617185 | 621442 | D | no | no | 9847 | 23930 | 25156 | 3390 | 0 | 0 | 0 | 2704 |
| 6482-60893F | 6482 | 608936 | 612509 | F | GCGCTAT | no | 9847 | 15681 | 19245 | 3574 | 911 | 0 | 0 | 2323 |
| 6406-4469X | 6406 | 4469 | 6544 | X | GCCTTAG | GCCTTAG | 8836 | 146 | 2221 | 2076 | 168 | 168 | 97.6 | 284 |
| 6498-25411X | 6498 | 2541172 | 2544793 | X | CTTGTAC | CTTGTAC | 10383 | 26876 | 30497 | 2423 | 168 | 168 | 99.4 | 276 |

SI Table 3.2. Continuation. III. 2-TIR copies

| TAG | Scaffold | Start | End | Subfamily | TSD1 | TSD2 | Contig_1 | Start | End | Total_length | TIR1 length | TIR2 length | TIR identity |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 6433-41007C | 6433 | 41007 | 42966 | C | ATACAAC | ATACAAC | 8990<br>8989 | 1<br>4387 | 1321<br>5010 | 929 | 335 | 339 | 98.2 |
| 6473-11762C | 6473 | 11762829 | 11764731 | C | ATTTGAA | ATTTGAA | 9647 | 4614 | 6516 | 1903 | 578 | 578 | 100 |
| 6500-31884C | 6500 | 31884435 | 31886401 | C | ATACTAC | ATACTAC | 10790 | 49101 | 51067 | 1967 | 556 | 494 | 81.6 |
| 6540-61321C | 6540 | 6132112 | 6133394 | C | GTCTGGC | GTCTGGC | 10985 | 125776 | 127058 | 1283 | 18 | 18 | 100 |
| 6680-24265C | 6680 | 24265741 | 24266577 | C | GTTCGGC | GTTCGGC | 11684 | 12780 | 13616 | 837 | 133 | 134 | 94 |
| 6482-26902q | 6482 | 269026 | 270625 | Chimeric E/C | GTGATAT | AATACAC | 9832 | 28557 | 30156 | 1600 | 577 | 565 | 67 |
| 6500-30179q | 6500 | 30179877 | 30181717 | Chimeric F/C | GTAGTAT | CGTAGAT | 10737 | 2699 | 4539 | 1841 | 432 | 566 | 53.9 |
| 6500-30733q | 6500 | 30733241 | 30734538 | Chimeric D/F | no | TCTTTGG | 10753 | 9551 | 10848 | 1298 | 399 | 535 | 60.5 |
| 6540-55852q | 6540 | 558528 | 561650 | Chimeric E/F | - | ATTTTAG | 10925 | 93044 | 96166 | 2443 | 897 | 1050 | 43.5 |
| 6541-16186q | 6541 | 1618681 | 1620248 | Chimeric F/C | CTTTTAG | GTAACAC | 11252 | 5173 | 6740 | 1568 | 754 | 526 | 80.7 |
| 6541-16912q | 6541 | 1691275 | 1695391 | Chimeric F/C | ATATAAC | GTCTTAA | 11256 | 5973 | 10089 | 1865 | 441 | 454 | 76.5 |
| 4124-318E | 4124 | 318 | 3097 | E | GTAGTAA | GTAGTAA | 4796<br>4795 | 1<br>318 | 3832<br>880 | 1866 | 676 | 559 | 99.1 |
| 4198-1393E | 4198 | 1393 | 3341 | E | GCTATAC | GCTATAC | 4932<br>4931 | 1<br>1393 | 914<br>1464 | 1949 | 72 | 72 | 100 |
| 4502-5732E | 4502 | 5732 | 6311 | E | GTTGTAT | CCTTAAT | 5475 | 5732 | 6311 | 580 | 195 | 195 | 95.9 |
| 6115-956E | 6115 | 956 | 2582 | E | ATATGGC | ATATGGC | 7618 | 956 | 2582 | 1927 | 477 | 478 | 96.4 |
| 6482-45393E | 6482 | 453934 | 454728 | E | GTCAGAC | GTCAGAC | 9840 | 16922 | 17716 | 795 | 53 | 53 | 98.1 |
| 6498-19996E | 6498 | 1999631 | 2001782 | E | ATATAAG | ATATAAG | 10352 | 110278 | 112924 | 697 | 145 | 145 | 99.3 |
| 6500-31360E | 6500 | 31360321 | 31362480 | E | - | GTTTTAT | 10770 | 81508 | 83667 | 2160 | 481 | 446 | 74.4 |
| 1776-3477F | 1776 | 3477 | 4795 | F | GAAGAAC | - | 1899 | 3477 | 4170 | 1319 | 291 | 291 | 90.8 |
| 3792-475F | 3792 | 475 | 3832 | F | GTACCGC | GTACCGC | 4241 | 475 | 3832 | 2193 | 846 | 1075 | 98.1 |
| 6473-16293F | 6473 | 16293472 | 16295724 | F | ATACAAT | ATTACAC | 9762<br>9763 | 19387<br>1 | 20553<br>656 | 2253 | 651 | 650 | 99.7 |
| 6482-21925F | 6482 | 2192546 | 2194452 | F | ATTTGAT | ATTTGAT | 9896 | 8068 | 9974 | 1907 | 812 | 812 | 100 |
| 6482-25792F | 6482 | 2579294 | 2581638 | F | ATTGAGT | ATTGAGT | 9911 | 1113 | 3457 | 2345 | 1032 | 1032 | 99.7 |
| 6496-25846F | 6496 | 25846816 | 25848819 | F | ACTCTAT | ACTCTAT | 10271<br>10270 | 1<br>145320 | 1156<br>146046 | 2004 | 727 | 727 | 100 |

SI Table 3.2. Continuation. III. 2-TIR copies

| TAG | Scaffold | Start | End | Subfamily | TSD1 | TSD2 | Contig_1 | Start | End | Total_length | TIR1 length | TIR2 length | TIR identity |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 6497-23418F | 6497 | 234188 | 236613 | F | GTACCGC | GTACCGC | 10309 | 10744 | 13169 | 2426 | 1147 | 1153 | 98.8 |
| 6498-18188F | 6498 | 1818872 | 1820002 | F | ATCAAAT | GTGAAAC | 13050 | 8248 | 9378 | 1032 | 23 | 23 | 91.3 |
| 6498-29668F | 6498 | 2966893 | 2969965 | F | GTAATAG | - | 10404 | 26506 | 29578 | 2137 | 665 | 646 | 82.8 |
| 6498-32815F | 6498 | 3281538 | 3283440 | F | GTAGTAT | GTAGTAT | 10415 | 25282 | 27184 | 1903 | 810 | 810 | 99.8 |
| 6500-30329F | 6500 | 30329586 | 30332946 | F | CTATAAC | TACATAT/TGCTAAT | 10741 | 55570 | 58930 | 3361 | 985 | 1214 | 97.8 |
| 6500-30494F | 6500 | 30494802 | 30496424 | F | ATTTTAC | TTACGCA | 10742 | 42243 | 43865 | 1344 | 396 | 502 | 91.7 |
| 6500-30596F | 6500 | 30596827 | 30599409 | F | GTCGTGG | GTCGTGG | 10744 | 15274 | 17856 | 2583 | 1264 | 1038 | 99.3 |
| 6500-30684F | 6500 | 30684259 | 30686266 | F | ATAGCGT | TTGAACC | 10751 | 6539 | 8546 | 2008 | 753 | 957 | 97.4 |
| 6500-30873F | 6500 | 30873698 | 30875357 | F | GTTATGC | GTTATGC | 10758 | 55001 | 56660 | 1660 | 485 | 626 | 91.7 |
| 6500-30976F | 6500 | 30976506 | 30978415 | F | ATAGTAG | ATAGTAG | 10762 | 1 | 940 | 1910 | 944 | 760 | 93.1 |
| | | | | | | | 10761 | 9087 | 10031 | | | | |
| 6500-31107F | 6500 | 31107017 | 31109152 | F | CTTAAAT | CTTAAAT | 10764 | 53125 | 53803 | 2136 | 677 | 679 | 99.8 |
| | | | | | | | 10765 | 1 | 1465 | | | | |
| 6500-31694F | 6500 | 31694898 | 31696939 | F | no | GTATCAG | 10387 | 1 | 2042 | 2042 | 1020 | 892 | 99.5 |
| 6540-56432F | 6540 | 564326 | 566215 | F | TGTACAT | ATGTACA | 10925 | 98843 | 100731 | 1889 | 803 | 806 | 99 |
| 6540-79670F | 6540 | 796704 | 798194 | F | GTTCGTG | CCAGACA | 10931 | 1 | 676 | 1491 | 108 | 109 | 30.2 |
| | | | | | | | 10930 | 17100 | 17422 | | | | |
| 6540-57500F | 6540 | 5750063 | 5751976 | F | CTTTAAC | CTTTAAC | 10984 | 336595 | 338508 | 1914 | 892 | 892 | 100 |
| 6540-33286F | 6540 | 33286261 | 33288174 | F | ATAAAAA | ATAAAAA | 11176 | 117065 | 118978 | 1914 | 892 | 892 | 100 |
| | | | | | | | 11193 | 1 | 975 | | | | |
| 6541-17831F | 6541 | 178316 | 180001 | F | ATAAGAC | ATAAGAC | 11192 | 15669 | 16353 | 1686 | 682 | 683 | 98.7 |
| 6541-10035F | 6541 | 1003587 | 1007838 | F | ATATAAG | CCCATAT | 11229 | 9895 | 14146 | 1876 | 712 | 688 | 93 |
| 6541-12491F | 6541 | 1249195 | 1251094 | F | CTAATAT | CTAATAT | 11238 | 17034 | 18933 | 1900 | 811 | 811 | 99.1 |
| 6541-15113F | 6541 | 1511326 | 1513666 | F | CTTTGTG | CTTTGTG | 11249 | 1 | 1329 | 2341 | 751 | 964 | 95.3 |
| | | | | | | | 11248 | 12735 | 13721 | | | | |
| 6680-24420F | 6680 | 24420206 | 24421090 | F | GTAGTAT | GTAGTAT | 11687 | 39553 | 40437 | 885 | 86 | 86 | 93 |

SI Table 3.2. Continuation. III. 2-TIR copies

| TAG | Scaffold | Start | End | Subfamily | TSD1 | TSD2 | Contig_1 | Start | End | Total_length | TIR1 length | TIR2 length | TIR identity |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 6680-24422F | 6680 | 24422223 | 24425258 | F | GTACACA | GTACACA | 11687 | 41570 | 44605 | 1451 | 335 | 335 | 97 |
| 6498-95069X | 6498 | 950693 | 951185 | X | TCCATAT | TCCATAT | 10339 | 90483 | 90975 | 493 | 90 | 88 | 98.8 |
| 6498-29033X | 6498 | 2903343 | 2904985 | X | CTTATAT | CTTATAT | 10400 | 3858 | 5500 | 1643 | 423 | 423 | 100 |
| 6500-29395X | 6500 | 29395284 | 29396934 | X | ATAATAC | TATAAAC | 10722 | 198610 | 200260 | 1651 | 423 | 423 | 99.1 |

SI Table 3.2. Continuation. IV. 2 Recombinant TIR or 2 longer TIR copies.

| TAG | Scaffold | Start | End | Subfamily | TSD1 | TSD2 | Contig_1 | Start | End | Total_length | TIR1 length | TIR2 length | TIR identity | TPase |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 6498-21770C | 6498 | 2177013 | 2181005 | C | GTTGAGC | GTTGAGC | 10367 | 3528 | 7520 | 3119 | 1107 | 1107 | 99.1 | 599 |
| 6500-31083C | 6500 | 31083091 | 31089863 | C | TTTATAT | TTTATAT | 10764 | 29199 | 35971 | 2469 | 815 | 715 | 86.4 | 57 |
| 6500-31371C | 6500 | 31371001 | 31374119 | C | ATAGTAG | CTACTAT | 10770 | 92188 | 95306 | 3119 | 1107 | 1107 | 98.7 | 602 |
| 6500-29973q | 6500 | 29973001 | 29975284 | Chimeric F/X | TAGGTAA | ATACAAC | 10735 | 14871 | 17154 | 2284 | 590 | 746 | 49.8 | 0 |
| 6500-30183q | 6500 | 30183437 | 30184591 | Chimeric F/X | ATAATAC | CATATAT | 10737 | 6259 | 7413 | 1155 | 449 | 714 | 59.3 | 0 |
| 6541-99710q | 6541 | 997100 | 998743 | Chimeric E/F | ACCATAC | GTACAGC | 11229 | 3408 | 5051 | 1644 | 46 | 47 | 87.2 | 0 |
| 6680-24427q | 6680 | 24427759 | 24434511 | Chimeric E/X | TTTGGGT | ATGTTAA | 11687<br>11688 | 47106<br>1 | 52345<br>552 | 2643 | 555 | 552 | 85.5 | 0 |
| 6680-24440q | 6680 | 24440484 | 24442275 | Chimeric E/X | TTTTGGT | ATGTTAA | 11688 | 6525 | 8316 | 1792 | 608 | 606 | 86.8 | 0 |
| 6540-10358D | 6540 | 1035815 | 1037361 | D | ATTGGGG | ATTGGGG | 10940 | 39859 | 41405 | 1547 | 458 | 458 | 96.5 | 0 |
| 6680-24069D | 6680 | 24069812 | 24072777 | D | TTATGAG | TTATGAG | 11679 | 95650 | 98615 | 2174 | 1012 | 1014 | 98.9 | 0 |
| 4503-1178E | 4503 | 1178 | 7564 | E | TCGTGAC | TCGTGAC | 5476 | 1178 | 7564 | 1991 | 954 | 979 | 98.1 | 0 |
| 6395-2229E | 6395 | 2229 | 6819 | E | CTATAAC | CTATAAC | 8783 | 2229 | 6819 | 2151 | 991 | 1028 | 95.6 | 0 |
| 6473-10080E | 6473 | 1008070 | 1010341 | E | GCGCTGA | GCGCTGA | 9253<br>9252<br>9834 | 1<br>9335<br>20890 | 1207<br>10394<br>25520 | 2272 | 1002 | 1002 | 99.7 | 0 |
| 6482-36252E | 6482 | 362528 | 364455 | E | - | GTTTAC | 9835 | 327 | 959 | 1928 | 741 | 741 | 98.4 | 0 |
| 6496-15292E | 6496 | 15292514 | 15294879 | E | TAAGTGG | TAAGTGG | 10177 | 520481 | 522846 | 2366 | 1075 | 1075 | 100 | 0 |
| 6498-95355E | 6498 | 953555 | 955919 | E | GCCAAAG | GCCAAAG | 10339 | 93345 | 95709 | 2365 | 1075 | 1074 | 99.9 | 0 |
| 6498-29938E | 6498 | 2993866 | 2995242 | E | CTTGTAC | CTTGTAC | 10405 | 19460 | 20836 | 1377 | 193 | 193 | 100 | 0 |
| 6500-29804E | 6500 | 29804958 | 29807163 | E | TCATTAC | TCATTAC | 10727 | 44829 | 47034 | 2206 | 1029 | 1027 | 99.6 | 0 |
| 6500-30306E | 6500 | 30306361 | 30308225 | E | GTGGTAT | GTGGTAT | 10741 | 32345 | 34209 | 1865 | 645 | 678 | 96 | 0 |
| 6500-30702E | 6500 | 30702790 | 30704984 | E | TGTATAC | TGTATAC | 10751 | 25070 | 27264 | 2195 | 1021 | 1026 | 99.3 | 0 |
| 6500-31202E | 6500 | 31202553 | 31204509 | E | ACATCAA | ACATCAA | 10766<br>10767 | 25837<br>1 | 26845<br>913 | 1957 | 832 | 819 | 96.6 | 0 |
| 6500-31506E | 6500 | 31506397 | 31509717 | E | GTAAAAA | GTAAAAC | 10774 | 18407 | 21727 | 3321 | 1048 | 1043 | 97.4 | 0 |
| 6500-31516E | 6500 | 31516211 | 31518161 | E | ATACTAG | GTACAGG | 10774<br>10775 | 28221<br>1 | 29239<br>907 | 1951 | 962 | 906 | 92.2 | 0 |
| 6500-31920E | 6500 | 31920296 | 31922494 | E | TCGAAAC | TCGAAAC | 10790 | 84962 | 87962 | 2199 | 1028 | 1022 | 98.2 | 0 |
| 6500-32268E | 6500 | 32268286 | 32271024 | E | ATTATAG | ATTATAG | 10803 | 46034 | 48772 | 2184 | 1016 | 1017 | 99.2 | 0 |

SI Table 3.2. Continuation. IV. Recombinant TIR or 2 longer TIR copies (continuation).

| TAG | Scaffold | Start | End | Subfamily | TSD1 | TSD2 | Contig_1 | Start | End | Total_length | TIR1 length | TIR2 length | TIR identity | Tpase |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 6540-11650E | 6540 | 1165033 | 1168349 | E | GATACAC | GATACAC | 10945 | 2811 | 6127 | 2180 | 998 | 1034 | 96.2 | 0 |
| 6540-13720E | 6540 | 1372066 | 1373808 | E | ATATAAT | ATATAAT | 10949 | 9737 | 1479 | 1743 | 771 | 771 | 99.2 | 0 |
| 6540-14510E | 6540 | 14510521 | 14512886 | E | CTTTTGT | CTTTTGT | 11044 | 136297 | 138662 | 2366 | 1075 | 1075 | 100 | 0 |
| 6540-31163E | 6540 | 31163990 | 31166355 | E | CTTAAAC | TTAGTGC | 11157 | 380472 | 382837 | 2366 | 1075 | 1075 | 99.3 | 0 |
| 6541-10420E | 6541 | 1042036 | 1043771 | E | TTAATGC | TTAATGC | 11229 | 48344 | 50079 | 1736 | 769 | 971 | 96.5 | 0 |
| 6541-10885E | 6541 | 1088506 | 1090501 | E | ATAGAGC | ATAGAGC | 11232<br>11231 | 1<br>24008 | 1014<br>24917 | 1996 | 910 | 915 | 98.1 | 0 |
| 6541-20142E | 6541 | 2014221 | 2016575 | E | GTATCAA | GTATCAA | 11267 | 12789 | 15143 | 1794 | 613 | 623 | 98.1 | 0 |
| 6328-16507F | 6328 | 1650720 | 1653101 | F | GTGCAGC | GTGCAGC | 8189 | 12012 | 14393 | 2382 | 894 | 894 | 99.9 | 0 |
| 6496-23195F | 6496 | 23195067 | 23197892 | F | GTATTTT | GTATTTT | 10246 | 221912 | 224737 | 2826 | 1116 | 1116 | 99.6 | 211 |
| 6500-29864F | 6500 | 29864896 | 29868109 | F | GTATTAT | GTATTAT | 10727 | 104767 | 107980 | 3214 | 1260 | 1241 | 95.2 | 0 |
| 6500-30351X | 6500 | 30351497 | 30353286 | X | CTATAAC | CTATAAC | 10741 | 77481 | 79270 | 1790 | 606 | 606 | 99.5 | 0 |
| 6680-24283X | 6680 | 24283772 | 24285562 | X | GCTAAAG | ATTAAAG | 11684 | 30811 | 32601 | 1791 | 606 | 606 | 98.3 | 0 |
| 6680-24520X | 6680 | 24520907 | 24522561 | X | ATAAGAC | ATAAGAC | 11693 | 28363 | 30017 | 1655 | 548 | 566 | 95.2 | 0 |
| 6680-24538X | 6680 | 24538620 | 24540276 | X | GTTACGG | GTTACGG | 11694 | 11280 | 12936 | 1657 | 550 | 566 | 95.1 | 0 |

SI Table 3.2. Continuation. V. Solo-TIR copies

| TAG | Scaffold | Start | End | Subfamily | TSD1 | Contig_1 | Start | End | Total_length |
|---|---|---|---|---|---|---|---|---|---|
| 4159-2383C | 4159 | 2383 | 3089 | C | no | 4862 | 93 | 745 | 707 |
| 4315-36359C | 4315 | 36359 | 37136 | C | ATTTAGG | 5157 | 597 | 1374 | 778 |
| 6475-6418C | 6475 | 6418 | 6949 | C | GTTATGC | 9793 | 6418 | 6949 | 532 |
| 6500-29798C | 6500 | 29798764 | 29799506 | C | - | 10727 | 39377 | 38635 | 743 |
| 6540-59683C | 6540 | 596833 | 597648 | C | GTTGAAC | 10925 | 131349 | 132164 | 816 |
| 6540-13434C | 6540 | 13434776 | 13435834 | C | ATACCC | 11040 | 106510 | 108568 | 1059 |
| 3967-5428D | 3967 | 5428 | 5995 | D | GTATTGA | 4504 | 449 | 1016 | 568 |
| 4302-1710D | 4302 | 1710 | 2167 | D | TTCACGA | 5129 | 23 | 480 | 458 |
| 5820-1010D | 5820 | 1010 | 1528 | D | GCTTTAT | 7167 | 1010 | 1528 | 519 |
| 6115-1D | 6115 | 1 | 291 | D | ATTTAAG | 7618 | 1 | 291 | 291 |
| 6422-3900D | 6422 | 3900 | 4471 | D | TTGATGT | 8929 | 3900 | 4471 | 572 |
| 6439-76259D | 6439 | 76259 | 76829 | D | GATAAAT | 9016 | 2195 | 2765 | 571 |
| 6482-25268D | 6482 | 2526809 | 2527340 | D | CTACTAC | 9907 | 14465 | 14996 | 532 |
| 6498-25609D | 6498 | 2560957 | 2561558 | D | TCATAAC | 10383 | 46934 | 47262 | 602 |
| 6500-30590D | 6500 | 30590766 | 30591289 | D | CTTCTAG | 10744 | 9213 | 9736 | 524 |
| 6541-24219D | 6541 | 2421943 | 2422508 | D | ATCGTTC ? | 11283 | 10937 | 11502 | 885 |
| 3878-2398E | 3878 | 2398 | 2973 | E | ATAATAG | 4340 | 2398 | 2973 | 576 |
| 4315-35763E | 4315 | 35763 | 36259 | E | GCGCAAC | 9252 | 9355 | 9851 | 497 |
| 4552-6419E | 4552 | 6419 | 7232 | E | CCATAAA | 5552 | 3434 | 4074 | 814 |
| 4621-5761E | 4621 | 5761 | 6344 | E | CTTCTAG | 5655 | 474 | 1057 | 584 |
| 6070-5751E | 6070 | 5751 | 6783 | E | TCGTGAC | 7517 | 5751 | 6783 | 1033 |
| 6320-38399E | 6320 | 38399 | 38982 | E | GTTCTGC | 8092 | 3568 | 4151 | 584 |
| 6329-49349E | 6329 | 49349 | 50504 | E | TTACTAC | 8308 | 1 | 1156 | 1156 |
| 6404-43168E | 6404 | 43168 | 43745 | E | GTTGAAG | 8826 | 832 | 1409 | 578 |
| 6498-24079E | 6498 | 2407995 | 2408421 | E | GTTCTAT | 10376 | 26216 | 26642 | 427 |
| 6498-26098E | 6498 | 2609857 | 2611989 | E | GTTTTGA | 10385 | 7751 | 8354 | 1149 |
| 6498-28362E | 6498 | 2836237 | 2836700 | E | TTGAAAG | 10397 | 7407 | 7870 | 464 |

SI Table 3.2. Continuation. V. Solo-TIR copies (continuation)

| TAG | Scaffold | Start | End | Subfamily | TSD1 | Contig_1 | Start | End | Total_length |
|---|---|---|---|---|---|---|---|---|---|
| 6498-31200E | 6498 | 3120041 | 3121220 | E | CAGTTGG | 10408 | 4700 | 5879 | 1180 |
| 6500-31339E | 6500 | 31339017 | 31339980 | E | ATATTAT | 10770 | 60532 | 61167 | 964 |
| 6500-31499E | 6500 | 31499776 | 31500354 | E | CATTAAC | 10774 | 11786 | 12364 | 579 |
| 6500-31817E | 6500 | 31817847 | 31818422 | E | GTCACGA | 10789 | 10934 | 11509 | 576 |
| 6540-75029E | 6540 | 750291 | 750804 | E | ACCATAC | 10928 | 42449 | 42962 | 514 |
| 6540-89813E | 6540 | 898138 | 898666 | E | CTTATAT | 10934 | 44264 | 44792 | 529 |
| 6540-10067E | 6540 | 1006735 | 1007752 | E | no | 10940 | 10779 | 11429 | 1018 |
| 6680-23161E | 6680 | 23161869 | 23162539 | E | ATATAAG | 11659 | 26002 | 26672 | 821 |
| 6498-17302F | 6498 | 1730250 | 1731243 | F | CTGTTAC | 10349 | 8459 | 9452 | 994 |
| 6498-23818F | 6498 | 2381827 | 2382890 | F | ATTAAAT | 10376 | 48 | 1111 | 1064 |
| 6498-25144F | 6498 | 2514476 | 2515268 | F | GCAAAAT | 10383 | 180 | 972 | 793 |
| 6498-25221F | 6498 | 2522128 | 2522920 | F | GCAAAAT | 10383 | 7832 | 8624 | 793 |
| 6498-27869F | 6498 | 2786970 | 2787863 | F | ATCATAT | 10394 | 5133 | 6204 | 894 |
| 6498-30224F | 6498 | 3022490 | 3023060 | F | no | 10406 | 18750 | 19320 | 571 |
| 6500-29965F | 6500 | 29965273 | 29966306 | F | GTAGTAC | 10735 | 7404 | 7734 | 1034 |
| 6500-29967F | 6500 | 29967217 | 29968358 | F | GTGCTAT | 10735 | 9087 | 10228 | 1142 |
| 6500-29976F | 6500 | 29976999 | 29977829 | F | TAAGTAC | 10735 | 18869 | 19699 | 831 |
| 6500-30981F | 6500 | 30981230 | 30981940 | F | no | 10762 | 3155 | 4465 | 711 |
| 6500-31888F | 6500 | 31888888 | 31889062 | F | GTATAAT | 10790 | 53554 | 53728 | 175 |
| 6500-32144F | 6500 | 32144419 | 32145123 | F | TTATAAT | 10797 | 25558 | 26210 | 705 |
| 6540-32266F | 6540 | 322669 | 323530 | F | TCACTAC | 10921 | 4012 | 4873 | 862 |
| 6540-46643F | 6540 | 466436 | 467246 | F | TTTAAAG | 10925 | 952 | 1762 | 811 |
| 6540-62798F | 6540 | 627982 | 628716 | F | ATATTGA | 10925 | 162498 | 163232 | 735 |
| 6540-69428F | 6540 | 694288 | 695126 | F | GTTCAGA | 10927 | 19474 | 20312 | 839 |
| 6540-75429F | 6540 | 754292 | 755195 | F | GTAGTAT | 10928 | 46450 | 47353 | 904 |
| 6540-10727F | 6540 | 1072704 | 1073438 | F | CTTATAT | 10941 | 6593 | 7327 | 735 |
| 6540-73206F | 6540 | 7320643 | 7321437 | F | GTGGAAC | 10998 | 72162 | 72956 | 795 |
| 6541-83575F | 6541 | 835755 | 836619 | F | ATTATAT | 11224 | 13291 | 14155 | 865 |
| 6541-10932F | 6541 | 1093209 | 1093801 | F | GTACAGA | 11232 | 3722 | 4314 | 593 |

SI Table 3.2. Continuation. V. Solo-TIR copies (continuation)

| TAG | Scaffold | Start | End | Subfamily | TSD1 | Contig_1 | Start | End | Total_length |
|---|---|---|---|---|---|---|---|---|---|
| 6541-24225F | 6541 | 2422509 | 2423290 | F | GTTCAGG | 11283 | 11503 | 12284 | 782 |
| 6680-23160F | 6680 | 23160719 | 23161539 | F | GTTATAA | 11659 | 24852 | 25672 | 671 |
| 6680-23219F | 6680 | 23219687 | 23220569 | F | CTCTAAC | 11661 | 17802 | 18684 | 883 |
| 6680-23825F | 6680 | 23825194 | 23825885 | F | GCAGAAA | 11672 | 39965 | 40656 | 692 |
| 6680-24145F | 6680 | 24145587 | 24146659 | F | GTACAGA | 11680 | 28664 | 29736 | 1073 |
| 6493-38387X | 6493 | 38387 | 39006 | X | GTAATAT | 9982 | 1 | 620 | 620 |
| 6500-31891X | 6500 | 31891331 | 31891606 | X | no | 10790 | 55997 | 56272 | 276 |
| 6540-72269X | 6540 | 722695 | 723349 | X | ATATGAA | 10928 | 14931 | 15507 | 655 |

SI Table **3.3**. Chromosomal distribution of *Galileo* copies in *D. mojavensis*.

| CAF1_scaffold | Galileo_start | Galileo_end | Galileo_Group | Galileo_Type | Galileo_length | GenBank Scaffold Acc | Scaffold_length_(bp) | Chr_arm |
|---|---|---|---|---|---|---|---|---|
| 6328 | 1650720 | 1653101 | F | 2TIR | 2382 | CH933812.1 | 4453435 | X |
| 6473 | 1008070 | 1010341 | E | Longer_2TIR | 2272 | CH933810.1 | 16943266 | X |
| 6473 | 11762829 | 11764731 | C | 2TIR | 1903 | CH933810.1 | 16943266 | X |
| 6473 | 16293472 | 16295724 | F | 2TIR | 2253 | CH933810.1 | 16943266 | X |
| 6482 | 269026 | 270625 | Chimeric | 2TIR | 1600 | CH933815.1 | 2735782 | X |
| 6482 | 362528 | 364455 | E | Longer_2TIR | 1928 | CH933815.1 | 2735782 | X |
| 6482 | 453934 | 454728 | E | 2TIR | 795 | CH933815.1 | 2735782 | X |
| 6482 | 608936 | 612509 | F | NC_DD | 3574 | CH933815.1 | 2735782 | X |
| 6482 | 614003 | 617184 | D | NC_DD | 3182 | CH933815.1 | 2735782 | X |
| 6482 | 617185 | 621442 | D | NC_DD | 3390 | CH933815.1 | 2735782 | X |
| 6482 | 2192546 | 2194452 | F | 2TIR | 1907 | CH933815.1 | 2735782 | X |
| 6482 | 2526809 | 2527340 | D | SOLO | 532 | CH933815.1 | 2735782 | X |
| 6482 | 2579294 | 2581638 | F | 2TIR | 2345 | CH933815.1 | 2735782 | X |
| 6496 | 15292514 | 15294879 | E | Longer_2TIR | 2366 | CH933808.1 | 26866924 | 5 |
| 6496 | 23195067 | 23197892 | F | NC_DD | 2826 | CH933808.1 | 26866924 | 5 |
| 6496 | 25846816 | 25848819 | F | 2TIR | 2004 | CH933808.1 | 26866924 | 5 |
| 6498 | 950693 | 951185 | X | 2TIR | 493 | CH933813.1 | 3408170 | 6 |
| 6498 | 953555 | 955919 | E | Longer_2TIR | 2365 | CH933813.1 | 3408170 | 6 |
| 6498 | 1730250 | 1731243 | F | SOLO | 994 | CH933813.1 | 3408170 | 6 |
| 6498 | 1818872 | 1820002 | F | 2TIR | 1032 | CH933813.1 | 3408170 | 6 |
| 6498 | 1999631 | 2001782 | E | 2TIR | 697 | CH933813.1 | 3408170 | 6 |
| 6498 | 2177013 | 2181005 | C | Longer_2TIR | 3119 | CH933813.1 | 3408170 | 6 |
| 6498 | 2253149 | 2269701 | F | NC | 4036 | CH933813.1 | 3408170 | 6 |
| 6498 | 2381827 | 2382890 | F | SOLO | 1064 | CH933813.1 | 3408170 | 6 |
| 6498 | 2386095 | 2392524 | D | NC | 5721 | CH933813.1 | 3408170 | 6 |
| 6498 | 2407995 | 2408421 | E | SOLO | 427 | CH933813.1 | 3408170 | 6 |
| 6498 | 2514476 | 2515268 | F | SOLO | 793 | CH933813.1 | 3408170 | 6 |
| 6498 | 2522128 | 2522920 | F | SOLO | 793 | CH933813.1 | 3408170 | 6 |

SI Table 3.3. Chromosomal distribution of *Galileo* copies in *D. mojavensis* (continuation).

| CAF1_scaffold | Galileo_start | Galileo_end | Galileo_Group | Galileo_Type | Galileo_length | GenBank Scaffold Acc | Scaffold_length_(bp) | Chr_arm |
|---|---|---|---|---|---|---|---|---|
| 6498 | 2541172 | 2544793 | X | NC_DD | 2423 | CH933813.1 | 3408170 | 6 |
| 6498 | 2560957 | 2561558 | D | SOLO | 602 | CH933813.1 | 3408170 | 6 |
| 6498 | 2609857 | 2611989 | E | SOLO | 1149 | CH933813.1 | 3408170 | 6 |
| 6498 | 2786970 | 2787863 | F | SOLO | 894 | CH933813.1 | 3408170 | 6 |
| 6498 | 2836237 | 2836700 | E | SOLO | 464 | CH933813.1 | 3408170 | 6 |
| 6498 | 2903343 | 2904985 | X | 2TIR | 1643 | CH933813.1 | 3408170 | 6 |
| 6498 | 2966893 | 2969965 | F | 2TIR | 2137 | CH933813.1 | 3408170 | 6 |
| 6498 | 2993866 | 2995242 | E | 2TIR | 1377 | CH933813.1 | 3408170 | 6 |
| 6498 | 3022490 | 3023060 | F | SOLO | 571 | CH933813.1 | 3408170 | 6 |
| 6498 | 3120041 | 3121220 | E | SOLO | 1180 | CH933813.1 | 3408170 | 6 |
| 6498 | 3281538 | 3283440 | F | 2TIR | 1903 | CH933813.1 | 3408170 | 6 |
| 6500 | 29395284 | 29396934 | X | 2TIR | 1651 | CH933807.1 | 32352404 | 3 |
| 6500 | 29798764 | 29799506 | C | SOLO | 743 | CH933807.1 | 32352404 | 3 |
| 6500 | 29804958 | 29807163 | E | Longer_2TIR | 2206 | CH933807.1 | 32352404 | 3 |
| 6500 | 29864896 | 29868109 | F | Longer_2TIR | 3214 | CH933807.1 | 32352404 | 3 |
| 6500 | 29965273 | 29966306 | F | SOLO | 1034 | CH933807.1 | 32352404 | 3 |
| 6500 | 29967217 | 29968358 | F | SOLO | 1142 | CH933807.1 | 32352404 | 3 |
| 6500 | 29973001 | 29975284 | Chimeric | 2TIR | 2284 | CH933807.1 | 32352404 | 3 |
| 6500 | 29976999 | 29977829 | F | SOLO | 831 | CH933807.1 | 32352404 | 3 |
| 6500 | 30179877 | 30181717 | Chimeric | 2TIR | 1841 | CH933807.1 | 32352404 | 3 |
| 6500 | 30183437 | 30184591 | Chimeric | 2TIR | 1155 | CH933807.1 | 32352404 | 3 |
| 6500 | 30306361 | 30308225 | E | Longer_2TIR | 1865 | CH933807.1 | 32352404 | 3 |
| 6500 | 30329586 | 30332946 | F | 2TIR | 3361 | CH933807.1 | 32352404 | 3 |
| 6500 | 30351497 | 30353286 | X | Longer_2TIR | 1790 | CH933807.1 | 32352404 | 3 |
| 6500 | 30494802 | 30496424 | F | 2TIR | 1344 | CH933807.1 | 32352404 | 3 |
| 6500 | 30590766 | 30591289 | D | SOLO | 524 | CH933807.1 | 32352404 | 3 |

SI Table 3.3. Chromosomal distribution of *Galileo* copies in *D. mojavensis* (continuation).

| CAF1_scaffold | Galileo_start | Galileo_end | Galileo_Group | Galileo_Type | Galileo_length | GenBank Scaffold Acc | Scaffold_length_(bp) | Chr_arm |
|---|---|---|---|---|---|---|---|---|
| 6500 | 30596827 | 30599409 | F | 2TIR | 2583 | CH933807.1 | 32352404 | 3 |
| 6500 | 30684259 | 30686266 | F | 2TIR | 2008 | CH933807.1 | 32352404 | 3 |
| 6500 | 30702790 | 30704984 | E | Longer_2TIR | 2195 | CH933807.1 | 32352404 | 3 |
| 6500 | 30733241 | 30734538 | Chimeric | 2TIR | 1298 | CH933807.1 | 32352404 | 3 |
| 6500 | 30856283 | 30862823 | C | NC | 5989 | CH933807.1 | 32352404 | 3 |
| 6500 | 30873698 | 30875357 | F | 2TIR | 1660 | CH933807.1 | 32352404 | 3 |
| 6500 | 30976506 | 30978415 | F | 2TIR | 1910 | CH933807.1 | 32352404 | 3 |
| 6500 | 30981230 | 30981940 | F | SOLO | 711 | CH933807.1 | 32352404 | 3 |
| 6500 | 31083091 | 31089863 | C | NC_DD | 2469 | CH933807.1 | 32352404 | 3 |
| 6500 | 31107017 | 31109152 | F | 2TIR | 2136 | CH933807.1 | 32352404 | 3 |
| 6500 | 31202553 | 31204509 | E | Longer_2TIR | 1957 | CH933807.1 | 32352404 | 3 |
| 6500 | 31288762 | 31295303 | C | NC | 5836 | CH933807.1 | 32352404 | 3 |
| 6500 | 31339017 | 31339980 | E | SOLO | 964 | CH933807.1 | 32352404 | 3 |
| 6500 | 31360321 | 31362480 | E | 2TIR | 2160 | CH933807.1 | 32352404 | 3 |
| 6500 | 31371001 | 31374119 | C | Longer_2TIR | 3119 | CH933807.1 | 32352404 | 3 |
| 6500 | 31458921 | 31465167 | D | NC | 4130 | CH933807.1 | 32352404 | 3 |
| 6500 | 31499776 | 31500354 | E | SOLO | 579 | CH933807.1 | 32352404 | 3 |
| 6500 | 31506397 | 31509717 | E | Longer_2TIR | 3321 | CH933807.1 | 32352404 | 3 |
| 6500 | 31516211 | 31518161 | E | Longer_2TIR | 1951 | CH933807.1 | 32352404 | 3 |
| 6500 | 31694898 | 31696939 | F | 2TIR | 2042 | CH933807.1 | 32352404 | 3 |
| 6500 | 31817847 | 31818422 | E | SOLO | 576 | CH933807.1 | 32352404 | 3 |
| 6500 | 31884435 | 31886401 | C | 2TIR | 1967 | CH933807.1 | 32352404 | 3 |
| 6500 | 31888888 | 31889062 | F | SOLO | 175 | CH933807.1 | 32352404 | 3 |
| 6500 | 31891331 | 31891606 | X | SOLO | 276 | CH933807.1 | 32352404 | 3 |
| 6500 | 31920296 | 31922494 | E | Longer_2TIR | 2199 | CH933807.1 | 32352404 | 3 |
| 6500 | 31981325 | 31986443 | C | NC_DD | 5119 | CH933807.1 | 32352404 | 3 |
| 6500 | 32144419 | 32145123 | F | SOLO | 705 | CH933807.1 | 32352404 | 3 |
| 6500 | 32268286 | 32271024 | E | Longer_2TIR | 2184 | CH933807.1 | 32352404 | 3 |

SI Table 3.3. Chromosomal distribution of *Galileo* copies in *D. mojavensis* (continuation).

| CAF1_scaffold | Galileo_start | Galileo_end | Galileo_Group | Galileo_Type | Galileo_length | GenBank Scaffold Acc | Scaffold_length_(bp) | Chr_arm |
|---|---|---|---|---|---|---|---|---|
| 6540 | 322669 | 323530 | F | SOLO | 862 | CH933806.1 | 34148556 | 2 |
| 6540 | 414493 | 419539 | X | NC | 5047 | CH933806.1 | 34148556 | 2 |
| 6540 | 466436 | 467246 | F | SOLO | 811 | CH933806.1 | 34148556 | 2 |
| 6540 | 558528 | 561650 | Chimeric | 2TIR | 2443 | CH933806.1 | 34148556 | 2 |
| 6540 | 564326 | 566215 | F | 2TIR | 1889 | CH933806.1 | 34148556 | 2 |
| 6540 | 596833 | 597648 | F | SOLO | 816 | CH933806.1 | 34148556 | 2 |
| 6540 | 627982 | 628716 | F | SOLO | 735 | CH933806.1 | 34148556 | 2 |
| 6540 | 694288 | 695126 | F | SOLO | 839 | CH933806.1 | 34148556 | 2 |
| 6540 | 722695 | 723349 | X | SOLO | 655 | CH933806.1 | 34148556 | 2 |
| 6540 | 750291 | 750804 | E | SOLO | 514 | CH933806.1 | 34148556 | 2 |
| 6540 | 754292 | 755195 | F | SOLO | 904 | CH933806.1 | 34148556 | 2 |
| 6540 | 796704 | 798194 | F | 2TIR | 1491 | CH933806.1 | 34148556 | 2 |
| 6540 | 898138 | 899882 | Chimeric | SOLO | 1285 | CH933806.1 | 34148556 | 2 |
| 6540 | 1006735 | 1007752 | E | SOLO | 1018 | CH933806.1 | 34148556 | 2 |
| 6540 | 1035815 | 1037361 | D | 2TIR | 1547 | CH933806.1 | 34148556 | 2 |
| 6540 | 1072704 | 1073438 | F | SOLO | 735 | CH933806.1 | 34148556 | 2 |
| 6540 | 1165033 | 1168349 | E | Longer_2TIR | 2180 | CH933806.1 | 34148556 | 2 |
| 6540 | 1175880 | 1182997 | D | NC | 5433 | CH933806.1 | 34148556 | 2 |
| 6540 | 1372066 | 1373808 | E | Longer_2TIR | 1743 | CH933806.1 | 34148556 | 2 |
| 6540 | 5750063 | 5751976 | F | 2TIR | 1914 | CH933806.1 | 34148556 | 2 |
| 6540 | 6132112 | 6133394 | C | 2TIR | 1283 | CH933806.1 | 34148556 | 2 |
| 6540 | 7320643 | 7321437 | F | SOLO | 795 | CH933806.1 | 34148556 | 2 |
| 6540 | 13434776 | 13435834 | C | SOLO | 1059 | CH933806.1 | 34148556 | 2 |
| 6540 | 14510521 | 14512886 | E | Longer_2TIR | 2366 | CH933806.1 | 34148556 | 2 |
| 6540 | 31163990 | 31166355 | E | Longer_2TIR | 2366 | CH933806.1 | 34148556 | 2 |
| 6540 | 33286261 | 33288174 | F | 2TIR | 1914 | CH933806.1 | 34148556 | 2 |
| 6541 | 178316 | 180001 | F | 2TIR | 1686 | CH933817.1 | 2543558 | X |

SI Table 3.3. Chromosomal distribution of *Galileo* copies in *D. mojavensis* (continuation).

| CAF1_scaffold | Galileo_start | Galileo_end | Galileo_Group | Galileo_Type | Galileo_length | GenBank Scaffold Acc | Scaffold_length_(bp) | Chr_arm |
|---|---|---|---|---|---|---|---|---|
| 6541 | 835755 | 836619 | F | SOLO | 865 | CH933817.1 | 2543558 | X |
| 6541 | 997100 | 998743 | Chimeric | 2TIR | 1644 | CH933817.1 | 2543558 | X |
| 6541 | 1003587 | 1007838 | F | 2TIR | 1876 | CH933817.1 | 2543558 | X |
| 6541 | 1042036 | 1043771 | E | Longer_2TIR | 1736 | CH933817.1 | 2543558 | X |
| 6541 | 1088506 | 1090501 | E | Longer_2TIR | 1996 | CH933817.1 | 2543558 | X |
| 6541 | 1093209 | 1093801 | F | SOLO | 593 | CH933817.1 | 2543558 | X |
| 6541 | 1141978 | 1149130 | Chimeric | NC | 6239 | CH933817.1 | 2543558 | X |
| 6541 | 1249195 | 1251094 | F | 2TIR | 1900 | CH933817.1 | 2543558 | X |
| 6541 | 1511326 | 1513666 | F | 2TIR | 2341 | CH933817.1 | 2543558 | X |
| 6541 | 1618681 | 1620248 | Chimeric | 2TIR | 1568 | CH933817.1 | 2543558 | X |
| 6541 | 1644296 | 1649755 | D | NC | 5460 | CH933817.1 | 2543558 | X |
| 6541 | 1691275 | 1695391 | Chimeric | 2TIR | 1865 | CH933817.1 | 2543558 | X |
| 6541 | 2014221 | 2016575 | E | 2TIR | 1794 | CH933817.1 | 2543558 | X |
| 6541 | 2421943 | 2422508 | D | SOLO | 885 | CH933817.1 | 2543558 | X |
| 6541 | 2422509 | 2423290 | F | SOLO | 782 | CH933817.1 | 2543558 | X |
| 6680 | 23160719 | 23161539 | F | SOLO | 671 | CH933809.1 | 24764193 | 4 |
| 6680 | 23161869 | 23162539 | E | SOLO | 821 | CH933809.1 | 24764193 | 4 |
| 46680 | 23219687 | 23220569 | F | SOLO | 883 | CH933809.1 | 24764193 | 4 |
| 6680 | 23825194 | 23825885 | F | SOLO | 692 | CH933809.1 | 24764193 | 4 |
| 6680 | 24069812 | 24072777 | D | Longer_2TIR | 2174 | CH933809.1 | 24764193 | 4 |
| 6680 | 24145587 | 24146659 | F | SOLO | 1073 | CH933809.1 | 24764193 | 4 |
| 6680 | 24265741 | 24266577 | C | 2TIR | 837 | CH933809.1 | 24764193 | 4 |
| 6680 | 24283772 | 24285562 | X | Longer_2TIR | 1791 | CH933809.1 | 24764193 | 4 |
| 6680 | 24420206 | 24421090 | F | 2TIR | 885 | CH933809.1 | 24764193 | 4 |
| 6680 | 24422223 | 24425258 | F | 2TIR | 1451 | CH933809.1 | 24764193 | 4 |
| 6680 | 24427759 | 24434511 | Chimeric | 2TIR | 2643 | CH933809.1 | 24764193 | 4 |
| 6680 | 24440484 | 24442275 | Chimeric | 2TIR | 1792 | CH933809.1 | 24764193 | 4 |
| 6680 | 24520907 | 24522561 | X | Longer_2TIR | 1655 | CH933809.1 | 24764193 | 4 |
| 6680 | 24538620 | 24540276 | X | Longer_2TIR | 1657 | CH933809.1 | 24764193 | 4 |

SI Table **3.4**. Intrachromosomal distribution of *Galileo* elements.
Chromosome X

| Scaffold | Scaffold_length | GenBank_acc | Chr_arm | Galileo_start | Galileo_end | Galileo_subfam | Galileo_type | Galileo_length | Chr_region |
|---|---|---|---|---|---|---|---|---|---|
| 6482 | 2735782 | CH933815.1 | X | 269026 | 270625 | Chimeric | 2TIR | 1600 | Central |
| 6482 | 2735782 | CH933815.1 | X | 362528 | 364455 | E | Longer_2TIR | 1928 | Central |
| 6482 | 2735782 | CH933815.1 | X | 453934 | 454728 | E | 2TIR | 795 | Central |
| 6482 | 2735782 | CH933815.1 | X | 608936 | 612509 | F | NC_DD | 3574 | Central |
| 6482 | 2735782 | CH933815.1 | X | 614003 | 617184 | D | NC_DD | 3182 | Central |
| 6482 | 2735782 | CH933815.1 | X | 617185 | 621442 | D | NC_DD | 3390 | Central |
| 6482 | 2735782 | CH933815.1 | X | 2192546 | 2194452 | F | 2TIR | 1907 | Central |
| 6482 | 2735782 | CH933815.1 | X | 2526809 | 2527340 | D | SOLO | 532 | 2 |
| 6482 | 2735782 | CH933815.1 | X | 2579294 | 2581638 | F | 2TIR | 2345 | 2 |

| Scf_6482 | Proportion | Region Start | Region End | *Galileo* Obs | *Galileo* Exp | | |
|---|---|---|---|---|---|---|---|
| Region 1 | 10.00% | 1 | 273578 | 0 | 0.9 | Chi square test | |
| Central region | 80.00% | 273579 | 2462203 | 7 | 7.2 | **P-val=** | **0.3246524674** |
| Region 2 | 10.00% | 2462204 | 2735782 | 2 | 0.9 | | |
| | | | | 9 | 9 | | |

SI Table 3.4. Intrachromosomal distribution of *Galileo* elements (continuation).

| Scaffold | GenBank_acc | Chr_arm | Scaffold_length | Galileo_start | Galileo_end | Galileo_subfam | Galileo_type | Galileo_length | Region |
|---|---|---|---|---|---|---|---|---|---|
| 6541 | CH933817.1 | X | 2543558 | 178316 | 180001 | F | 2TIR | 1686 | 1 |
| 6541 | CH933817.1 | X | 2543558 | 835755 | 836619 | F | SOLO | 865 | Central |
| 6541 | CH933817.1 | X | 2543558 | 997100 | 998743 | Chimeric | 2TIR | 1644 | Central |
| 6541 | CH933817.1 | X | 2543558 | 1003587 | 1007838 | F | 2TIR | 1876 | Central |
| 6541 | CH933817.1 | X | 2543558 | 1042036 | 1043771 | E | Longer_2TIR | 1736 | Central |
| 6541 | CH933817.1 | X | 2543558 | 1088506 | 1090501 | E | Longer_2TIR | 1996 | Central |
| 6541 | CH933817.1 | X | 2543558 | 1093209 | 1093801 | F | SOLO | 593 | Central |
| 6541 | CH933817.1 | X | 2543558 | 1141978 | 1149130 | Chimeric | NC | 6239 | Central |
| 6541 | CH933817.1 | X | 2543558 | 1249195 | 1251094 | F | 2TIR | 1900 | Central |
| 6541 | CH933817.1 | X | 2543558 | 1511326 | 1513666 | F | 2TIR | 2341 | Central |
| 6541 | CH933817.1 | X | 2543558 | 1618681 | 1620248 | Chimeric | 2TIR | 1568 | Central |
| 6541 | CH933817.1 | X | 2543558 | 1644296 | 1649755 | D | NC | 5460 | Central |
| 6541 | CH933817.1 | X | 2543558 | 1691275 | 1695391 | Chimeric | 2TIR | 1865 | Central |
| 6541 | CH933817.1 | X | 2543558 | 2014221 | 2016575 | E | 2TIR | 1794 | Central |
| 6541 | CH933817.1 | X | 2543558 | 2421943 | 2422508 | D | SOLO | 885 | 2 |
| 6541 | CH933817.1 | X | 2543558 | 2422509 | 2423290 | F | SOLO | 782 | 2 |

| Scf_6541 | Proportion | Region Start | Region End | *Galileo* Obs | *Galileo* Exp | | |
|---|---|---|---|---|---|---|---|
| Region 1 | 10.00% | 1 | 254356 | 1 | 1.6 | Chi square test | |
| Central region | 80.00% | 254357 | 2289202 | 13 | 12.8 | **P-val=** | **0.8486889772** |
| Region 2 | 10.00% | 2289203 | 2543558 | 2 | 1.6 | | |
| | | | | 16 | 16 | | |

SI Table 3.4.. Intrachromosomal distribution of *Galileo* elements (continuation).
Chromosome 2

| Scaffold | GenBank_acc | Chr_arm | Scaffold_length | Galileo_start | Galileo_end | Galileo_subfam | Galileo_type | Galileo_length | Region |
|---|---|---|---|---|---|---|---|---|---|
| 6540 | CH933806.1 | 2 | 34148556 | 322669 | 323530 | F | SOLO | 862 | 1 |
| 6540 | CH933806.1 | 2 | 34148556 | 414493 | 419539 | X | NC | 5047 | 1 |
| 6540 | CH933806.1 | 2 | 34148556 | 466436 | 467246 | F | SOLO | 811 | 1 |
| 6540 | CH933806.1 | 2 | 34148556 | 558528 | 561650 | Chimeric | 2TIR | 2443 | 1 |
| 6540 | CH933806.1 | 2 | 34148556 | 564326 | 566215 | F | 2TIR | 1889 | 1 |
| 6540 | CH933806.1 | 2 | 34148556 | 596833 | 597648 | F | SOLO | 816 | 1 |
| 6540 | CH933806.1 | 2 | 34148556 | 627982 | 628716 | F | SOLO | 735 | 1 |
| 6540 | CH933806.1 | 2 | 34148556 | 694288 | 695126 | F | SOLO | 839 | 1 |
| 6540 | CH933806.1 | 2 | 34148556 | 722695 | 723349 | X | SOLO | 655 | 1 |
| 6540 | CH933806.1 | 2 | 34148556 | 750291 | 750804 | E | SOLO | 514 | 1 |
| 6540 | CH933806.1 | 2 | 34148556 | 754292 | 755195 | F | SOLO | 904 | 1 |
| 6540 | CH933806.1 | 2 | 34148556 | 796704 | 798194 | F | 2TIR | 1491 | 1 |
| 6540 | CH933806.1 | 2 | 34148556 | 898138 | 899882 | Chimeric | SOLO | 1285 | 1 |
| 6540 | CH933806.1 | 2 | 34148556 | 1006735 | 1007752 | E | SOLO | 1018 | 1 |
| 6540 | CH933806.1 | 2 | 34148556 | 1035815 | 1037361 | D | 2TIR | 1547 | 1 |
| 6540 | CH933806.1 | 2 | 34148556 | 1072704 | 1073438 | F | SOLO | 735 | 1 |
| 6540 | CH933806.1 | 2 | 34148556 | 1165033 | 1168349 | E | Longer_2TIR | 2180 | 1 |
| 6540 | CH933806.1 | 2 | 34148556 | 1175880 | 1182997 | D | NC | 5433 | 1 |
| 6540 | CH933806.1 | 2 | 34148556 | 1372066 | 1373808 | E | Longer_2TIR | 1743 | 1 |
| 6540 | CH933806.1 | 2 | 34148556 | 5750063 | 5751976 | F | 2TIR | 1914 | Central |
| 6540 | CH933806.1 | 2 | 34148556 | 6132112 | 6133394 | C | 2TIR | 1283 | Central |
| 6540 | CH933806.1 | 2 | 34148556 | 7320643 | 7321437 | F | SOLO | 795 | Central |
| 6540 | CH933806.1 | 2 | 34148556 | 13434776 | 13435834 | C | SOLO | 1059 | Central |
| 6540 | CH933806.1 | 2 | 34148556 | 14510521 | 14512886 | E | Longer_2TIR | 2366 | Central |
| 6540 | CH933806.1 | 2 | 34148556 | 31163990 | 31166355 | E | Longer_2TIR | 2366 | 2 |
| 6540 | CH933806.1 | 2 | 34148556 | 33286261 | 33288174 | F | 2TIR | 1914 | 2 |

SI Table 3.4. Intrachromosomal distribution of *Galileo* elements (continuation).

| Scf_6540 | Proportion | Region Start | Region End | Galileos Obs | Galileos Exp | | |
|---|---|---|---|---|---|---|---|
| | | | | | | | **7.956133869452** |
| Centromeric region 1 | 10.00% | 1 | 3414856 | 19 | 2.6 | Chi square test | **44E-026** |
| Central region | 80.00% | 3414857 | 30733700 | 5 | 20.8 | | |
| Telomeric fraction 2 | 10.00% | 30733701 | 34148556 | 2 | 2.6 | | |
| | | | | 26 | 26 | | |

Chromosome 3

| Scaffold | GenBank_acc | Chr_arm | Scaffold_length | Galileo_start | Galileo_end | Galileo_subfam | Galileo_type | Galileo_length | Region |
|---|---|---|---|---|---|---|---|---|---|
| 6500 | CH933807.1 | 3 | 32352404 | 29395284 | 29396934 | X | 2TIR | 1651 | 1 |
| 6500 | CH933807.1 | 3 | 32352404 | 29798764 | 29799506 | C | SOLO | 743 | 1 |
| 6500 | CH933807.1 | 3 | 32352404 | 29804958 | 29807163 | E | Longer_2TIR | 2206 | 1 |
| 6500 | CH933807.1 | 3 | 32352404 | 29864896 | 29868109 | F | Longer_2TIR | 3214 | 1 |
| 6500 | CH933807.1 | 3 | 32352404 | 29965273 | 29966306 | F | SOLO | 1034 | 1 |
| 6500 | CH933807.1 | 3 | 32352404 | 29967217 | 29968358 | F | SOLO | 1142 | 1 |
| 6500 | CH933807.1 | 3 | 32352404 | 29973001 | 29975284 | Chimeric | 2TIR | 2284 | 1 |
| 6500 | CH933807.1 | 3 | 32352404 | 29976999 | 29977829 | F | SOLO | 831 | 1 |
| 6500 | CH933807.1 | 3 | 32352404 | 30179877 | 30181717 | Chimeric | 2TIR | 1841 | 1 |
| 6500 | CH933807.1 | 3 | 32352404 | 30183437 | 30184591 | Chimeric | 2TIR | 1155 | 1 |
| 6500 | CH933807.1 | 3 | 32352404 | 30306361 | 30308225 | E | Longer_2TIR | 1865 | 1 |
| 6500 | CH933807.1 | 3 | 32352404 | 30329586 | 30332946 | F | 2TIR | 3361 | 1 |
| 6500 | CH933807.1 | 3 | 32352404 | 30351497 | 30353286 | X | Longer_2TIR | 1790 | 1 |
| 6500 | CH933807.1 | 3 | 32352404 | 30494802 | 30496424 | F | 2TIR | 1344 | 1 |
| 6500 | CH933807.1 | 3 | 32352404 | 30590766 | 30591289 | D | SOLO | 524 | 1 |
| 6500 | CH933807.1 | 3 | 32352404 | 30596827 | 30599409 | F | 2TIR | 2583 | 1 |
| 6500 | CH933807.1 | 3 | 32352404 | 30684259 | 30686266 | F | 2TIR | 2008 | 1 |
| 6500 | CH933807.1 | 3 | 32352404 | 30702790 | 30704984 | E | Longer_2TIR | 2195 | 1 |
| 6500 | CH933807.1 | 3 | 32352404 | 30733241 | 30734538 | Chimeric | 2TIR | 1298 | 1 |

SI Table 3.4.. Intrachromosomal distribution of *Galileo* elements (continuation).

| Scaffold | GenBank_acc | Chr_arm | Scaffold_length | Galileo_start | Galileo_end | Galileo_subfam | Galileo_type | Galileo_length | Region |
|---|---|---|---|---|---|---|---|---|---|
| 6500 | CH933807.1 | 3 | 32352404 | 30856283 | 30862823 | C | NC | 5989 | 1 |
| 6500 | CH933807.1 | 3 | 32352404 | 30873698 | 30875357 | F | 2TIR | 1660 | 1 |
| 6500 | CH933807.1 | 3 | 32352404 | 30976506 | 30978415 | F | 2TIR | 1910 | 1 |
| 6500 | CH933807.1 | 3 | 32352404 | 30981230 | 30981940 | F | SOLO | 711 | 1 |
| 6500 | CH933807.1 | 3 | 32352404 | 31083091 | 31089863 | C | NC_DD | 2469 | 1 |
| 6500 | CH933807.1 | 3 | 32352404 | 31107017 | 31109152 | F | 2TIR | 2136 | 1 |
| 6500 | CH933807.1 | 3 | 32352404 | 31202553 | 31204509 | E | Longer_2TIR | 1957 | 1 |
| 6500 | CH933807.1 | 3 | 32352404 | 31288762 | 31295303 | C | NC | 5836 | 1 |
| 6500 | CH933807.1 | 3 | 32352404 | 31339017 | 31339980 | E | SOLO | 964 | 1 |
| 6500 | CH933807.1 | 3 | 32352404 | 31360321 | 31362480 | E | 2TIR | 2160 | 1 |
| 6500 | CH933807.1 | 3 | 32352404 | 31371001 | 31374119 | C | Longer_2TIR | 3119 | 1 |
| 6500 | CH933807.1 | 3 | 32352404 | 31458921 | 31465167 | D | NC | 4130 | 1 |
| 6500 | CH933807.1 | 3 | 32352404 | 31499776 | 31500354 | E | SOLO | 579 | 1 |
| 6500 | CH933807.1 | 3 | 32352404 | 31506397 | 31509717 | E | Longer_2TIR | 3321 | 1 |
| 6500 | CH933807.1 | 3 | 32352404 | 31516211 | 31518161 | E | Longer_2TIR | 1951 | 1 |
| 6500 | CH933807.1 | 3 | 32352404 | 31694898 | 31696939 | F | 2TIR | 2042 | 1 |
| 6500 | CH933807.1 | 3 | 32352404 | 31817847 | 31818422 | E | SOLO | 576 | 1 |
| 6500 | CH933807.1 | 3 | 32352404 | 31884435 | 31886401 | C | 2TIR | 1967 | 1 |
| 6500 | CH933807.1 | 3 | 32352404 | 31888888 | 31889062 | F | SOLO | 175 | 1 |
| 6500 | CH933807.1 | 3 | 32352404 | 31891331 | 31891606 | X | SOLO | 276 | 1 |
| 6500 | CH933807.1 | 3 | 32352404 | 31920296 | 31922494 | E | Longer_2TIR | 2199 | 1 |
| 6500 | CH933807.1 | 3 | 32352404 | 31981325 | 31986443 | C | NC_DD | 5119 | 1 |
| 6500 | CH933807.1 | 3 | 32352404 | 32144419 | 32145123 | F | SOLO | 705 | 1 |
| 6500 | CH933807.1 | 3 | 32352404 | 32268286 | 32271024 | E | Longer_2TIR | 2184 | 1 |

| Scf_6500 | Proportion | Region Start | Region End | Observed | Expected | | |
|---|---|---|---|---|---|---|---|
| Centromeric (1) | 10.00% | 29117164 | 32352404 | 43 | 4.3 | Chi square test | |
| Central | 80.00% | 3235241 | 29117163 | 0 | 34.4 | **P-val=** | **9.20487195758081E-085** |
| Telomeric (2) | 10.00% | 1 | 3235240 | 0 | 4.3 | | |
| | | | | 43 | 43 | | |

SI Table 3.4.. Intrachromosomal distribution of *Galileo* elements (continuation).
Chromosome 4

| Scaffold | GenBank_acc | Chr_arm | Scaffold_length | Galileo_start | Galileo_end | Galileo_subfam | Galileo_type | Galileo_length | Region |
|---|---|---|---|---|---|---|---|---|---|
| 6680 | CH933809.1 | 4 | 24764193 | 23160719 | 23161539 | F | SOLO | 671 | 1 |
| 6680 | CH933809.1 | 4 | 24764193 | 23161869 | 23162539 | E | SOLO | 821 | 1 |
| 6680 | CH933809.1 | 4 | 24764193 | 23219687 | 23220569 | F | SOLO | 883 | 1 |
| 6680 | CH933809.1 | 4 | 24764193 | 23825194 | 23825885 | F | SOLO | 692 | 1 |
| 6680 | CH933809.1 | 4 | 24764193 | 24069812 | 24072777 | D | Longer_2TIR | 2174 | 1 |
| 6680 | CH933809.1 | 4 | 24764193 | 24145587 | 24146659 | F | SOLO | 1073 | 1 |
| 6680 | CH933809.1 | 4 | 24764193 | 24265741 | 24266577 | C | 2TIR | 837 | 1 |
| 6680 | CH933809.1 | 4 | 24764193 | 24283772 | 24285562 | X | Longer_2TIR | 1791 | 1 |
| 6680 | CH933809.1 | 4 | 24764193 | 24420206 | 24421090 | F | 2TIR | 885 | 1 |
| 6680 | CH933809.1 | 4 | 24764193 | 24422223 | 24425258 | F | 2TIR | 1451 | 1 |
| 6680 | CH933809.1 | 4 | 24764193 | 24427759 | 24434511 | Chimeric | 2TIR | 2643 | 1 |
| 6680 | CH933809.1 | 4 | 24764193 | 24440484 | 24442275 | Chimeric | 2TIR | 1792 | 1 |
| 6680 | CH933809.1 | 4 | 24764193 | 24520907 | 24522561 | X | Longer_2TIR | 1655 | 1 |
| 6680 | CH933809.1 | 4 | 24764193 | 24538620 | 24540276 | X | Longer_2TIR | 1657 | 1 |

| Scf_6680 | Proportion | Region Start | Region End | *Galileo* Obs | *Galileo* Exp | | |
|---|---|---|---|---|---|---|---|
| Telomeric (3) | 10.00% | 0 | 2476419 | 0 | 1.4 | Chi square test | |
| Central (2) | 80.00% | 2476420 | 22287773 | 0 | 11.2 | **P-value=** | **4.35961000006307E-028** |
| Centromeric (1) | 10.00% | 22287774 | 24764193 | 14 | 1.4 | | |
| | | | | 14 | 14 | | |

SI Table 3.4. Intrachromosomal distribution of *Galileo* elements (continuation).
Chromosome 5

| Scaffold | GenBank_acc | Chr_arm | Scaffold_length | Galileo_start | Galileo_end | Galileo_subfam | Galileo_type | Galileo_length | Region |
|---|---|---|---|---|---|---|---|---|---|
| 6496 | CH933808.1 | 5 | 26866924 | 15292514 | 15294879 | E | Longer_2TIR | 2366 | Central |
| 6496 | CH933808.1 | 5 | 26866924 | 23195067 | 23197892 | F | NC_DD | 2826 | Central |
| 6496 | CH933808.1 | 5 | 26866924 | 25846816 | 25848819 | F | 2TIR | 2004 | 3 |

| Region | Proportion | Region Start | Region End |
|---|---|---|---|
| Telomeric (1) | 10.00% | 0 | 2686692 |
| Central | 80.00% | 2686693 | 24180231 |
| Centromeric (2) | 10.00% | 24180232 | 26866924 |

No enough copies for a Chi square test

Chromosome 6

| Scaffold | GenBank_acc | Chr_arm | Scaffold_length | Galileo_start | Galileo_end | Galileo_subfam | Galileo_type | Galileo_length | Region |
|---|---|---|---|---|---|---|---|---|---|
| 6498 | CH933813.1 | 6 | 3408170 | 950693 | 951185 | X | 2TIR | 493 | Central |
| 6498 | CH933813.1 | 6 | 3408170 | 953555 | 955919 | E | Longer_2TIR | 2365 | Central |
| 6498 | CH933813.1 | 6 | 3408170 | 1730250 | 1731243 | F | SOLO | 994 | Central |
| 6498 | CH933813.1 | 6 | 3408170 | 1818872 | 1820002 | F | 2TIR | 1032 | Central |
| 6498 | CH933813.1 | 6 | 3408170 | 1999631 | 2001782 | E | 2TIR | 697 | Central |
| 6498 | CH933813.1 | 6 | 3408170 | 2177013 | 2181005 | C | Longer_2TIR | 3119 | Central |
| 6498 | CH933813.1 | 6 | 3408170 | 2253149 | 2269701 | F | NC | 4036 | Central |
| 6498 | CH933813.1 | 6 | 3408170 | 2381827 | 2382890 | F | SOLO | 1064 | Central |
| 6498 | CH933813.1 | 6 | 3408170 | 2386095 | 2392524 | D | NC | 5721 | Central |
| 6498 | CH933813.1 | 6 | 3408170 | 2407995 | 2408421 | E | SOLO | 427 | Central |
| 6498 | CH933813.1 | 6 | 3408170 | 2514476 | 2515268 | F | SOLO | 793 | Central |
| 6498 | CH933813.1 | 6 | 3408170 | 2522128 | 2522920 | F | SOLO | 793 | Central |
| 6498 | CH933813.1 | 6 | 3408170 | 2541172 | 2544793 | X | NC_DD | 2423 | Central |
| 6498 | CH933813.1 | 6 | 3408170 | 2560957 | 2561558 | D | SOLO | 602 | Central |
| 6498 | CH933813.1 | 6 | 3408170 | 2609857 | 2611989 | E | SOLO | 1149 | Central |

SI Table 3.4. Intrachromosomal distribution of *Galileo* elements (continuation).

| Scaffold | GenBank_acc | Chr_arm | Scaffold_length | Galileo_start | Galileo_end | Galileo_subfam | Galileo_type | Galileo_length | Region |
|---|---|---|---|---|---|---|---|---|---|
| 6498 | CH933813.1 | 6 | 3408170 | 2786970 | 2787863 | F | SOLO | 894 | Central |
| 6498 | CH933813.1 | 6 | 3408170 | 2836237 | 2836700 | E | SOLO | 464 | Central |
| 6498 | CH933813.1 | 6 | 3408170 | 2903343 | 2904985 | X | 2TIR | 1643 | Central |
| 6498 | CH933813.1 | 6 | 3408170 | 2966893 | 2969965 | F | 2TIR | 2137 | Central |
| 6498 | CH933813.1 | 6 | 3408170 | 2993866 | 2995242 | E | 2TIR | 1377 | Central |
| 6498 | CH933813.1 | 6 | 3408170 | 3022490 | 3023060 | F | SOLO | 571 | Central |
| 6498 | CH933813.1 | 6 | 3408170 | 3120041 | 3121220 | E | SOLO | 1180 | 2 |
| 6498 | CH933813.1 | 6 | 3408170 | 3281538 | 3283440 | F | 2TIR | 1903 | 2 |

| Region | Proportion | Region Start | Region End | *Galileo* Obs | *Galileo* Exp | | |
|---|---|---|---|---|---|---|---|
| Region 1 | 10.00% | 0 | 340817 | 0 | 2.3 | Chi square test | |
| Central Region | 80.00% | 340818 | 3067352 | 21 | 18.4 | P-val= | 0.2583962888 |
| Region 2 | 10.00% | 3067353 | 3408170 | 2 | 2.3 | | |
| | | | | 23 | 23 | | |

SI Table **3.5**. Nearest genes to *Galileo* copies.

| Scaffold | Start | Type | Group | Gene | *Galileo* position | Distance | D. melanogaster orthologous gene | Molecular function | Biological process |
|---|---|---|---|---|---|---|---|---|---|
| 6540 | 5750063 | 2TIR | F | Dmoj-GI24072 | downstream | 29 | Unknown | Unknown | Unknown |
| 6540 | 31163990 | Longer_2TIR | E | Dmoj-GI10679 | upstream | 69 | Unknown | Unknown | Unknown |
| 6500 | 29395284 | 2TIR | X | Dmoj-GI18249 | downstream | 131 | Dmel\CG2614 | Methyl-transferase (InterProScan) | Metabolic process |
| 6540 | 14510521 | Longer_2TIR | E | Dmoj-tRNA:GI25221 | upstream | 144 | tRNA | tRNA | tRNA |
| | | | | Dmoj-tRNA:GI25222 | downstream | 153 | tRNA | tRNA | tRNA |
| 6540 | 6132112 | 2TIR | C | Dmoj-GI23502 | downstream | 147 | Unknown | Unknown | Unknown |
| | | | | Dmoj-GI23503 | upstream | 219 | CSN5 | NEDD8 activating enzyme activity | Biological regulation; neuron differentiation; system development; multicellular organism reproduction; macromolecule modification; cellular component organization or biogenesis; localization; gamete generation; anterior/posterior axis specification; sensory organ development; dorsal/ventral axis specification |
| 6496 | 23195067 | NC_DD | F | Dmoj-GI18468 | upstream | 148 | Dmel\CG7922 | Helicase activity | ATP-dependent RNA helicase activity |
| | | | | Dmoj-GI18348 | upstream | 152 | Dmel\CG9890 | Zinc ion binding | Unknown |
| 6496 | 25846816 | 2TIR | F | Dmoj-GI21310 | downstream | 371 | Nop60B (Nucleolar protein at 60B ) | Pseudouridylate synthase activity | Wing disc development; ribosome biogenesis; germ cell development; rRNA processing; pseudouridine synthesis |

SI Table 3.5. Nearest genes to *Galileo* copies (continuation).

| Scaffold | Start | Type | Group | Gene | *Galileo* position | Distance | D. melanogaster orthologous gene | Molecular function | Biological process |
|---|---|---|---|---|---|---|---|---|---|
| 6540 | 33286261 | 2TIR | F | Dmoj-GI21981 | upstream | 165 | Orc2 | DNA-binding | Mitotic chromosome condensation; DNA-dependent DNA replication initiation; cell proliferation; eggshell chorion gene amplification; mitotic spindle organization; DNA replication; chromosome condensation |
| | | | | Dmoj-GI10788 | downstream | 371 | T-cp1 | Unfolded protein binding | Mitotic spindle organization; phagocytosis, engulfment |
| 6680 | 23825194 | SOLO | F | Dmoj-GI13965 | downstream | 209 | GNBP1 (Gram-negative bacteria binding protein 1) | Protein binding | Peptidoglycan binding; immune response; peptidoglycan catabolic process; defense response to Gram-positive bacterium |
| 6498 | 2786970 | SOLO | F | Dmoj-GI14139 | upstream | 445 | Nmdyn-D6 | Nucleoside diphosphate kinase activity | Nucleoside diphosphate phosphorylation; GTP biosynthetic process; CTP biosynthetic process; UTP biosynthetic process |
| 6328 | 1650720 | 2TIR | F | Dmoj-GI16179 | upstream | 463 | Dmel\CG4332 | Unknown | Unknown |
| 6540 | 1072704 | SOLO | F | Dmoj-GI23814 | upstream | 486 | Dmel\CG16899 // FoxP | Sequence-specific DNA binding transcription factor activity | Regulation of transcription, DNA-dependent |
| 6540 | 31163990 | Longer_2TIR | E | Dmoj-GI10680 | downstream | 426 | His4r | DNA binding; | Chromatin assembly or disassembly |

SI Table **3.6**. Intronic *Galileo* copies

| Scaffold | Start | End | Type | Group | Gene | *FB* gene name | Intron length | D. melanogaster orthologous |
|---|---|---|---|---|---|---|---|---|
| 6473 | 11762829 | 11764731 | 2TIR | C | Dmoj\GI15819 | *Fb*gn0138568 | 5920 | Dmel\CG9572 |
| 6482 | 269026 | 270625 | 2TIR | Chimeric | Dmoj\GI14384 | *FB*gn0137136 | 119822 | Unknown |
| 6482 | 362528 | 364455 | Longer_2TIR | E | Dmoj\GI14384 | *FB*gn0137136 | 119822 | Unknown |
| 6482 | 614003 | 617184 | NC_DD | D | Dmoj\GI14397 | *FB*gn0137149 | 16289 | Unknown |
| 6482 | 617185 | 621442 | NC_DD | D | Dmoj\GI14397 | *FB*gn0137149 | 16289 | Unknown |
| 6482 | 2579294 | 2581638 | 2TIR | F | Dmoj\GI14475 | *FB*gn0137227 | 48340 | Dmel\S6kII |
| 6498 | 2407995 | 2408421 | SOLO | E | Dmoj\GI14130 | *FB*gn0136884 | 1478 | Dmel\C12.2 |
| 6498 | 2903343 | 2904985 | 2TIR | X | Dmoj\GI14010 | *FB*gn0136764 | 55397 | Unknown |
| 6498 | 2993866 | 2995242 | 2TIR | E | Dmoj\GI14008 | *FB*gn0136762 | 172415 | Dmel\CG32627//NnaD |
| 6498 | 3022490 | 3023060 | SOLO | F | Dmoj\GI14008 | *FB*gn0136762 | 172415 | Dmel\CG32627//NnaD |
| 6498 | 3120041 | 3121220 | SOLO | E | Dmoj\GI14008 | *FB*gn0136762 | 172415 | Dmel\CG32627//NnaD |
| 6500 | 30733241 | 30734538 | 2TIR | Chimeric | Dmoj\GI18277 | *FB*gn0141016 | 65803 | Unknown |
| 6500 | 31339017 | 31339980 | SOLO | E | Dmoj\GI18740 | *FB*gn0141479 | 15508 | Dmel\CG5708 |
| 6500 | 31884435 | 31886401 | 2TIR | C | Dmoj\GI18594 | *FB*gn0141333 | 9452 | Dmel\Cdk5alpha |
| 6500 | 31888888 | 31889062 | SOLO | F | Dmoj\GI18594 | *FB*gn0141333 | 9452 | Dmel\Cdk5alpha |
| 6500 | 31891331 | 31891606 | SOLO | X | Dmoj\GI18594 | *FB*gn0141333 | 9452 | Dmel\Cdk5alpha |
| 6540 | 694288 | 695126 | SOLO | F | Dmoj\GI23792 | *FB*gn0146517 | 69549 | Unknown |
| 6540 | 722695 | 723349 | SOLO | X | Dmoj\GI23792 | *FB*gn0146517 | 69549 | Unknown |
| 6541 | 835755 | 836619 | SOLO | F | Dmoj\GI14178 | *FB*gn0136931 | 11198 | Dmel\Stim |
| 6541 | 1042036 | 1043771 | Longer_2TIR | E | Dmoj\GI14176 | *FB*gn0136929 | 47317 | Dmel\CG8578 |
| 6541 | 1249195 | 1251094 | 2TIR | F | Dmoj\GI14213 | *Fb*gn0136966 | 8704 | Unknown |
| 6541 | 1511326 | 1513666 | 2TIR | F | Dmoj\GI14170 | *FB*gn0136923 | 16452 | Dmel\Ranbp16 |
| 6680 | 24283772 | 24285562 | Longer_2TIR | X | Dmoj\GI11297 | *FB*gn0134058 | 37777 | Dmel\Pka-C3 |

# V.- DISCUSSION

# 1.- *Galileo* and the *P-element* superfamily of transposons

*Galileo* was discovered by our research group in *D. buzzatii* (Cáceres et al. 1999, 2001).The first *Galileo* sequences did not harbour any coding region neither presented any significant identity to any know TE. Thus, *Galileo* was tentatively classified as a class II *Foldback*-like tranposon, due to its structure, which was mainly composed by long internally repetitive TIR (Cáceres et al. 2001; Casals et al. 2005). In the present thesis, the putatively complete copy of *Galileo* with transposase-coding segment was isolated from *D. buzzatii.* In addition, similar nearly-complete elements were detected in 6 of the 12 sequenced *Drosophila* genomes. These observations provided valuable information for a new classification of the transposon. The transposase analysis showed significant identity to the *P-element* and *1360* transposases along with the same functional protein domains. This fact allowed a functional classification of *Galileo* in the *P-element* superfamily of DNA transposons (Class II, subclass I, TIR elements order, Wicker et al. 2007) which predicts a similar transposition reaction. Conceivably, all the *P-element* superfamily members transpose through a cut-and-paste reaction, where transposon staggered ends are generated after the transposon excision and TSD appear after the transposon insertion.

In this sense, the TSD present different lengths among the *P-element* superfamily members. The *P-element* generates 8-bp palindromic TSD, whereas *Galileo* and *1360* present palindromic TSD of 7-bp. Although the length of the TSD can be used as a diagnostic trait for TE classification (Wicker et al. 2007), there is variability in its length within several transposon superfamilies, such as *MuDR*, *CACTA*, *Merlin*, *Banshee* (reviewed in Feschotte & Pritham 2007). Likewise, TIR length is also a variable trait within different transposon superfamilies, such as, *MuDR*, *Tc1/mariner*, *PIF-harbinger*, (Feschotte & Pritham 2007; Wicker et al. 2007). Despite the length differences, it is noteworthy that TIR and TSD ends of *P-element*, *Galileo* and *1360* start with CA sequence. Since the transposase binding site is not located at the very end in *Galileo* and *P-element,* the reason of this conservation could be the need of this sequence for the endonuclease reaction of the transposon excision.

| Element | Total length | TIR | Transposase coding segment | Introns | Protein residues | TSD |
|---|---|---|---|---|---|---|
| P-element | 2907 | 31 | 2256 | 3 | 751 | 8 |
| 1360 | 3614 | 31 | 2564 | no | 863 | 7 |
| Galileo | 5407 | 1229 | 2739 | no | 912 | 7 |

Table **1.1**: Comparison of different features of *P-element*, *1360* and *Galileo*. Both *P-element* and *1360* are from *D. melanogaster* and *Galileo* corresponds to the synthetic copy from *D. buzzatii* (Marzo et al. 2008). *P-element* accession number: K06779; *1360* accession number: AE014135 (*D. melanogaster* dot chromosome, coordinates 809591-813204).

Regarding the transposase of this superfamily, *Galileo* and *1360* putative proteins harbour the same domains present in the *P-element* transposase. From our analysis, the *Galileo* THAP domain is longer than the other THAP domains (such as *P-element* or THAP1, see Results-Chapter two) and presents a longer N-terminal region as well. This longer THAP domain sequence could be related to the longer binding site of *Dbuz\Galileo*G. Despite its increased length, in accordance to other traits of the transposon, the *Galileo* binding site sequence conserves the proposed consensus nucleotides (Campagne et al. 2010; Sabogal et al. 2010). Thus, we can conclude that the THAP domain of *Galileo* presents significant amino acid identity with other THAP domains and there is also similarity in the recognised nucleotide sequence.

After the THAP domain, there is a coiled coil region where the transposase interacts with other transposase monomers for assembling a transposase multimer. This multimer is a tetramer in the *P-element* (Tang et al. 2007). Presumably, *Galileo* would interact in the same way, although a different number of units in the multimer could be expected, similarly to other superfamilies of transposons, such as *Tc1/Mariner*, where *Mos1* acts as a dimer and *Hermes* as an hexamer (Hickman et al. 2005; Richardson et al. 2006). Since we have only predicted these regions using computational tools, further experimental analysis with the purified transposase would be very interesting.

The next domain that appears in the *P-element* transposase is the GTP binding domain. The GTP acts as an allosteric co-factor and it is not hydrolysed during the reaction (Rio 2002; Tang et al. 2005). Recently, this domain has been delimited by Sabogal & Rio (2010) after isolating it and checking that the GTP binding activity remained. These residues can be located in the *Galileo* transposase when aligned with the *P-element* transposase. In the *P-element* the GTP binding domain is located in residues 275 to 409, and in *Duz\Galileo*G, in residues 403 to 519. This region presents

27% aminoacid identity when the two transposases are aligned (21.3% identity fot the entire protein). Thus, *Galileo* seems to harbour this domain as well. Experimental evidences would be needed to corroborate the involvement of GTP in *Galileo* transposition reaction and to conclude that GTP would be an important cofactor in the *Galileo* transposition like in the *P-element* transposition.

The last domain in the transposase of the *P-element* superfamily is the catalytic domain, which is characterised by a high proportion of acidic residues and performs cuts in the DNA through an endonuclease reaction (Rio 2002). The catalytic domain of almost all DNA transposons shared the DDE signature with integrases of retroelements, however, the *P-element* did not seem to present it (Hickman et al. 2010). Recently, Yuan and Wessler (2011) have studied systematically a broad sample of transposases of different superfamilies with the aim of uncovering conserved residues not detected before. This way, they have found the DDE motif in the *P-element* superfamily among other superfamilies. The residues proposed by Yuan & Wessler (2011) are not in agreement with those proposed by Rio (2002). In this this work, the catalytic domain of the *Galileo* transposase was found and the key catalytic residues identified (Results – Chapter 1) based on Rio 2002. However, Yuan and Wessler (2011) suggested other key catalytic residues in *Dbuz**Galileo*G transposase, which are D337, D426 and E651. There is only one residue in common with those proposed by Rio (2002), E651. Since the proposed residues are highly conserved among the superfamily transposases, including *Galileo* from different species, it would be very interesting to corroborate experimentally its key role in the transposition reaction along with the $Mg^{++}$ conjugation.

The catalytic domain cuts the transposon at the very end of the TIR, thus, the conservation found in this region must be very important for the proper cut of the transposon. This fact could be the reason why the most conserved region of the different *Galileo* subfamilies is the end of the TIR, especially the nine terminal nucleotides: CACTACCAA (CACT**G**CCAA in C, D, E and X *D. mojavensis* subfamilies).
However, when the different families of the *P-element* superfamily are compared, this conservation is only found in the first two residues of the TIR (CA). Although there are few residues conserved, they might be a trait of a common catalytic domain. Maybe,

the fact that other cut-and-paste transposon TIR start with CA (such as some hAT, *CACTA* or *transib*, Feschotte & Pritham 2007; Yuan & Wessler 2011) is another trait of the shared DDE domain (Hickman et al. 2010; Yuan & Wessler 2011).

Since these three elements, *Galileo*, *1360* and *P-element*, probably share a common ancestor, we could hypothesise which of them could be the most similar to the ancestor of the group. Since the three main members of the *P-element* superfamily are contained in *Drosophila* genus species, the species distribution of these elements would shed some light on the evolutionary relationships among *P-element*, *1360* and *Galileo,* at least in this host genus. The *P-element* does not exist in the *Drosophila* subgenus but *Galileo* and *1360* have been found in the two main subgenera of the *Drosophila* genus. This could be indicating a more ancient origin of *Galileo* and *1360* in the whole genus, which would be in agreement with the lack of complete functional copies found so far. However, since more than 2000 species make up the *Drosophila* genus, the study of more species could uncover very different landscapes.

To sum up, *Galileo* classification is strong and well-supported. The variation in TSD and TIR length does not represent any classification conflict. From our experience, we corroborate that the most powerful criterion for transposon classification is the transposase similarity, which is where the transposition mechanism reside.

## 2.- Long TIR and transposon evolution

Since transposons do not present any selective constraint for the host, they evolve neutrally, with the only requirement of keeping the transposase affinity. Furthermore, since the cell would be repressing the TE activity, the mobile elements would be more able to avoid the cell repression if they are freer to change. However, there is a region with some constraint, the coding sequence. Thus, the higher conservation found in the transposase region, where homology is detected, is in agreement with the transposon selective constraint that would keep it active in the genome. Thanks to this conservation, it is possible to relate divergent transposons in superfamilies, such as the case of *Galileo, P-element* and *1360* (Feschotte & Pritham 2007; Jurka et al. 2007; Wicker et al. 2007).

Excision of cut-and-paste transposons generates double-strand breaks which have to be repaired by the cell machinery. This repair is one of the mechanisms that cut-and-paste transposon use for their proliferation, along with the coupling of transpositional activity to S phase of the cell cycle (Craig et al. 2002; Feschotte & Pritham 2007). On the one hand, as in the case of *P-element*, the staggered ends can join through non-homologous end joining (NHEJ) and a footprint of the transposon would remain at the donor site (Engels et al. 1990). On the other hand, this double strand break can be fixed through a gap repair process using the sister chromatid (G2 cellular stage) or the homologous chromosome (G1 cellular stage) as template through a synthesis-dependent strand annealing (SDSA, Formosa & Alberts 1986). This way, transposon sequence could be both, restored at the donor site or completely erased, depending on the content of the template sequence. (Engels et al. 1990; Rio 2002). Furthermore, besides the sister chromatid or the homologous chromosome, any copy of the transposon could be used as template as well in the gap repair process (Hastings 1988; Gloor et al. 1991). The interruption of the SDSA process would cause a deleted copy (Engels et al. 1990; Gloor et al. 1991; Plasterk 1991; Hsia & Schnable 1996; Dray & Gloor 1997; Rubin & Levy 1997). This way, transposon copies get shorter and there is no selective constraint that would prevent it. Moreover, shorter copies can exhibit a higher transposition rate (as long as the sequences needed for transposase binding and cutting are kept in the copy) than the complete ones and they could outnumber the longest ones (Yang et al. 2009; Atkinson & Chalmers 2010).

The spreading of the incomplete copies would have two effects: on the one hand, the insertion of short copies would have a lower impact in the new genomic location than the longer ones. These insertions would be less harmful for the host and these copies would have advantage over the longer ones, favouring again the spreading of shorter copies. On the other hand, the more transposase target sequences which no transposase production, the less transposition rate, due to the lack of all the required transposase monomers in a given copy at a given time. This is a titration effect which down-regulates the transposition rate and it would be another reason for the short copies be less deleterious than the longest ones. Nevertheless, all these mechanisms seem to be a death sentence for the transposon. This fate, however, could be overcome by the arrival of new TEs through horizontal transfer or by reactivation of formerly inactive

copies (Kidwell 1992; Silva et al. 2004; Sánchez-Gracia et al. 2005; Loreto et al. 2008). We would like to propose that cut-and-paste transposon reactivation could be enhanced by long TIR.

Long-TIR elements have arisen in several transposon superfamilies besides the *P-element* superfamily (Feschotte & Pritham 2007). For example, relatively long TIR elements have been reported in the *Tc1*/*mariner* superfamily as well, such as *Sleeping Beauty* (225 bp), *Tc3* (462 bp) and *Minos* (245 bp) (Collins et al. 1989; Franz & Savakis 1991; Ivics et al. 1997). Another example is the *Phantom* transposon, which has recently been classified as a member of the *Mutator* superfamily (Marquez & Pritham 2010). The TIR of *Phantom* are longer than other related families and present different structures, from simple long TIR to long internally repetitive TIR which resemble the *Foldback* structure . Since the TIR seems a dynamic trait in transposons, it is not a reliable character for classification (Marzo et al. 2008; Marquez & Pritham 2010).

Long TIR could have a negative effect for transposons, because the more distance between the two TIR, the less efficiency in transposition reaction (Atkinson & Chalmers 2010). Furthermore, DNA secondary structures appear with repetitive sequences rendering more chances of DNA breaks during replication. However, since the long TIR appear in different superfamilies they may entail some benefit for the transposon, although they could be a shared trait only by chance. Maybe, the long TIR expands a region without disrupting the promoter sequences and the CDS of the transposon. This way, new binding sites or other transposition enhancing sequences could be located in a longer TIR. Direct repeats, which correspond to binding sites, have been found in different transposons, such as *Sleeping Beauty*, *Bari*, *Herves* (Cui et al. 2002; Moschetti et al. 2008; Kahlon et al. 2011) and in *Galileo* we have strong evidences that its direct repeats would be binding sites as well. The existence of several binding sites in each TIR or transposon end could be useful for a more efficient recruitment of the transposition machinery, where the different binding sites could be driving the transposition proteins to the transposon ends.

Another positive effect of long TIR could be the fact that longer TIR are more prone to recombine and suffer gene conversion. Although it could be a drawback at first

sight, because ectopic recombination, along with deletion, are the main forces to prevent TE spreading (Petrov 2002; Petrov et al. 2003, 2010), gene conversion could favour, for example, the formation of highly identical TIR. Although in some transposons an asymmetry in the binding sites is needed for the transposition reaction (*P-element* and *Herves* for example Rio 2002; Kahlon et al. 2011), maybe other groups, such as long TIR elements (*Galileo*, *Sleeping Beauty*, *Phantom*) transpose better with highly identical and symmetrical binding sites. It would be very interesting to test how identity between the two long TIRs of a transposon affects the transposition reaction.

The possibility that TIRs could behave similarly to segmental duplications provides the transposon with a faster change rate which could result in new sequences that could scape the titration down-regulation and start new transposition bursts. This phenomenon could be considered transposon reactivation, being more useful for the transposon survival compared to punctual mutations, which would take very long to generate new transposon subfamilies or variants. In this sense, conversion and recombination have been found intimately related with transposons in different organisms, such as *Wolbachia* endosymbiont (Cordaux 2009; Ling & Cordaux 2010), and other procaryotes (Redder & Garrett 2006; Beare et al. 2009), yeast (Roeder 1983), and metazoans, such as humans (Schwartz et al. 1998; Lee et al. 2008) or *D. buzzatii,* where inversions have been generated through TE ectopic recombination (Cáceres et al. 1999; Casals et al. 2003; Delprat et al. 2009). Furthermore, the *Galileo* TIR length dynamics we have found in *D. mojavensis* could be the result of this process as well (see Chapter 3 of Results). Thus, TEs evolution seems linked to recombination and conversion where transposon long TIR would favour this association. This could be the reason of the convergence of this trait in different superfamilies of cut-and-paste transposons.

# VI.- CONCLUSIONS

The following conclusions can be drawn from this work:

1. *Galileo* is a class II element (DNA transposon) belonging to subclass 1 order TIR and *P-element* superfamily.

2. Putative complete copies in *D. buzzatii* are 5.4-kb long and contain long TIR (1.2 kb), a transposase-coding segment (2.7 kb) and spacing regions.

3. Similarly to the *P-element* transposase, the *Galileo* transposase contains the following domains: THAP DNA binding domain, coiled coil region, GTP binding domain and catalytic domain similarly to the *P-element* transposase.

4. The common traits between *Galileo* and *P-element* are the palindromic structure of the TSD, the beginning of the TIR sequences (17 out of 31 bp including the first two nucleotides CA) and the similarity in the transposase sequences along with equal disposition of the same protein domains in it.

5. The main differences between *Galileo* and *P-element* are: the length of the TSD, where *P-element* present 8-bp and *Galileo*, 7-bp; the TIR length, where *P-element* present 31-bp and *Galileo* from ~500 bp to ~1,2 kb; the length of the putative binding site, where *P-element* presents 10-11 bp binding site and *Galileo* presents 18-bp.

6. *Galileo* is found, besides *D. buzzatii*, in six of the 12 sequenced genomes: *D. mojavensis*, *D. virilis*, *D. willistoni*, *D. ananassae*, *D. pesudoobscura* and *D. persimilis*. This means that *Galileo* is found in the two main subgenera of *Drosophila* genus, *Sophophora* and *Drosophila* and it is likely widespread in the genus.

7. *Galileo* presents different subfamilies within the genomes of *D. mojavensis* (*Galileo*C, *Galileo*D, *Galileo*E, *Galileo*F, *Galileo*X) and *D. virilis* (*Galileo*A and *Galileo*B). Similarly, the *D. buzzatii* elements *Galileo*, *Kepler* and *Newton* can be considered as subfamilies of *Galileo* in this species (*Dbuz**Galileo*G, *Dbuz**Galileo*N, *Dbuz**Galileo*K, prespectively).

8. The transposase phylogeny generated with consensus transposases of the *Galileo* elements found in each genome, presents a topology that differ from the species phylogeny. This incongruence could be due to horizontal transfer,

incomplete lineage sorting or phylogenetic artefacts, such as long branch attraction as a result of the high divergence the sequences analysed.

9. The transposase THAP DNA binding domains of *Dbuz\Galileo*G, *Dmoj\Galileo*C, *Dmoj\Galileo*D and *Dana\Galileo* have been successfully reconstructed and expressed in vitro. They present specific binding activity for *Galileo* TIR sequences.

10. The DNA binding domain of *Dbuz\Galileo*G was isolated and it was located in nucleotides 63-80 of the *Galileo* TIR. This 18-bp sequence shows similarity to the binding sites of other THAP domains, such as those of *P-element* transposase or human THAP1 protein.

11. No *Galileo* transposase activity has been detected in our *in vivo* transposition experiments

12. Within the genome of *D. mojavensis*, *Galileo* presents, besides its nucleotide variability, huge structural variation in its copies. The TIR is the most variable region in length and structure of the element. This structural dynamism may be explained by several mechanisms, including deletion, duplication, recombination and conversion.

13. *D. mojavensis* genome contains five different *Galileo* subfamilies, four of them harbour transposase coding regions (none of them coding for a functional protein) and the fifth presents a putative chimeric origin.

14. The accumulation of linages through time (LTT) in the phylogeny of *D. mojavensis Galileo* elements shows an exponential increase of copies without any trace of evident deceleration or stationary rate. This suggests that the element is still active in *D. mojavensis* genome or has been active until very recently.

# VII.- APPENDIXES

Vol 450 | 8 November 2007 | doi:10.1038/nature06341 nature

ARTICLES

# Evolution of genes and genomes on the *Drosophila* phylogeny

*Drosophila* 12 Genomes Consortium*

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

As one might expect from a genus with species living in deserts, in the tropics, on chains of volcanic islands and, often, commensally with humans, *Drosophila* species vary considerably in their morphology, ecology and behaviour[1]. Species in this genus span a wide range of global distributions: the 12 sequenced species originate from Africa, Asia, the Americas and the Pacific Islands, and also include cosmopolitan species that have colonized the planet (*D. melanogaster* and *D. simulans*) as well as closely related species that live on single islands (*D. sechellia*)[2]. A variety of behavioural strategies is also encompassed by the sequenced species, ranging in feeding habit from generalist, such as *D. ananassae*, to specialist, such as *D. sechellia*, which feeds on the fruit of a single plant species.

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

## Genome assembly, annotation and alignment

**Genome sequencing and assembly.** We used the previously published sequence and updated assemblies for two *Drosophila* species, *D. melanogaster*[3,4] (release 4) and *D. pseudoobscura*[5] (release 2), and generated DNA sequence data for 10 additional *Drosophila* genomes by whole-genome shotgun sequencing[6,7]. These species were chosen to span a wide variety of evolutionary distances, from closely related pairs such as *D. sechellia*/*D. simulans* and *D. persimilis*/*D. pseudoobscura* to the distantly related species of the *Drosophila* and *Sophophora* subgenera. Whereas the time to the most recent common ancestor of the sequenced species may seem small on an evolutionary timescale, the evolutionary divergence spanned by the genus *Drosophila* exceeds that of the entire mammalian radiation when generation time is taken into account, as discussed further in ref. 8. We sequenced seven of the new species (*D. yakuba*, *D. erecta*, *D. ananassae*, *D. willistoni*, *D. virilis*, *D. mojavensis* and *D. grimshawi*) to deep coverage (8.4× to 11.0×) to produce high quality draft sequences. We sequenced two species, *D. sechellia* and *D. persimilis*, to intermediate coverage (4.9× and 4.1×, respectively) under the assumption that the availability of a sister species sequenced to high coverage would obviate the need for deep sequencing without sacrificing draft genome quality. Finally, seven inbred strains of *D. simulans* were sequenced to low coverage (2.9× coverage from $w^{501}$ and ~1× coverage of six other strains) to provide population variation data[9]. Further details of the sequencing strategy can be found in Table 1, Supplementary Table 1 and section 1 in Supplementary Information.

We generated an initial draft assembly for each species using one of three different whole-genome shotgun assembly programs (Table 1). For *D. ananassae*, *D. erecta*, *D. grimshawi*, *D. mojavensis*, *D. virilis* and *D. willistoni*, we also generated secondary assemblies; reconciliation of these with the primary assemblies resulted in a 7–30% decrease in the estimated number of misassembled regions and a 12–23% increase in the N50 contig size[10] (Supplementary Table 2). For *D. yakuba*, we generated 52,000 targeted reads across low-quality regions and gaps to improve the assembly. This doubled the mean contig and scaffold sizes and increased the total fraction of high quality bases (quality score (Q) > 40) from 96.5% to 98.5%. We improved the initial 2.9× *D. simulans* $w^{501}$ whole-genome shotgun assembly by filling assembly gaps with contigs and unplaced reads from the ~1× assemblies of the six other *D. simulans* strains, generating a 'mosaic' assembly (Supplementary Table 3). This integration markedly improved the *D. simulans* assembly: the N50 contig size of the mosaic assembly, for instance, is more than twice that of the initial $w^{501}$ assembly (17 kb versus 7 kb).

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

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

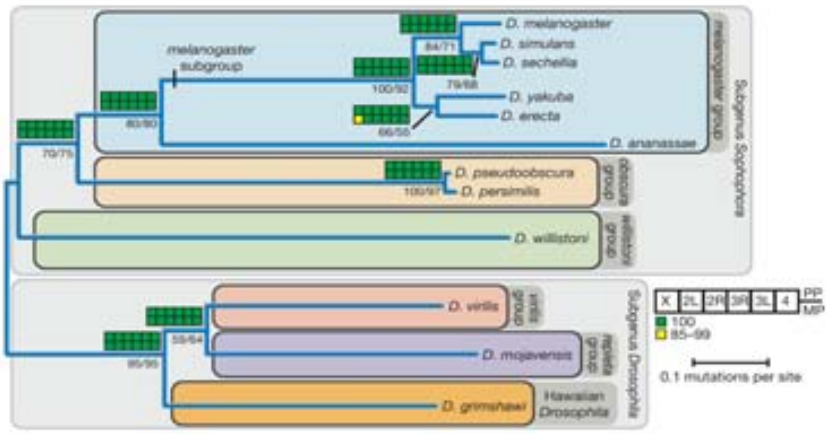



**Figure 1 | Phylogram of the 12 sequenced species of *Drosophila*.** Phylogram derived using pairwise genomic mutation distances and the neighbour-joining method[152,153]. Numbers below nodes indicate the per cent of genes supporting a given relationship, based on evolutionary distances estimated from fourfold-degenerate sites (left of solidus) and second codon positions (right of solidus). Coloured blocks indicate support from bayesian (posterior probability (PP), upper blocks) and maximum parsimony (MP; bootstrap values, lower blocks) analyses of data partitioned by chromosome arm. Branch lengths indicate the number of mutations per site (at fourfold-degenerate sites) using the ordinary least squares method. See ref. 154 for a discussion of the uncertainties in the *D. yakuba*/*D. erecta* clade.

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

The mitochondrial (mt)DNA of *D. melanogaster*, *D. sechellia*, *D. simulans* (siII), *D. mauritiana* (maII) and *D. yakuba* have been previously sequenced[11,12]. For the remaining species (except *D. pseudoobscura*, the DNA from which was prepared from embryonic nuclei), we were able to assemble full mitochondrial genomes, excluding the A+T-rich control region (Supplementary Information section 2)[13]. In addition, the genome sequences of three *Wolbachia* endosymbionts (*Wolbachia wSim*, *Wolbachia wAna* and *Wolbachia wWil*) were assembled from trace archives, in *D. simulans*, *D. ananassae* and *D. willistoni*, respectively[14]. All of the genome sequences described here are available in FlyBase (www.flybase.org) and GenBank (www.ncbi.nlm.nih.gov) (Supplementary Tables 4 and 5).

**Repeat and transposable element annotation.** Repetitive DNA sequences such as transposable elements pose challenges for whole-genome shotgun assembly and annotation. Because the best approach to transposable element discovery and identification is still an active and unresolved research question, we used several repeat libraries and computational strategies to estimate the transposable element/repeat content of the 12 *Drosophila* genome assemblies (Supplementary Information section 3). Previously curated transposable element libraries in *D. melanogaster* provided the starting point for our analysis; to limit the effects of ascertainment bias, we also developed *de novo* repeat libraries using PILER-DF[15,16] and ReAS[17]. We used four transposable element/repeat detection methods (RepeatMasker, BLASTER-TX, RepeatRunner and CompTE) in conjunction with these transposable element libraries to identify repetitive elements in non-*melanogaster* species. We assessed the accuracy of each method by calibration with the estimated 5.5% transposable element content in the *D. melanogaster* genome, which is based on a high-resolution transposable element annotation[18] (Supplementary Fig. 1). On the basis of our results, we suggest a hybrid strategy for new genome sequences, employing translated BLAST with general transposable element libraries and RepeatMasker with species-specific ReAS libraries to estimate the upper and lower bound on transposable element content.

**Protein-coding gene annotation.** We annotated protein-coding sequences in the 11 non-*melanogaster* genomes, using four different *de novo* gene predictors (GeneID[19], SNAP[20], N-SCAN[21] and CONTRAST[22]); three homology-based predictors that transfer annotations from *D. melanogaster* (GeneWise[23], Exonerate[24], GeneMapper[25]); and one predictor that combined *de novo* and homology-based evidence (Gnomon[26]). These gene prediction sets

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

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

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

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

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

* Repeat coverage calculated as the fraction of scaffolds >200 kb covered by repeats, estimated as the midpoint between BLASTER-tx + PILER and RepeatMasker + ReAS (Supplementary Information section 3). †Total genome size estimated as the sum of base pairs in genomic scaffold >200,000 bp. ‡Genome size estimates based on flow cytometry[36].

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

**Quality of gene models.** As the first step in assessing the quality of the GLEAN-R gene models, we used expression data from microarray experiments on adult flies, with arrays custom-designed for *D. simulans*, *D. yakuba*, *D. ananassae*, *D. pseudoobscura*, *D. virilis* and *D. mojavensis*[29] (GEO series GSE6640; Supplementary Information section 4.2). We detected expression significantly above negative controls (false-discovery-rate-corrected Mann–Whitney U (MWU) $P < 0.001$) for 77–93% of assayed GLEAN-R models, representing 50–68% of the total GLEAN-R predictions in each species (Supplementary Table 6). Evolutionarily conserved gene models are much more likely to be expressed than lineage-specific ones (Fig. 2). Although these data cannot confirm the detailed structure of gene models, they do suggest that the majority of GLEAN-R models contain sequence that is part of a poly-adenylated transcript. Approximately 20% of transcription in *D. melanogaster* seems to be unassociated with protein-coding genes[30], and our microarray experiments fail to detect conditionally expressed genes. Thus, transcript abundance cannot conclusively establish the presence or absence of a protein-coding gene. Nonetheless, we believe these expression data increase our confidence in the reliability of the GLEAN-R models, particularly those supported by homology evidence (Fig. 2).

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

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

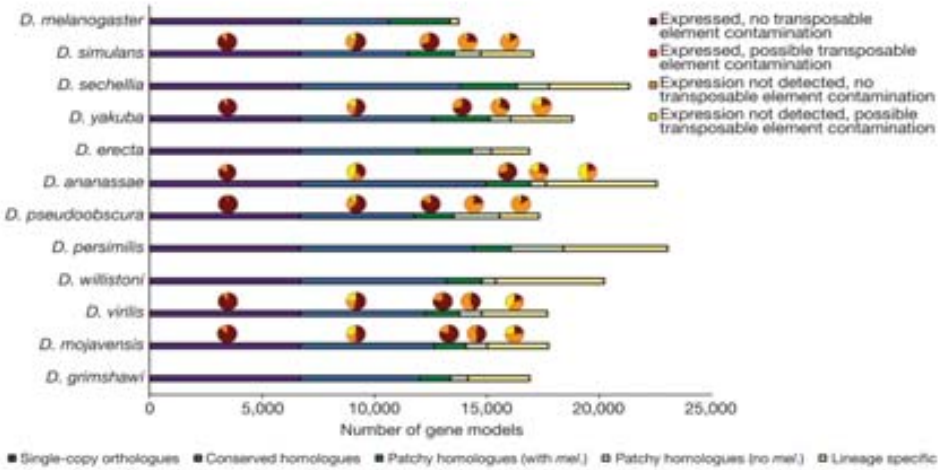



**Figure 2 | Gene models in 12 *Drosophila* genomes.** Number of gene models that fall into one of five homology classes: single-copy orthologues in all species (single-copy orthologues), conserved in all species as orthologues or paralogues (conserved homologues), a *D. melanogaster* homologue, but not found in all species (patchy homologues with *mel.*), conserved in at least two species but without a *D. melanogaster* homologue (patchy homologues, no *mel.*), and found only in a single lineage (lineage specific). For those species with expression data[29], pie charts indicate the fraction of genes in each homology class that fall into one of four evidence classes (see text for details).

extension of the reciprocal BLAST method[31] applicable to multiple species simultaneously (Supplementary Information section 5.1). Because the FRB algorithm does not integrate syntenic information, we also used a second approach based on Synpipe (Supplementary Information section 5.2), a tool for synteny-aided orthology assignment[32]. To generate a reconciled set of homology calls, pairwise Synpipe calls (between each species and *D. melanogaster*) were mapped to GLEAN-R models, filtered to retain only 1:1 relationships, and added to the FRB calls when they did not conflict and were non-redundant. This reconciled FRB + Synpipe set of homology calls forms the basis of our subsequent analyses. There were 8,563 genes with single-copy orthologues in the *melanogaster* group and 6,698 genes with single-copy orthologues in all 12 species; similar numbers of genes were also obtained with an independent approach[33]. Most single-copy orthologues are expressed and are free from potential transposable element contamination, suggesting that the reconciled orthologue set contains robust and high-quality gene models (Fig. 2).

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

**Coding gene alignment and filtering.** Investigating the molecular evolution of orthologous and paralogous genes requires accurate multi-species alignments. Initial amino acid alignments were generated using TCOFFEE[34] and converted to nucleotide alignments (Supplementary Table 8). To reduce biases in downstream analyses, a simple computational screen was developed to identify and mask problematic regions of each alignment (Supplementary Information section 6). Overall, 2.8% of bases were masked in the *melanogaster* group alignments, and 3.0% of bases were masked in the full 12 species alignments, representing 8.5% and 13.8% of alignment columns, respectively. The vast majority of masked bases are masked in no more than one species (Supplementary Fig. 3), suggesting that the masking procedure is not simply eliminating rapidly evolving regions of the genome. We find an appreciably higher frequency of masked bases in lower-quality *D. simulans* and *D. sechellia* assemblies, compared to the more divergent (from *D. melanogaster*) but higher-quality *D. erecta* and *D. yakuba* assemblies, suggesting a higher error rate in accurately predicting and aligning gene models in lower-quality assemblies (Supplementary Information section 6 and Supplementary Fig. 3). We used masked versions of the alignments, including only the longest *D. melanogaster* transcripts for all subsequent analysis unless otherwise noted.

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

## Evolution of genome structure

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

To investigate overall conservation of genome architecture at an intermediate scale, we analysed synteny relationships across species using Synpipe[32] (Supplementary Information section 9.1). Synteny block size and average number of genes per block varies across the phylogeny as expected, with the number of blocks increasing and the average size of blocks decreasing with increasing evolutionary distance from *D. melanogaster* (A. Bhutkar, S. Russo, T. F. Smith and W. M. Gelbart, personal communication) (Supplementary Fig. 4). We inferred 112 syntenic blocks between *D. melanogaster* and *D. sechellia* (with an average of 122 genes per block), compared to 1,406 syntenic blocks between *D. melanogaster* and *D. grimshawi* (with an average of 8 genes per block). On average, 66% of each genome assembly was covered by syntenic blocks, ranging from 68% in *D. sechellia* to 58% in *D. grimshawi*.

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

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

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

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

Previous analysis has revealed heterogeneity in rearrangement rates among close relatives: careful inspection of 29 inversions that differentiate the chromosomes of *D. melanogaster* and *D. yakuba* revealed that 28 were fixed in the lineage leading to *D. yakuba*, and only one was fixed on the lineage leading to *D. melanogaster*[39]. Rearrangement rates are also heterogeneous across the genome among the 12 species: simulations reject a random-breakage model, which assumes that all sites are free to break in inversion events, but fail to reject a model of coldspots and hotspots for breakpoints (S. Schaeffer, personal communication). Furthermore, inversions seem to have played important roles in the process of speciation in at least some of these taxa[40].

One particularly striking example of the dynamic nature of genome micro-structure in *Drosophila* is the homeotic *homeobox (Hox)* gene cluster(s)[41]. *Hox* genes typically occur in genomic clusters, and this clustering is conserved across many vertebrate and invertebrate taxa, suggesting a functional role for the precise and collinear arrangement of these genes. However, several cluster splits have been previously identified in *Drosophila*[42,43], and the 12 *Drosophila* genome sequences provide additional evidence against the functional importance of *Hox* gene clustering in *Drosophila*. There are seven different gene arrangements found across 13 *Drosophila* species (the 12 sequenced genomes and *D. buzzatii*), with no species retaining the inferred ancestral gene order[44]. It thus seems that, in *Drosophila*, *Hox* genes do not require clustering to maintain proper function, and are a powerful illustration of the dynamism of genome structure across the sequenced genomes.

**Transposable element evolution.** Mobile, repetitive transposable element sequences are a particularly dynamic component of eukaryotic genomes. Transposable element/repeat content (in scaffolds >200 kb) varies by over an order of magnitude across the genus, ranging from ~2.7% in *D. simulans* and *D. grimshawi* to ~25% in *D. ananassae* (Table 2, and Supplementary Fig. 1). These data support the lower euchromatic transposable element content in *D. simulans* relative to *D. melanogaster*[45], and reveal that euchromatic transposable element/repeat content is generally similar within the *melanogaster* subgroup. Within the *Drosophila* subgenus, *D. grimshawi* has the lowest transposable element/repeat content, possibly relating to its ecological status as an island endemic, which may minimize the chance for horizontal transfer of transposable element families. Finally, the highest levels of transposable element/repeat content are found in *D. ananassae* and *D. willistoni*. These species also have the highest numbers of pseudo-transfer (t)RNA genes (Table 2), indicating a potential relationship between pseudo-tRNA genesis and repetitive DNA, as has been established in the mouse genome[36].

Different classes of transposable elements can vary in abundance owing to a variety of host factors, motivating an analysis of the intragenomic ecology of transposable elements in the 12 genomes. In *D. melanogaster*, long terminal repeat (LTR) retrotransposons have the highest abundance, followed by LINE (long interspersed nuclear element)-like retrotransposons and terminal inverted repeat (TIR) DNA-based transposons[18]. An unbiased, conservative approach (Supplementary Information section 3) for estimating the rank order abundance of major transposable element classes suggests that these abundance trends are conserved across the entire genus (Supplementary Fig. 5). Two exceptions are an increased abundance of TIR elements in *D. erecta* and a decreased abundance of LTR elements in *D. pseudoobscura*; the latter observation may represent an assembly artefact because the sister species *D. persimilis* shows typical LTR abundance. Given that individual instances of transposable element repeats and transposable element families themselves are not conserved across the genus, the stability of abundance trends for different classes of transposable elements is striking and suggests common mechanisms for host–transposable element co-evolution in *Drosophila*.

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

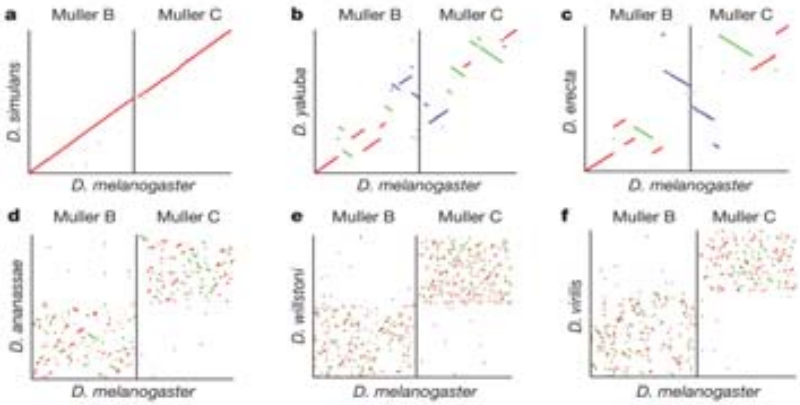



**Figure 3 | Synteny plots for Muller elements B and C with respect to *D. melanogaster* gene order.** The horizontal axis shows *D. melanogaster* gene order for Muller elements B and C, and the vertical axis maps homologous locations[52,53] in individual species (**a**–**f** in increasing evolutionary distance from *D. melanogaster*). Left to right on the *x* axis is from telomere to centromere for Muller element B, followed by Muller element C from centromere to telomere. Red and green lines represent syntenic segments in the same or reverse orientation along the chromosome relative to *D. melanogaster*, respectively. Blue segments show gene transposition of genes from one element to the other.

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

The use of these 12 genomes also facilitated the discovery of transposable element lineages not yet documented in *Drosophila*, specifically the P instability factor (*PIF*) superfamily of DNA transposons. Our analysis indicates that there are four distinct lineages of this transposon in *Drosophila*, and that this element has indeed colonized many of the sequenced genomes[53]. This superfamily is particularly intriguing given that *PIF*-transposase-like genes have been implicated in the origin of at least seven different genes during the *Drosophila* radiation[53], suggesting that not only do transposable elements affect the evolution of genome structure, but that their domestication can play a part in the emergence of novel genes.

*D. melanogaster* maintains its telomeres by occasional targeted transposition of three telomere-specific non-LTR retrotransposons (*HeT-A*, *TART* and *TAHRE*) to chromosome ends[54,55] and not by the more common mechanism of telomerase-generated G-rich repeats[56]. Multiple telomeric retrotransposons have originated within the genus, where they now maintain telomeres, and recurrent loss of most of the ORF2 from telomeric retrotransposons (for example, *TAHRE*) has given rise to half-telomeric-retrotransposons (for example, *HeT-A*) during *Drosophila* evolution[57]. The phylogenetic relationship among these telomeric elements is congruent with the species phylogeny, suggesting that they have been vertically transmitted from a common ancestor[57].

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

Of 78 previously reported micro (mi)RNA genes, 71 (91%) are highly conserved across the entire genus, with the remaining seven genes (*mir-2b-1*, *-289*, *-303*, *-310*, *-311*, *-312* and *-313*) restricted to the subgenus *Sophophora* (Supplementary Information section 7.2). All the species contain similar numbers of spliceosomal snRNA genes (Table 2), including at least one copy each of the four U12-dependent (minor) spliceosomal RNAs, despite evidence for birth and death of these genes and the absence of stable subtypes[61]. The unusual, lineage-specific expansion in size of U11 snRNA, previously described in *Drosophila*[61,62], is even more extreme in *D. willistoni*. We annotated 99 copies of the 5S ribosomal (r)RNA gene in a cluster in *D. melanogaster*, and between 13 and 73 partial 5S rRNA genes in clusters in the other genomes. Finally, we identified members of several other classes of ncRNA genes, including the RNA components of the RNase P (1 per genome) and the signal recognition particle (SRP) RNA complexes (1–3 per genome), suggesting that these functional RNAs are involved in similar biological processes throughout the genus. We were only able to locate the *roX* (RNA on X)[63,64] genes involved in dosage compensation using nucleotide homology in the *melanogaster* subgroup, although analyses incorporating structural information have identified *roX* genes in other members of the genus[65].

We investigated the evolution of rRNA genes in the 12 sequenced genomes, using trace archives to locate sequence variants within the transcribed portions of these genes. This analysis revealed moderate levels of variation that are not distributed evenly across the rRNA genes, with fewest variants in conserved core coding regions, more variants in coding expansion regions, and higher still variant abundances in non-coding regions. The level and distribution of sequence variation in rRNA genes are suggestive of concerted evolution, in which recombination events uniformly distribute variants throughout the rDNA loci, and selection dictates the frequency to which variants can expand[66].

**Protein-coding gene family evolution.** For a general perspective on how the protein-coding composition of these 12 genomes has changed, we examined gene family expansions and contractions in the 11,434 gene families (including those of size one in each species) predicted to be present in the most recent common ancestor of the two subgenera. We applied a maximum likelihood model of gene gain and loss[67] to estimate rates of gene turnover. This analysis suggests that gene families expand or contract at a rate of 0.0012 gains and losses per gene per million years, or roughly one fixed gene gain/loss across the genome every 60,000 yr[68]. Many gene families (4,692 or 41.0%) changed in size in at least one species, and 342 families showed significantly elevated ($P < 0.0001$) rates of gene gain and loss compared to the genomic average, indicating that non-neutral processes may play a part in gene family evolution. Twenty-two families exhibit rapid copy number evolution along the branch leading to *D. melanogaster* (eighteen contractions and four expansions; Supplementary Table 10). The most common Gene Ontology (GO) terms among families with elevated rates of gain/loss include 'defence response', 'protein binding', 'zinc ion binding', 'proteolysis', and 'trypsin activity'. Interestingly, genes involved in 'defence response' and 'proteolysis' also show high rates of protein evolution (see below). We also found heterogeneity in overall rates of gene gain and loss across lineages, although much of this variation could result from interspecific differences in assembly quality[68].

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

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

expression studies on seven of the species show that species-restricted genes are more likely to exhibit male-biased expression[29]. Further, these genes are significantly more tissue-specific in expression (as measured by τ; ref. 74) (MWU, $P = 9.6 \times 10^{-6}$), and this pattern is not solely driven by genes with testis-specific expression patterns.

### Protein-coding gene evolution

**Positive selection and selective constraints in *Drosophila* genomes.** To study the molecular evolution of protein-coding genes, we estimated rates of synonymous and non-synonymous substitution in 8,510 single-copy orthologues within the six *melanogaster* group species using PAML[75] (Supplementary Information section 11.1); synonymous site saturation prevents analysis of more divergent comparisons. We investigate only single-copy orthologues because when paralogues are included, alignments become increasingly problematic. Rates of amino acid divergence for single-copy orthologues in all 12 species were also calculated; these results are largely consistent with the analysis of non-synonymous divergence in the *melanogaster* group, and are not discussed further.

To understand global patterns of divergence and constraint across functional classes of genes, we examined the distributions of $\omega$ ($=d_N/d_S$, the ratio of non-synonymous to synonymous divergence) across Gene Ontology categories (GO)[76], excluding GO annotations based solely on electronic support (Supplementary Information section 11.2). Most functional categories of genes are strongly constrained, with median estimates of $\omega$ much less than one. In general, functionally similar genes are similarly constrained: 31.8% of GO categories have significantly lower variance in $\omega$ than expected ($q$-value true-positive test[77]). Only 11% of GO categories had statistically significantly elevated $\omega$ (relative to the median of all genes with GO annotations) at a 5% false-discovery rate (FDR), suggesting either positive selection or a reduction in selective constraint. The GO categories with elevated $\omega$ include the biological process terms 'defence response', 'proteolysis', 'DNA metabolic process' and 'response to biotic stimulus'; the molecular function terms 'transcription factor activity', 'peptidase activity', 'receptor binding', 'odorant binding', 'DNA binding', 'receptor activity' and 'G-protein-coupled receptor activity'; and the cellular location term 'extracellular' (Fig. 4, and Supplementary Table 12). Similar results are obtained when $d_N$ is compared across GO categories, suggesting that in most cases differences in $\omega$ among GO categories is driven by amino acid rather than synonymous site substitutions. The two exceptions are the molecular function terms 'transcription factor activity' and 'DNA binding activity', for which we observe significantly decelerated $d_S$ (FDR $= 7.2 \times 10^{-4}$ for both; Supplementary Information section 11.2) and no significant differences in $d_N$.

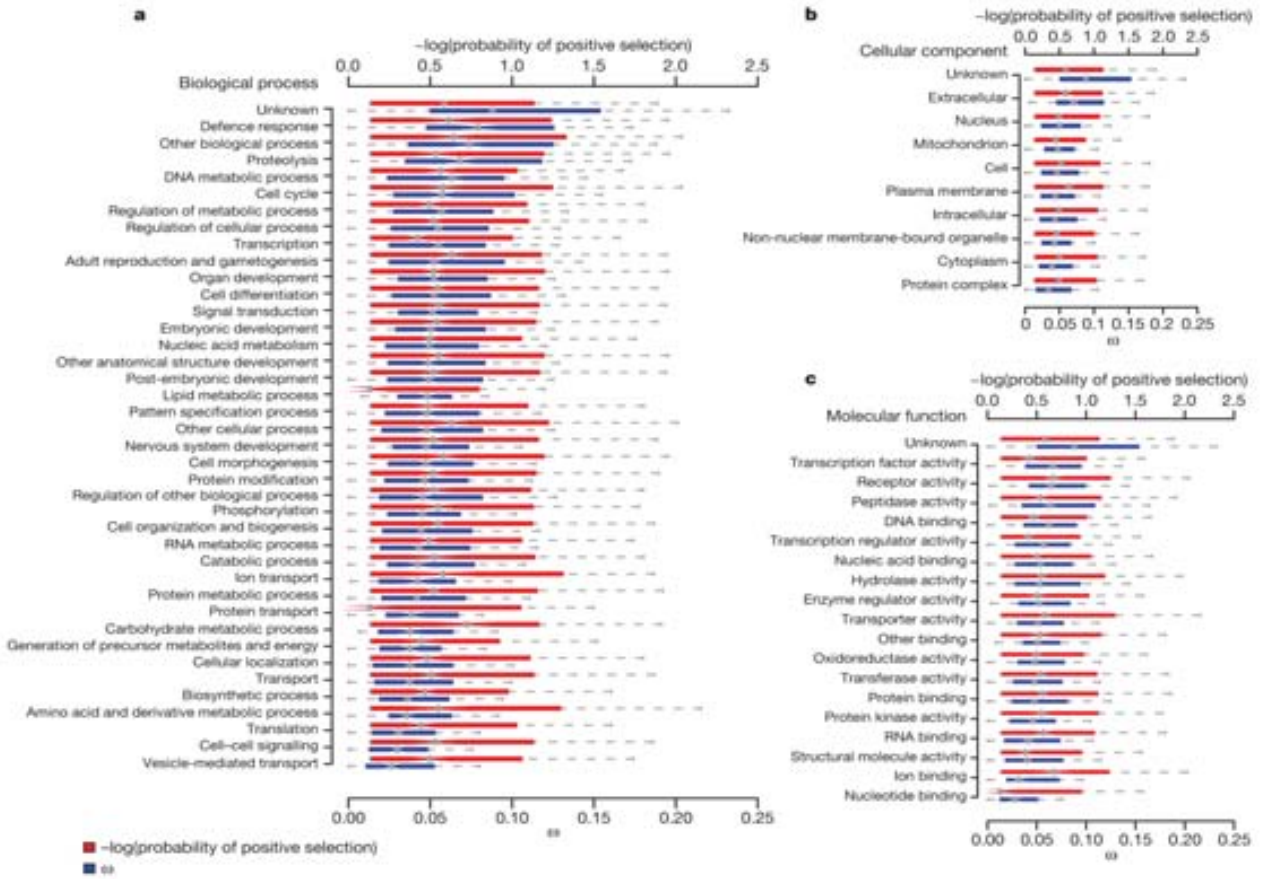



**Figure 4 | Patterns of constraint and positive selection among GO terms.** Distribution of average $\omega$ per gene and the negative $\log_{10}$ of the probability of positive selection (Supplementary Information section 11.2) for genes annotated with: **a**, biological process GO terms; **b**, cellular component GO terms; and **c**, molecular function GO terms. Only GO terms with 200 or more genes annotated are plotted. See Supplementary Table 12 for median values and significance. Note that most genes evolve under evolutionary constraint at most of their sites, leading to low values of $\omega$; even genes that experience positive selection do not typically have an average $\omega$ across all codons that exceeds one.

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

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

Previous work has suggested that a substantial fraction of non-synonymous substitutions in *Drosophila* were fixed through positive selection[81–85]. We estimate that 33.1% of single-copy orthologues in the *melanogaster* group have experienced positive selection on at least a subset of codons ($q$-value true-positive tests[77]) (Supplementary Information section 11.1). This may be an underestimate, because we have only examined single-copy orthologues, owing to difficulties in producing accurate alignments of paralogues by automated methods. On the basis of the 878 genes inferred to have experienced positive selection with high confidence (FDR < 10%), we estimated that an average of 2% of codons in positively selected genes have $\omega > 1$. Thus, several lines of evidence, based on different methodologies, suggest that patterns of amino acid fixation in *Drosophila* genomes have been shaped extensively by positive selection.

The presence of functional domains within a protein may lead to heterogeneity in patterns of constraint and adaptation along its length. Among genes inferred to be evolving by positive selection at a 10% FDR, 63.7% ($q$-value true-positive tests[77]) show evidence for spatial clustering of positively selected codons (Supplementary Information section 11.2). Spatial heterogeneity in constraint is further supported by contrasting $\omega$ for codons inside versus outside defined InterPro domains (genes lacking InterPro domains are treated as 'outside' a defined InterPro domain). Codons within InterPro domains were significantly more conserved than codons outside InterPro domains (median $\omega$: 0.062 InterPro domains, 0.084 outside InterPro domains; MWU, $P < 2.2 \times 10^{-16}$; Supplementary Information section 11.2). Similarly, there were significantly more positively selected codons outside of InterPro domains than inside domains (FET $P < 2.2 \times 10^{-16}$), suggesting that in addition to being more constrained, codons in protein domains are less likely to be targets of positive selection (Supplementary Fig. 6).

**Factors affecting the rate of protein evolution in *Drosophila*.** The sequenced genomes of the *melanogaster* group provide unprecedented statistical power to identify factors affecting rates of protein evolution. Previous analyses have suggested that although the level of gene expression consistently seems to be a major determinant of variation in rates of evolution among proteins[86,87], other factors probably play a significant, if perhaps minor, part[88–91]. In *Drosophila*, although highly expressed genes do evolve more slowly, breadth of expression across tissues, gene essentiality and intron number all also independently correlate with rates of protein evolution, suggesting that the additional complexities of multicellular organisms are important factors in modulating rates of protein evolution[78]. The presence of repetitive amino acid sequences has a role as well: non-repeat regions in proteins containing repeats evolve faster and show more evidence for positive selection than genes lacking repeats[92].

These data also provide a unique opportunity to examine the impact of chromosomal location on evolutionary rates. Population genetic theory predicts that for new recessive mutations, both purifying and positive selection will be more efficient on the X chromosome given its hemizygosity in males[93]. In contrast, the lack of recombination on the small, mainly heterochromatic dot chromosome[94,95] is expected to reduce the efficacy of selection[96]. Because codon bias, or the unequal usage of synonymous codons in protein-coding sequences, reflects weak but pervasive selection, it is a sensitive metric for evaluating the efficacy of purifying selection. Consistent with expectation, in all 12 species, we find significantly elevated levels of codon bias on the X chromosome and significantly reduced levels of codon bias on the dot chromosome[97]. Furthermore, X-chromosome-linked genes are marginally over-represented within the set of positively selected genes in the *melanogaster* group (FET, $P = 0.055$), which is consistent with increased rates of adaptive substitution on this chromosome. This analysis suggests that chromosomal context also serves to modulate rates of molecular evolution in protein-coding genes.

To examine further the impact of genomic location on protein evolution, we examined the subset of genes that have moved within or between chromosome arms[92,98]. Genes inferred to have moved between Muller elements have a significantly higher rate of protein evolution than genes inferred to have moved within a Muller element (MWU, $P = 1.32 \times 10^{-14}$) and genes that have maintained their genomic position (MWU, $P = 0.008$) (Supplementary Fig. 7). Interestingly, genes that move within Muller elements have a significantly lower rate of protein evolution than those for which genomic locations have been maintained (MWU, $P = 3.85 \times 10^{-14}$). It remains unclear whether these differences reflect underlying biases in the types of genes that move inter- versus intra-chromosomally, or whether they are due to *in situ* patterns of evolution in novel genomic contexts.

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

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

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

**Chemoreception.** *Drosophila* species have complex olfactory and gustatory systems used to identify food sources, hazards and mates, which depend on odorant-binding proteins, and olfactory/odorant and gustatory receptors (*Or*s and *Gr*s). The *D. melanogaster* genome has approximately 60 *Or*s, 60 *Gr*s and 50 odorant-binding protein genes. Despite overall conservation of gene number across the 12 species and widespread evidence for purifying selection within the *melanogaster* group, there is evidence that a subset of *Or* and *Gr* genes experiences positive selection[121–123]. Furthermore, clear lineage-specific differences are detectable between generalist and specialist species within the *melanogaster* subgroup. First, the two independently evolved specialists (*D. sechellia* and *D. erecta*) are losing *Gr* genes approximately five times more rapidly than the generalist species[121,124]. We believe this result is robust to sequence quality, because all pseudogenes and deletions were verified by direct re-sequencing and synteny-based orthologue searches, respectively. Generalists are expected to encounter the most diverse set of tastants and seem to have maintained the greatest diversity of gustatory receptors. Second, *Or* and *Gr* genes that remain intact in *D. sechellia* and *D. erecta* evolve significantly more rapidly along these two lineages ($\omega = 0.1556$ for *Or*s and 0.1874 for *Gr*s) than along the generalist lineages ($\omega = 0.1049$ for *Or*s and 0.1658 for *Gr*s; paired Wilcoxon, $P = 0.0003$ and 0.003, respectively[124]). There is some evidence that odorant-binding protein genes also evolve significantly faster in specialists compared to generalists[122]. This elevated $\omega$ reflects a trend observed throughout the genomes of the two specialists and is likely to result, at least in part, from demographic phenomena. However, the difference between specialist and generalist $\omega$ for *Or*/*Gr* genes (0.0292) is significantly greater than the difference for genes across the genome (0.0091; MWU, $P = 0.0052$)[121], suggesting a change in selective regime. Moreover, the observation that elevated $\omega$ as well as accelerated gene loss disproportionately affect groups of *Or* and *Gr* genes that respond to specific chemical ligands and/or are expressed during specific life stages suggests that rapid evolution at *Or*/*Gr* loci in specialists is related to the ecological shifts these species have sustained[121].

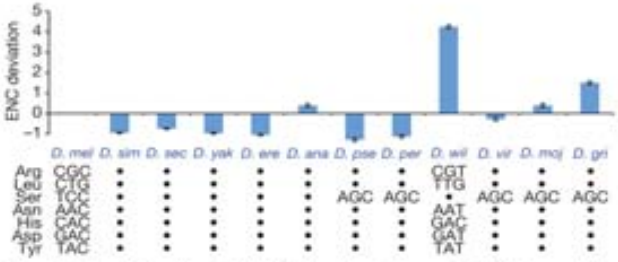



**Figure 5 | Deviations in codon bias from *D. melanogaster* in 11 *Drosophila* species.** The upper panel depicts differences in ENC (effective number of codons) between *D. melanogaster* and the 11 non-*melanogaster* species, calculated on a gene-by-gene basis. Note that increasing levels of ENC indicates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

**Detoxification/metabolism.** The larval food sources for many *Drosophila* species contain a cocktail of toxic compounds, and consequently *Drosophila* genomes encode a wide variety of detoxification proteins. These include members of the cytochrome P450 (P450), carboxyl/choline-esterase (CCE) and glutathione *S*-transferase (GST) multigene families, all of which also have critical roles in resistance to insecticides[125–127]. Among the P450s, the five enzymes associated with insecticide resistance are highly dynamic across the phylogeny, with 24 duplication events and 4 loss events since the last common ancestor of the genus, which is in striking contrast to genes with known developmental roles, eight of which are present as a single copy in all 12 species (C. Robin, personal communication). As with chemoreceptors, specialists seem to lose detoxification genes at a faster rate than generalists. For instance, *D. sechellia* has lost the most P450 genes; these 14 losses comprise almost one-third of all P450 loss events (Supplementary Table 13) (C. Robin, personal communication). Positive selection has been implicated in detoxification-gene evolution as well, because a search for positive selection among GSTs identified the parallel evolution of a radical glycine to lysine amino acid change in GSTD1, an enzyme known to degrade DDT[128]. Finally, although metabolic enzymes in general are highly constrained (median $\omega = 0.045$ for enzymes, 0.066 for non-enzymes; MWU, $P = 5.7 \times 10^{-24}$), enzymes involved in xenobiotic metabolism evolve significantly faster than other enzymes (median $\omega = 0.05$ for the xenobiotic group versus $\omega = 0.045$ overall, two-tailed permutation test, $P = 0.0110$; A. J. Greenberg, personal communication).

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

**Immunity/defence.** *Drosophila*, like all insects, possesses an innate immune system with many components analogous to the innate immune pathways of mammals, although it lacks an antibody-mediated adaptive immune system[131]. Immune system genes often evolve rapidly and adaptively, driven by selection pressures from pathogens and parasites[132–134]. The genus *Drosophila* is no exception: immune system genes evolve more rapidly than non-immune genes, showing both high total divergence rates and specific signs of positive selection[135]. In particular, 29% of receptor genes involved in phagocytosis seem to evolve under positive selection, suggesting that molecular co-evolution between *Drosophila* pattern recognition receptors and pathogen antigens is driving adaptation in the immune system[135]. Somewhat surprisingly, genes encoding effector proteins such as antimicrobial peptides are far less likely to exhibit adaptive sequence evolution. Only 5% of effector genes (and no antimicrobial peptides) show evidence of adaptive evolution, compared to 10% of genes genome-wide. Instead, effector genes seem to evolve by rapid duplication and deletion. Whereas 49% of genes genome-wide, 63%

of genes involved in pathogen recognition and 81% of genes implicated in immune-related signal transduction can be found as single-copy orthologues in all 12 species, only 40% of effector genes exist as single-copy orthologues across the genus ($\chi 2 = 41.13$, $P = 2.53 \times 10^{-8}$), suggesting rapid radiation of effector protein classes along particular lineages[135]. Thus, much of the *Drosophila* immune system seems to evolve rapidly, although the mode of evolution varies across immune-gene functional classes.

**Sex/reproduction.** Genes encoding sex- and reproduction-related proteins are subject to a wide array of selective forces, including sexual conflict, sperm competition and cryptic female choice, and to the extent that these selective forces are of evolutionary consequence, this should lead to rapid evolution in these genes[136] (for an overview see refs 137, 138). The analysis of 2,505 sex- and reproduction-related genes within the *melanogaster* group indicated that male sex- and reproduction-related genes evolve more rapidly at the protein level than genes not involved in sex or reproduction or than female sex- and reproduction-related genes (Supplementary Fig. 8). Positive selection seems to be at least partially responsible for these patterns, because genes involved in spermatogenesis have significantly stronger evidence for positive selection than do non-spermatogenesis genes (permutation test, $P = 0.0053$). Similarly, genes that encode components of seminal fluid have significantly stronger evidence for positive selection than 'non-sex' genes[139]. Moreover, protein-coding genes involved in male reproduction, especially seminal fluid and testis genes, are particularly likely to be lost or gained across *Drosophila* species[29,139].

**Evolutionary forces in the mitochondrial genome.** Functional elements in mtDNA are strongly conserved, as expected: tRNAs are relatively more conserved than the mtDNA overall (average pairwise nucleotide distance = 0.055 substitutions per site for tRNAs versus 0.125 substitutions per site overall). We observe a deficit of substitutions occurring in the stem regions of the stem-loop structure in tRNAs, consistent with strong selective pressure to maintain RNA secondary structure, and there is a strong signature of purifying selection in protein-coding genes[13]. However, despite their shared role in aerobic respiration, there is marked heterogeneity in the rates of amino acid divergence between the oxidative phosphorylation enzyme complexes across the 12 species (NADH dehydrogenase, 0.059 > ATPase, 0.042 > CytB, 0.037 > cytochrome oxidase, 0.020; mean pairwise $d_N$), which contrasts with the relative homogeneity in synonymous substitution rates. A model with distinct substitution rates for each enzyme complex rather than a single rate provides a significantly better fit to the data ($P < 0.0001$), suggesting complex-specific selective effects of mitochondrial mutations[13].

## Non-coding sequence evolution

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

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

To link patterns of evolution with structural constraints, we inferred ancestral pre-miRNA sequences and deduced secondary structures at each ancestral node on the phylogeny (Supplementary Information section 12.1). Although conserved miRNA genes show little structural change (little change in free energy), the five *melanogaster* group-specific miRNA genes (*miR-303* and the *mir-310/311/312/313* cluster) have undergone numerous changes across the entire pre-miRNA sequence, including the ordinarily invariant mature miRNA. Patterns of polymorphism and divergence in these lineage-specific miRNA genes, including a high frequency of derived mutations, are suggestive of positive selection[140]. Although lineage-specific miRNAs may evolve under less constraint because they have fewer target transcripts in the genome, it is also possible that recent integration into regulatory networks causes accelerated rates of miRNA evolution.

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

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

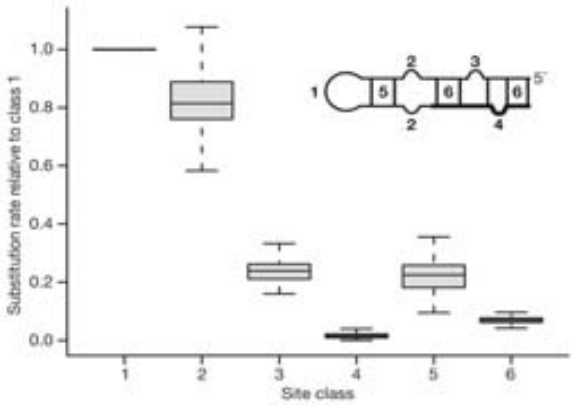



**Figure 6 | Substitution rate of site classes within miRNAs.** Bootstrap distributions of miRNA substitution rates. Structural alignments of miRNA precursor hairpins were partitioned into six site-classes (inset): (1) hairpin loops; unpaired sites (2) outside, (3) in the complementary region of, and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

Finally, we predicted a set of conserved intronic RNA structures and analysed patterns of compensatory nucleotide substitution in *D. melanogaster*, *D. yakuba*, *D. ananassae*, *D. pseudoobscura*, *D. virilis* and *D. mojavensis* (Supplementary Information section 13). Signatures of compensatory evolution in RNA helices are detected as covarying nucleotide sites or 'covariations' (that is, two Watson–Crick bases that interact in species A replaced by a different Watson–Crick pair in species B). The number of covariations (per base pair of a helix) depends on the physical distance between the interacting nucleotides (Supplementary Fig. 9), as has been observed for the RNA helices in the *Drosophila bicoid* 3′ UTR region[142]. Short-range pairings exhibit a higher average number of covariations with a larger variance among helices than longer-range pairings. The decrease in rate of covariation with increasing distance may be explained by physical properties of a helix, which may impose selective constraints on the evolution of covarying nucleotides within a helix. Alternatively, if individual mutations at each locus are deleterious but compensated by mutations at a second locus, given sufficiently strong selection against the first deleterious mutation these epistatic fitness interactions could generate the observed distance effect[143].

**Evolution of *cis*-regulatory DNAs.** Comparative analyses of *cis*-regulatory sequences may provide insights into the evolutionary forces acting on regulatory components of genes, shed light on the constraints of the *cis*-regulatory code and aid in annotation of new regulatory sequences. Here we rely on two recently compiled databases, and present results comparing *cis*-regulatory modules[144] and transcription factor binding sites (derived from DNase I footprints)[145] between *D. melanogaster* and *D. simulans* (Supplementary Information section 8). We estimated mean selective constraint (*C*, the fraction of mutations removed by natural selection) relative to the 'fastest evolving intron' sites at the 5′ end of short introns, which represent putatively unconstrained neutral standards (Supplementary Information section 8.2)[146]. Note that this approach ignores the contribution of positively selected sites, potentially underestimating the fraction of functionally relevant sites[147].

Consistent with previous findings, *Drosophila cis*-regulatory sequences are highly constrained[148,149]. Mean constraint within *cis*-regulatory modules is 0.643 (95% bootstrap confidence interval = 0.621–0.662) and within footprints is 0.692 (0.655–0.723), both of which are significantly higher than mean constraint in non-coding DNA overall (0.555 (0.546–0.563)) and significantly lower than constraint at non-degenerate coding sites (0.862 (0.856–0.868)) and ncRNA genes (0.864 (0.846–0.880)) (Supplementary Fig. 10). The high level of constraint in *cis*-regulatory sequences also extends into flanking sequences, only declining to constraint levels typical of non-coding DNA 40 bp away. This is consistent with previous findings that transcription factor binding sites tend to be found in larger blocks of constraint that cluster to form *cis*-regulatory modules[150]. To understand selective constraints on nucleotides within *cis*-regulatory sequences that have direct contact with transcription factors, we estimated the selective constraint for the best match to position weight matrices within each footprint[151]; core motifs in transcription-factor-binding sites have a mean constraint of 0.773 (0.729–0.814), significantly greater than the mean for the footprints as a whole, and approaching the level of constraint found at non-degenerate coding sites and in ncRNA genes (Supplementary Fig. 10).

We next examined the variation in selective constraint across *cis*-regulatory sequences. Surprisingly, we find no evidence that selective constraint is correlated with predicted transcription-factor-binding strength (estimated as the position weight matrix score *P*-value) (Spearman's $r = 0.0681$, $P = 0.0609$). We observe significant variation in constraint both among target genes (Kruskal–Wallis tests, footprints, $P < 0.0001$; and position weight matrix matches within footprints, $P = 0.0023$) and among chromosomes (*cis*-regulatory modules, $P = 0.0186$; footprints, $P = 0.0388$; and position weight matrix matches within footprints, $P = 0.0108$; Supplementary Table 16).

## Discussion and conclusion

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

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

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

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

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

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

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

## METHODS

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

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

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

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

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

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

**Received 19 July; accepted 5 October 2007.**


1. Markow, T. A. & O'Grady, P. M. Drosophila biology in the genomic age. *Genetics* doi:10.1534/genetics.107.074112 (in the press).
2. Powell, J. R. *Progress and Prospects in Evolutionary Biology: The Drosophila Model* (Oxford Univ. Press, Oxford, 1997).
3. Adams, M. D. *et al.* The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–2195 (2000).
4. Celniker, S. E. *et al.* Finishing a whole-genome shotgun: release 3 of the *Drosophila melanogaster* euchromatic genome sequence. *Genome Biol.* **3**, research0079.1–0079.14 (2002).
5. Richards, S. *et al.* Comparative genome sequencing of *Drosophila pseudoobscura*: chromosomal, gene, and *cis*-element evolution. *Genome Res.* **15**, 1–18 (2005).
6. Myers, E. W. *et al.* A whole-genome assembly of *Drosophila*. *Science* **287**, 2196–2204 (2000).
7. Fleischmann, R. D. *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496–512 (1995).
8. Stark et al.. Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. *Nature* doi:10.1038/nature06340 (this issue).
9. Begun, D. J. *et al.* Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *PLoS Biol.* **5**, e310, doi:10.1371/journal.pbio.0050310 (2007).
10. Zimin, A. V., Smith, D. R., Sutton, G. & Yorke, J. A. Assembly reconciliation. *Bioinformatics* (in the press).
11. Clary, D. O. & Wolstenholme, D. R. The mitochondrial DNA molecule of *Drosophila yakuba*: nucleotide sequence, gene organization, and genetic code. *J. Mol. Evol.* **22**, 252–271 (1985).
12. Ballard, J. W. When one is not enough: introgression of mitochondrial DNA in *Drosophila*. *Mol. Biol. Evol.* **17**, 1126–1130 (2000).
13. Montooth, K. L., Abt, D. N., Hoffman, J. & Rand, D. M. Evolution of the mitochondrial DNA across twelve species of *Drosophila*. *Mol. Biol. Evol.* (submitted).
14. Salzberg, S. *et al.* Serendipitous discovery of *Wolbachia* genomes in multiple *Drosophila* species. *Genome Biol.* **6**, R23 (2005).
15. Edgar, R. C. & Myers, E. W. PILER: identification and classification of genomic repeats. *Bioinformatics* **21**, i152–i158 (2005).
16. Smith, C. D. *et al.* Improved repeat identification and masking in Dipterans. *Gene* **389**, 1–9 (2007).
17. Li, Q. *et al.* ReAS: Recovery of ancestral sequences for transposable elements from the unassembled reads of a whole shotgun. *PloS Comput. Biol.* **1**, e43 (2005).
18. Bergman, C. M., Quesneville, H., Anxolabehere, D. & Ashburner, M. Recurrent insertion and duplication generate networks of transposable element sequences in the *Drosophila melanogaster* genome. *Genome Biol.* **7**, R112 (2006).
19. Guigo, R., Knudsen, S., Drake, N. & Smith, T. Prediction of gene structure. *J. Mol. Biol.* **226**, 141–157 (1992).
20. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
21. Gross, S. S. & Brent, M. R. Using multiple alignments to improve gene prediction. *J. Comput. Biol.* **13**, 379–393 (2006).
22. Gross, S. S., Do, C. B. & Batzoglou, S. in *BCATS 2005 Symposium Proc.* 82 (2005).
23. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14**, 988–995 (2004).
24. Slater, G. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).
25. Chatterji, S. & Pachter, L. Reference based annotation with GeneMapper. *Genome Biol.* **7**, R29 (2006).
26. Souvorov, A. *et al.* in *NCBI News Fall/Winter, NIH Publication No. 04-3272* (eds Benson, D & Wheeler, D) (2006).
27. Honeybee Genome Sequencing Consortium. Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature* **443**, 931–949 (2006).
28. Elsik, C. G. *et al.* Creating a honey bee consensus gene set. *Genome Biol.* **8**, R13 (2007).
29. Zhang, Y., Sturgill, D., Parisi, M., Kumar, S. & Oliver, B. Constraint and turnover in sex-biased gene expression in the genus *Drosophila*. *Nature* doi:10.1038/nature06323 (this issue).
30. Manak, J. R. *et al.* Biological function of unannotated transcription during the early development of *Drosophila melanogaster*. *Nature Genet.* **38**, 1151–1158 (2006).
31. Tatusov, R. L., Koonin, E. V. & Lipman, D. J. A genomic perspective on protein families. *Science* **278**, 631–637 (1997).
32. Bhutkar, A., Russo, S., Smith, T. F. & Gelbart, W. M. Techniques for multi-genome synteny analysis to overcome assembly limitations. *Genome Informatics* **17**, 152–161 (2006).
33. Heger, A. & Ponting, C. Evolutionary rate analyses of orthologues and paralogues from twelve *Drosophila* genomes. doi:10.1101/gr6249707 *Genome Res.* (in the press).
34. Notredame, C., Higgins, D. G. & Heringa, J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302**, 205–217 (2000).
35. Rat Genome Sequencing Project Consortium. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**, 493–521 (2004).
36. Waterston, R. H. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
37. Harrison, P. M., Milburn, D., Zhang, Z., Bertone, P. & Gerstein, M. Identification of pseudogenes in the *Drosophila melanogaster* genome. *Nucleic Acids Res.* **31**, 1033–1037 (2003).
38. Bosco, G., Campbell, P., Leiva-Neto, J. & Markow, T. Analysis of *Drosophila* species genome size and satellite DNA content reveals significant differences among strains as well as between species. *Genetics* doi:10.1534/Genetics107.075069 (in the press).
39. Ranz, J. *et al.* Principles of genome evolution in the *Drosophila melanogaster* species group. *PLoS Biol.* **5**, e152, doi:10.1371/journal.pbio.0050152 (2007).
40. Noor, M. A. F., Garfield, D. A., Schaeffer, S. W. & Machado, C. A. Divergence between the *Drosophila pseudoobscura* and *D. persimilis* genome sequences in relation to chromosomal inversions. *Genetics* doi:10.1534/genetics.107.070672 (in the press).
41. Lewis, E. B. A gene complex controlling segmentation in *Drosophila*. *Nature* **276**, 565–570 (1978).
42. Negre, B., Ranz, J. M., Casals, F., Caceres, M. & Ruiz, A. A new split of the *Hox* gene complex in *Drosophila*: relocation and evolution of the gene labial. *Mol. Biol. Evol.* **20**, 2042–2054 (2003).
43. Von Allmen, G. *et al.* Splits in fruitfly *Hox* gene complexes. *Nature* **380**, 116 (1996).
44. Negre, B. & Ruiz, A. HOM-C evolution in *Drosophila*: is there a need for *Hox* gene clustering? *Trends Genet.* **23**, 55–59 (2007).
45. Dowsett, A. P. & Young, M. W. Differing levels of dispersed repetitive DNA among closely related species of *Drosophila*. *Proc. Natl Acad. Sci.* **79**, 4570–4574 (1982).
46. Kapitonov, V. V. & Jurka, J. DNAREPI_DM. (Repbase Update Release 3.4, 1999).
47. Kapitonov, V. V. & Jurka, J. Molecular paleontology of transposable elements in the *Drosophila melanogaster* genome. *Proc. Natl Acad. Sci. USA* **100**, 6569–6574 (2003).

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

115. Akashi, H. *et al.* Molecular evolution in the *Drosophila melanogaster* species subgroup: Frequent parameter fluctuations on the timescale of molecular divergence. *Genetics* **172,** 1711–1726 (2006).
116. Bauer DuMont, V., Fay, J. C., Calabrese, P. P. & Aquadro, C. F. DNA variability and divergence at the *Notch* locus in *Drosophila melanogaster* and *D. simulans*: a case of accelerated synonymous site divergence. *Genetics* **167,** 171–185 (2004).
117. McVean, G. A. & Vieira, J. The evolution of codon preferences in *Drosophila*: a maximum-likelihood approach to parameter estimation and hypothesis testing. *J. Mol. Evol.* **49,** 63–75 (1999).
118. Nielsen, R., Bauer DuMont, V., Hubisz, M. J. & Aquadro, C. F. Maximum likelihood estimation of ancestral codon usage bias parameters in *Drosophila*. *Mol. Biol. Evol.* **24,** 228–235 (2007).
119. Begun, D. J. The frequency distribution of nucleotide variation in *Drosophila simulans*. *Mol. Biol. Evol.* **18,** 1343–1352 (2001).
120. Singh, N. S., Bauer DuMont, V. L., Hubisz, M. J., Nielsen, R. & Aquadro, C. F. Patterns of mutation and selection at synonymous sites in *Drosophila*. *Mol. Biol. Evol.* doi:10.1093/molbev.msm196 (in the press).
121. McBride, C. S. & Arguello, J. R. Five *Drosophila* genomes reveal non-neutral evolution and the signature of host specialization in the chemoreceptor superfamily. *Genetics* (in the press).
122. Vieira, F. G., Sanchez-Gracia, A. & Rozas, J. Comparative genomic analysis of the odorant-binding protein family in 12 *Drosophila* genomes: Purifying selection and birth-and-death evolution. *Genome Biol.* **8,** 235 (2007).
123. Gardiner, A., Barker, D., Butlne, R. K., Jordan, W. C. & Ritchie, M. G. *Drosophila* chemoreceptor evolution: Selection, specialisation and genome size. *Genome Biol.* (submitted).
124. McBride, C. S. Rapid evolution of smell and taste receptor genes during host specialization in *Drosophila sechellia*. *Proc. Natl Acad. Sci. USA* **104,** 4996–5001 (2007).
125. Ranson, H. *et al.* Evolution of supergene families associated with insecticide resistance. *Science* **298,** 179–181 (2002).
126. Tijet, N., Helvig, C. & Feyereisen, R. The cytochrome P450 gene superfamily in *Drosophila melanogaster*. *Gene* **262,** 189–198 (2001).
127. Claudianos, C. *et al.* A deficit of detoxification enzymes: pesticide sensitivity and environmental response in the honeybee. *Insect Mol. Biol.* **15,** 615–636 (2006).
128. Low, W. L. *et al.* Molecular evolution of glutathione *S*-transferases in the genus *Drosophila*. *Genetics* (in the press).
129. Castellano, S. *et al. In silico* identification of novel selenoproteins in the *Drosophila melanogaster* genome. *EMBO Rep.* **2,** 697–702 (2001).
130. Martin-Romero, F. J. *et al.* Selenium metabolism in *Drosophila*: selenoproteins, selenoprotein mRNA expression, fertility, and mortality. *J. Biol. Chem.* **276,** 29798–29804 (2001).
131. Lemaitre, B. & Hoffmann, J. The host defense of *Drosophila melanogaster*. *Annu. Rev. Immunol.* **25,** 697–743 (2007).
132. Hughes, A. L. & Nei, M. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* **335,** 167–170 (1988).
133. Murphy, P. M. Molecular mimicry and the generation of host defense protein diversity. *Cell* **72,** 823–826 (1993).
134. Schlenke, T. A. & Begun, D. J. Natural selection drives *Drosophila* immune system evolution. *Genetics* **164,** 1471–1480 (2003).
135. Sackton, T. B. *et al.* The evolution of the innate immune system across *Drosophila*. *Nature Genet.* (submitted).
136. Civetta, A. & Singh, R. S. High divergence of reproductive tract proteins and their association with postzygotic reproductive isolation in *Drosophila melanogaster* and *Drosophila virilis* group species. *J. Mol. Evol.* **41,** 1085–1095 (1995).
137. Civetta, A. Shall we dance or shall we fight? Using DNA sequence data to untangel controversies surrounding sexual selection. *Genome* **46,** 925–929 (2003).
138. Clark, N. L., Aagard, J. E. & Swanson, W. J. Evolution of reproductive proteins from animals and plants. *Reproduction* **131,** 11–22 (2006).
139. Haerty, W. *et al.* Evolution in the fast lane: rapidly evolving sex- and reproduction-related genes in *Drosophila* species. *Genetics* (in the press).
140. Lu, J. *et al.* Adaptive evolution of newly-emerged microRNA genes in *Drosophila*. *Mol. Biol. Evol.* (submitted).
141. Lai, E. C., Tomancak, P., Williams, R. W. & Rubin, G. M. Computational identification of *Drosophila* microRNA genes. *Genome Biol.* **4,** R42 (2003).
142. Parsch, J., Braverman, J. M. & Stephan, W. Comparative sequence analysis and patterns of covariation in RNA secondary structures. *Genetics* **154,** 909–921 (2000).
143. Stephan, W. The rate of compensatory evolution. *Genetics* **144,** 419–426 (1996).
144. Gallo, S. M., Li, L., Hu, Z. & Halfon, M. S. REDfly: a Regulatory Element Database for *Drosophila*. *Bioinformatics* **22,** 381–383 (2006).
145. Bergman, C. M., Carlson, J. W. & Celniker, S. E. *Drosophila* DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, *Drosophila melanogaster*. *Bioinformatics* **21,** 1747–1749 (2005).
146. Halligan, D. L. & Keightley, P. D. Ubiquitous selective constraints in the *Drosophila* genome revealed by a genome-wide interspecies comparison. *Genome Res.* **16,** 875–884 (2006).
147. Andolfatto, P. Adaptive evolution of non-coding DNA in *Drosophila*. *Nature* **437,** 1149–1152 (2005).
148. Bird, C. P., Stranger, B. E. & Dermitzakis, E. T. Functional variation and evolution of non-coding DNA. *Curr. Opin. Genet. Dev.* **16,** 559–564 (2006).
149. Wittkopp, P. J. Evolution of *cis*-regulatory sequence and function in Diptera. *Heredity* **97,** 139–147 (2006).
150. Ludwig, M. Z., Patel, N. H. & Kreitman, M. Functional analysis of *eve stripe 2* enhancer evolution in *Drosophila*. *Development* **125,** 949–958 (1998).
151. Down, A. T. A., Bergman, C. M., Su, J. & Hubbard, T. J. P. Large scale discovery of promoter motifs in *Drosophila melanogaster*. *PloS Comput. Biol.* **3,** e7 (2007).
152. Tamura, K., Subramanian, S. & Kumar, S. Temporal patterns of fruit fly (*Drosophila*) evolution revealed by mutation clocks. *Mol. Biol. Evol.* **21,** 36–44 (2004).
153. Kumar, S., Tamura, K. & Nei, M. MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment. *Brief. Bioinform.* **5,** 150–163 (2004).
154. Pollard, D. A., Iyer, V. N., Moses, A. M. & Eisen, M. B. Widespread discordance of gene trees with species tree in *Drosophila*: evidence for incomplete lineage sorting. *PLoS Genet.* **2,** e173 (2006).
155. Bhutkar, A., Gelbart, W. M. & Smith, T. F. Inferring genome-scale rearrangement phylogeny and ancestral gene order: A *Drosophila* case study. *Genome Biol.* (in the press).

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

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

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

**Author Information** Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to A.G.C. (ac347@cornell.edu), M.B.E. (mbeisen@lbl.gov), D.R.S. (douglas.smith@agencourt.com), C.M.B. (casey.bergman@manchester.ac.uk), W.G. (gelbart@morgan.harvard.edu), B.O. (oliver@helix.nih.gov), T.A.M. (tmarkow@public.ar1.arizona.edu), T.C.K. (kaufman@indiana.edu), M.K. (manoli@mit.edu), V.N.I. (venky@berkeley.edu), T.B.S. (tbs7@cornell.edu), A.M.L. (aml69@cornell.edu), D.A.P. (danielapollard@alum.bowdoin.edu), N.D.S. (nds25@cornell.edu), or collectively to 12flies@morgan.harvard.edu.

***Drosophila* 12 Genomes Consortium**

**Project Leaders** Andrew G. Clark[1], Michael B. Eisen[2,3], Douglas R. Smith[4], Casey M. Bergman[5], Brian Oliver[6], Therese A. Markow[7], Thomas C. Kaufman[8], Manolis Kellis[9,10] & William Gelbart[11,12]

**Annotation Coordination** Venky N. Iyer[13] & Daniel A. Pollard[14]

**Analysis/Writing Coordination** Timothy B. Sackton[1,15], Amanda M. Larracuente[1] & Nadia D. Singh[1]

**Sequencing, Assembly, Annotation and Analysis Contributors** Jose P. Abad[16], Dawn N. Abt[17], Boris Adryan[18], Montserrat Aguade[19], Hiroshi Akashi[20], Wyatt W. Anderson[21], Charles F. Aquadro[1], David H. Ardell[22], Roman Arguello[23], Carlo G. Artieri[24], Daniel A. Barbash[1], Daniel Barker[25], Paolo Barsanti[26], Phil Batterham[27], Serafim Batzoglou[28], Dave Begun[29], Arjun Bhutkar[11,30], Enrico Blanco[31], Stephanie A. Bosak[4], Robert K. Bradley[32], Adrianne D. Brand[4], Michael R. Brent[33], Angela N. Brooks[13], Randall H. Brown[33], Roger K. Butlin[34], Corrado Caggese[26], Brian R. Calvi[35], A. Bernardo de Carvalho[36], Anat Caspi[32], Sergio Castrezana[37], Susan E. Celniker[2], Jean L. Chang[10], Charles Chapple[38], Sourav Chatterji[38,39], Asif Chinwalla[40], Alberto Civetta[41], Sandra W. Clifton[40], Josep M. Comeron[42], James C. Costello[43], Jerry A. Coyne[23], Jennifer Daub[44], Robert G. David[4], Arthur L. Delcher[45], Kim Delehaunty[40], Chuong B. Do[28], Heather Ebling[4], Kevin Edwards[46], Thomas Eickbush[47], Jay D. Evans[48], Alan Filipski[49], Sven Findeiß[49,50], Eva Freyhult[22], Lucinda Fulton[40], Robert Fulton[40], Ana C. L. Garcia[51], Anastasia Gardiner[25], David A. Garfield[52], Barry E. Garvin[4], Greg Gibson[53], Don Gilbert[8], Sante Gnerre[10], Jennifer Godfrey[40], Robert Good[27], Valer Gotea[20], Brenton Gravely[54], Anthony J. Greenberg[1], Sam Griffiths-Jones[5,64], Samuel Gross[28], Roderic Guigo[31,55], Erik A. Gustafson[4], Wilfried Haerty[24], Matthew W. Hahn[8,43], Daniel L. Halligan[56], Aaron L. Halpern[57], Gillian M. Halter[20], Mira V. Han[43], Andreas Heger[58,59], LaDeana Hillier[40], Angie S. Hinrichs[60], Ian Holmes[32], Roger A. Hoskins[2], Melissa J. Hubisz[61], Dan Hultmark[62], Melanie A. Huntley[1], David B. Jaffe[10], Santosh Jagadeeshan[24], William R. Jeck[63], Justin Johnson[57], Corbin D. Jones[63], William C. Jordan[64], Gary H. Karpen[13,65], Eiko Kataoka[66], Peter D. Keightley[56], Pouya Kheradpour[9], Ewen F. Kirkness[57], Leonardo B. Koerich[36], Karsten Kristiansen[67], Dave

Kudrna[68], Rob J. Kulathinal[69], Sudhir Kumar[49,70], Roberta Kwok[8], Eric Lander[10], Charles H. Langley[29], Richard Lapoint[71], Brian P. Lazzaro[72], So-Jeong Lee[68], Lisa Levesque[41], Ruiqiang Li[67,73], Chiao-Feng Lin[20], Michael F. Lin[9,10], Kerstin Lindblad-Toh[10], Ana Llopart[42], Manyuan Long[73], Lloyd Low[27], Elena Lozovsky[69], Jian Lu[23], Meizhong Luo[68], Carlos A. Machado[7], Wojciech Makalowski[20], Mar Marzo[74], Muneo Matsuda[66], Luciano Matzkin[7], Bryant McAllister[42], Carolyn S. McBride[29], Brendan McKernan[8], Kevin McKernan[4], Maria Mendez-Lago[16], Patrick Minx[40], Michael U. Mollenhauer[20], Kristi Montooth[17], Stephen M. Mount[45,75], Xu Mu[20], Eugene Myers[76], Barbara Negre[77], Stuart Newfeld[70], Rasmus Nielsen[78], Mohamed A. F. Noor[52], Patrick O'Grady[71], Lior Pachter[38], Montserrat Papaceit[79], Matthew J. Parisi[4], Michael Parisi[6], Leopold Parts[9], Jakob S. Pedersen[60,79], Graziano Pesole[80], Adam M. Phillippy[45], Chris P. Ponting[58,59], Mihai Pop[45], Damiano Porcelli[26], Jeffrey R. Powell[81], Sonja Prohaska[49,82], Kim Pruitt[83], Marta Puig[74], Hadi Quesneville[84], Kristipati Ravi Ram[1], David Rand[17], Matthew D. Rasmussen[9], Laura K. Reed[53], Robert Reenan[85], Amy Reily[40], Karin A. Remington[57], Tania T. Rieger[86], Michael G. Ritchie[25], Charles Robin[27], Yu-Hui Rogers[57], Claudia Rohde[87], Julio Rozas[79], Marc J. Rubenfield[4], Alfredo Ruiz[74], Susan Russo[51,12], Steven L. Salzberg[45], Alejandro Sanchez-Gracia[19,88], David J. Saranga[4], Hajime Sato[66], Stephen W. Schaeffer[20], Michael C. Schatz[45], Todd Schlenke[89], Russell Schwartz[90], Carmen Segarra[16], Rama S. Singh[24], Laura Sirot[1], Marina Sirota[91], Nicholas B. Sisneros[68], Chris D. Smith[65,92], Temple F. Smith[30], John Spieth[40], Deborah E. Stage[47], Alexander Stark[9,10], Wolfgang Stephan[93], Robert L. Strausberg[57], Sebastian Strempel[93], David Sturgill[6], Granger Sutton[57], Granger G. Sutton[57], Wei Tao[4], Sarah Teichmann[38], Yoshiko N. Tobari[94], Yoshihiko Tomimura[95], Jason M. Tsolas[4], Vera L. S. Valente[51], Eli Venter[57], J. Craig Venter[57], Saverio Vicario[81], Filipe G. Vieira[79], Albert J. Vilella[79,96], Alfredo Villasante[16], Brian Walenz[57], Jun Wang[67,73], Marvin Wasserman[97], Thomas Watts[7], Derek Wilson[98], Richard K. Wilson[40], Rod A. Wing[68], Mariana F. Wolfner[1], Alex Wong[1], Gane Ka-Shu Wong[73,99], Chung-I Wu[23], Gabriel Wu[32], Daisuke Yamamoto[99], Hsiao-Pei Yang[1], Shiaw-Pyng Yang[40], James A. Yorke[100], Kiyohito Yoshida[101], Evgeny Zdobnov[102], Peili Zhang[67,12], Yu Zhang[6], Aleksey V. Zimin[100], Broad Institute Genome Sequencing Platform* & Broad Institute Whole Genome Assembly Team*

[1]Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York 14853, USA. [2]Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA. [3]Center for Integrative Genomics, Department of Molecular and Cell Biology, University of California at Berkeley, Berkeley, California 94720, USA. [4]Agencourt Bioscience Corporation, Beverly, Massachusetts 01915, USA. [5]Faculty of Life Sciences, University of Manchester, Manchester M13 9PT, UK. [6]Laboratory of Cellular and Developmental Biology, National Institutes of Health, Bethesda, Maryland 20892, USA. [7]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721, USA. [8]Department of Biology, Indiana University, Bloomington, Indiana 47405, USA. [9]Computer Science and Artificial Intelligence Laboratory, Cambridge, Massachusetts 02139, USA. [10]Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA. [11]Department of Molecular and Cellular Biology, Harvard University, Cambridge, Massachusetts 02138, USA. [12]FlyBase, The Biological Laboratories, Harvard University, Cambridge, Massachusetts 02138, USA. [13]Department of Molecular and Cell Biology, University of California at Berkeley, Berkeley, California 94720, USA. [14]Biophysics Graduate Group, University of California at Berkeley, Berkeley, California 94720, USA. [15]Field of Ecology and Evolutionary Biology, Cornell University, Ithaca, New York 14853, USA. [16]Centro de Biología Molecular Severo Ochoa, Universidad Autónoma de Madrid, Madrid 28049, Spain. [17]Department of Ecology and Evolutionary Biology, Brown University, Providence, Rhode Island 02912, USA. [18]Structural Studies Division, MRC Laboratory of Molecular Biology, Cambridge CB2 2QH, UK. [19]Departament de Genètica, Universitat de Barcelona, Barcelona 08071, Spain. [20]Department of Biology, Pennsylvania State University, University Park, Pennsylvania 16802, USA. [21]Department of Genetics, University of Georgia, Athens, Georgia 30602, USA. [22]Linnaeus Centre for Bioinformatics, Uppsala Universitet, Uppsala, SE-75124, Sweden. [23]Department of Ecology and Evolution, University of Chicago, Chicago, Illinois 60637, USA. [24]Department of Biology, McMaster University, Hamilton, Ontario, L8S 4K1, Canada. [25]School of Biology, University of St. Andrews, Fife KY16 9TH, UK. [26]Dipartimento di Genetica e Microbiologia dell'Università di Bari, Bari, 70126, Italy. [27]Department of Genetics, University of Melbourne, Melbourne 3010, Australia. [28]Computer Science Department, Stanford University, Stanford, California 94305, USA. [29]Section of Evolution and Ecology and Center for Population Biology, University of California at Davis, Davis, California 95616, USA. [30]BioMolecular Engineering Research Center, Boston University, Boston, Massachusetts 02215, USA. [31]Research Group in Biomedical Informatics, Institut Municipal d'Investigacio Medica, Universitat Pompeu Fabra, Barcelona 08003, Catalonia, Spain. [32]Department of Bioengineering, University of California at Berkeley, Berkeley, California 94720, USA. [33]Laboratory for Computational Genomics, Washington University, St Louis, Missouri 63108, USA. [34]Animal and Plant Sciences, The University of Sheffield, Sheffield S10 2TN, UK. [35]Department of Biology, Syracuse University, Syracuse, New York 13244, USA. [36]Departamento de Genética, Universidade Federal do Rio de Janeiro, Rio de Janeiro 21944-970, Brazil. [37]Tucson Stock Center, Tucson, Arizona 85721, USA. [38]Department of Mathematics, University of California at Berkeley, Berkeley, California 94720, USA. [39]Genome Center, University of California at Davis, Davis, California 95616, USA. [40]Genome Sequencing Center, Washington University School of Medicine, St Louis, Missouri 63108, USA. [41]Department of Biology, University of Winnipeg, Winnipeg, Manitoba R3B 2E9, Canada. [42]Department of Biological Sciences, University of Iowa, Iowa City, Iowa 52242, USA. [43]School of Informatics, Indiana University, Bloomington, Indiana 47405, USA. [44]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. [45]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, Maryland 20742, USA. [46]Department of Biological Sciences, Illinois State University, Normal, Illinois 61790, USA. [47]Department of Biology, University of Rochester, Rochester, New York 14627, USA. [48]Bee Research Lab, USDA-ARS, Beltsville, Maryland 20705, USA. [49]Center for Evolutionary Functional Genomics, Biodesign Institute, Arizona State University, Tempe, Arizona 85287, USA. [50]Department of Computer Science, University of Leipzig, Leipzig 04107, Germany. [51]Departamento de Genética, Universidade Federal do Rio Grande do Sul, Porto Alegre/RS 68011, Brazil. [52]Department of Biology, Duke University, Durham, New Carolina 27708, USA. [53]Department of Genetics, North Carolina State University, Raleigh, North Caroline 27695, USA. [54]Health Center, University of Connecticut, Farmington, Connecticut 06030, USA. [55]Center of Genomic Regulation, Barcelona 8003, Catalonia, Spain. [56]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3JT, UK. [57]J. Craig Venter Institute, Rockville, Maryland 20850, USA. [58]MRC Functional Genetics Unit, University of Oxford, Oxford OX1 3QX, UK. [59]Department of Physiology, Anatomy and Genetics, University of Oxford, Oxford OX1 3QX, UK. [60]Center for Biomolecular Science and Engineering, University of California at Santa Cruz, Santa Cruz, California 95064, USA. [61]Department of Human Genetics, University of Chicago, Chicago, Illinois 60637, USA. [62]Umeå Center for Molecular Pathogenesis, Umeå University, Umeå SE-90187, Sweden. [63]Department of Biology and Carolina Center for Genome Sciences, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. [64]Institute of Zoology, Regent's Park, London NW1 4RY, UK. [65]Drosophila Heterochromatin Genome Project, Department of Genome and Computational Biology, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA. [66]Kyorin University, School of Medicine, Mitaka, Tokyo 181-8611, Japan. [67]Department of Biochemistry and Molecular Biology, University of Southern Denmark, Odense M DK-5230, Denmark. [68]Arizona Genomics Institute, Department of Plant Sciences and BIO5, University of Arizona, Tucson, Arizona 85721, USA. [69]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138, USA. [70]School of Life Sciences, Arizona State University, Tempe, Arizona 85287, USA. [71]Department of Environmental Science, Policy and Management, University of California at Berkeley, Berkeley, California 94720, USA. [72]Department of Entomology, Cornell University, Ithaca, New York 14853, USA. [73]Beijing Genomics Institute at ShenZhen, ShenZhen 518083, China. [74]Departament Genètica i Microbiologia, Universitat Autònoma de Barcelona, Bellaterra 08193, Spain. [75]Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, Maryland 20742, USA. [76]Janelia Farm Research Campus, Howard Hughes Medical Institue, Ashburn, Viginia 20147-2408, USA. [77]Department of Zoology, University of Cambridge, Cambridge, CB2 3EJ, UK. [78]Institue of Biology, University of Copenhagen, DK-2100 Copenhagen Ø, Denmark. [79]Bioinformatics Centre, Department of Molecular Biology, University of Copenhagen, DK-2200 Copenhagen N, Denmark. [80]Dipartimento di Biochimica e Biologia Molecolare, Università di Bari and Istituto Tecnologie Biomediche del Consiglio Nazionale delle Ricerche, Bari 70126, Italy. [81]Department of Ecology and Evolutionary Biology, Yale University, New Haven, Connecticut 06520, USA. [82]Department of Biomedical Informatics, Arizona State University, Tempe, Arizona 85287, USA. [83]National Center for Biotechnology Information, National Institutes of Health, Bethesda, Maryland 20894, USA. [84]Bioinformatics and Genomics Laboratory, Institut Jacques Monod, Paris, 75251, France. [85]Department of Molecular Biology, Cell Biology and Biochemistry, Brown University, Providence, Rhode Island 02912, USA. [86]Departamento de Genética, Centro de Ciências Biológicas, Universidade Federal de Pernambuco, Recife/PE 68011, Brazil. [87]Centro Acadêmico de Vitória, Universidade Federal de Pernambuco, Vitória de Santo Antão/PE, Brazil. [88]Cajal Institute, CSIC, Madrid 28002, Spain. [89]Department of Biology, Emory University, Atlanta, Georgia 30322, USA. [90]Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA. [91]Biomedical Informatics, Stanford University, Stanford, California 94305, USA. [92]Department of Biology, San Francisco State University, San Francisco, California 94132, USA. [93]Department of Biology, University of Munich, 82152 Planegg-Martinsried, Germany. [94]Institute of Evolutionary Biology, Setagaya-ku, Tokyo 158-0098, Japan. [95]Shiba Gakuen, Minato-ku, Tokyo 105-0011, Japan. [96]European Bioinformatics Institute, Hinxton, CB10 1SD, UK. [97]Department of Biology, City University of New York at Queens, Flushing, New York 11367, USA. [98]Department of Biological Sciences and Department of Medicine, University of Alberta, Edmonton, Alberta T6G 2E9, Canada. [99]Department of Developmental Biology and Neurosciences, Tohoku University, Sendai 980-8578, Japan. [100]Institute for Physical Science and Technology, University of Maryland, College Park, Maryland 20742, USA. [101]Hokkaido University, EESBIO, Sapporo, Hokkaido 060-0810, Japan. [102]Faculty of Medecine, Universite de Geneve, Geneva CH-1211, Switzerland.

***Broad Institute Genome Sequencing Platform** Jennifer Baldwin[10], Amr Abdouelleil[10], Jamal Abdulkadir[10], Adal Abebe[10], Brikti Abera[10], Justin Abreu[10], St Christophe Acer[10], Lynne Aftuck[10], Allen Alexander[10], Peter An[10], Erica Anderson[10], Scott Anderson[10], Harindra Arachi[10], Marc Azer[10], Pasang Bachantsang[10], Andrew Barry[10], Tashi Bayul[10], Aaron Berlin[10], Daniel Bessette[10], Toby Bloom[10], Jason Blye[10], Leonid Boguslavskiy[10], Claude Bonnet[10], Boris Boukhgalter[10], Imane Bourzgui[10], Adam Brown[10], Patrick Cahill[10], Sheridon Channer[10], Yama Cheshatsang[10], Lisa Chuda[10], Mieke Citroen[10], Alville Collymore[10], Patrick Cooke[10], Maura Costello[10], Katie D'Aco[10], Riza Daza[10], Georgius De Haan[10], Stuart DeGray[10], Christina DeMaso[10], Norbu Dhargay[10], Kimberly Dooley[10], Erin Dooley[10], Missole Doricent[10], Passang Dorje[10], Kunsang Dorjee[10], Alan Dupes[10], Richard Elong[10], Jill Falk[10], Abderrahim Farina[10], Susan Faro[10], Diallo Ferguson[10], Sheila Fisher[10], Chelsea D. Foley[10], Alicia Franke[10], Dennis Friedrich[10], Loryn Gadbois[10], Gary Gearin[10], Christina R. Gearin[10], Georgia Giannoukos[10], Tina Goode[10], Joseph Graham[10], Edward Grandbois[10], Sharleen Grewal[10], Kunsang Gyaltsen[10], Nabil Hafez[10], Birhane Hagos[10], Jennifer Hall[10], Charlotte Henson[10], Andrew Hollinger[10], Tracey Honan[10], Monika D. Huard[10], Leanne Hughes[10], Brian Hurhula[10], M Erii Husby[10], Asha Kamat[10], Ben Kanga[10],

Seva Kashin[10], Dmitry Khazanovich[10], Peter Kisner[10], Krista Lance[10], Marcia Lara[10], William Lee[10], Niall Lennon[10], Frances Letendre[10], Rosie LeVine[10], Alex Lipovsky[10], Xiaohong Liu[10], Jinlei Liu[10], Shangtao Liu[10], Tashi Lokyitsang[10], Yeshi Lokyitsang[10], Rakela Lubonja[10], Annie Lui[10], Pen MacDonald[10], Vasilia Magnisalis[10], Kebede Maru[10], Charles Matthews[10], William McCusker[10], Susan McDonough[10], Teena Mehta[10], James Meldrim[10], Louis Meneus[10], Oana Mihai[10], Atanas Mihalev[10], Tanya Mihova[10], Rachel Mittelman[10], Valentine Mlenga[10], Anna Montmayeur[10], Leonidas Mulrain[10], Adam Navidi[10], Jerome Naylor[10], Tamrat Negash[10], Thu Nguyen[10], Nga Nguyen[10], Robert Nicol[10], Choe Norbu[10], Nyima Norbu[10], Nathaniel Novod[10], Barry O'Neill[10], Sahal Osman[10], Eva Markiewicz[10], Otero L. Oyono[10], Christopher Patti[10], Pema Phunkhang[10], Fritz Pierre[10], Margaret Priest[10], Sujaa Raghuraman[10], Filip Rege[10], Rebecca Reyes[10], Cecil Rise[10], Peter Rogov[10], Keenan Ross[10], Elizabeth Ryan[10], Sampath Settipalli[10], Terry Shea[10], Ngawang Sherpa[10], Lu Shi[10], Diana Shih[10], Todd Sparrow[10], Jessica Spaulding[10], John Stalker[10], Nicole Stange-Thomann[10], Sharon Stavropoulos[10], Catherine Stone[10], Christopher Strader[10], Senait Tesfaye[10], Talene Thomson[10], Yama Thoulutsang[10], Dawa Thoulutsang[10], Kerri Topham[10], Ira Topping[10], Tsamla Tsamla[10], Helen Vassiliev[10], Andy Vo[10], Tsering Wangchuk[10], Tsering Wangdi[10], Michael Weiand[10], Jane Wilkinson[10], Adam Wilson[10], Shailendra Yadav[10], Geneva Young[10], Qing Yu[10], Lisa Zembek[10], Danni Zhong[10], Andrew Zimmer[10] & Zac Zwirko[10] **Broad Institute Whole Genome Assembly Team** David B. Jaffe[10], Pablo Alvarez[10], Will Brockman[10], Jonathan Butler[10], CheeWhye Chin[10], Sante Gnerre[10], Manfred Grabherr[10], Michael Kleber[10], Evan Mauceli[10] & Iain MacCallum[10]

# VIII.- REFERENCES

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., et al. (2000). The Genome Sequence of *Drosophila* melanogaster. *Science*, 287, 2185-2195.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25, 3389-3402.

Alves, E., Ballesteros, I., Linacero, R. & Vázquez, A.M. (2005). RYS1, a *foldback* transposon, is activated by tissue culture and shows preferential insertion points into the rye genome. *Theor Appl Genet*, 111, 431-436.

Anxolabéhère, D., Kidwell, M.G. & Periquet, G. (1988). Molecular characteristics of diverse populations are consistent with the hypothesis of a recent invasion of *Drosophila* melanogaster by mobile *P-elements*. *Mol. Biol. Evol*, 5, 252-269.

Anxolabehere, D., Nouaud, D. & Periquet, G. (1985). Séquences homologues à l'élément P chez des espèces de *Drosophila* du groupe obscura et chez Scaptomyza pallida (Drosophilidae). *Genet Sel Evol*, 17, 579.

Ashburner, M., Golic, K.G. & Hawley, R.S. (2005). *Drosophila: a laboratory handbook*. Cold Spring Harbor Laboratory Press.

Atkinson, H. & Chalmers, R. (2010). Delivering the goods: viral and non-viral gene therapy systems and the inherent limits on cargo DNA and internal sequences. *Genetica*, 138, 485-498.

Bächli, G. (2007). Taxodros: The Database on taxonomy of Drosophilidae v1.03. *http://taxodros.unizh.ch/*.

Bachmann, A. & Knust, E. (2008). The use of *P-element* transposons to generate transgenic flies. *Methods Mol. Biol*, 420, 61-77.

Badal, M., Portela, A., Baldrich, E., Marcos, R., Cabré, O. & Xamena, N. (2006a). An *FB*-NOF mediated duplication of the *white* gene is responsible for the zeste 1 phenotype in some *Drosophila* melanogaster unstable strains. *Mol Genet Genomics*, 275, 35-43.

Badal, M., Portela, A., Xamena, N. & Cabré, O. (2006b). Molecular and bioinformatic analysis of the *FB*-NOF transposable element. *Gene*, 371, 130-135.

Bailey, J.A. & Eichler, E.E. (2006). Primate segmental duplications: crucibles of evolution, diversity and disease. *Nat Rev Genet*, 7, 552-564.

Barker, J.S.F. & Starmer, W.T. (1982). Population genetics of Opuntia breeding *Drosophila* in Australia. In: *Ecological Genetics and Evolution. The Cactus-Yeast-Drosophila Model System* (eds. Barker, J.S.F. & Starmer, W.T.). Academic Press, Sydney, pp. 209-224.

Barker, J.S.F., Starmer, W.T. & MacIntyre, R.J. (1990). *Ecological and evolutionary genetics of Drosophila*. Plenum Press.

Barraclough, T.G. & Nee, S. (2001). Phylogenetics and speciation. *Trends in Ecology & Evolution*, 16, 391-399.

Bartolomé, C., Maside, X. & Charlesworth, B. (2002). On the Abundance and Distribution of Transposable Elements in the Genome of *Drosophila* melanogaster. *Molecular Biology and Evolution*, 19, 926 -937.

Beall & Rio. (1997). *Drosophila P-element* transposase is a novel site-specific endonuclease. *Genes Dev*, 11, 2137-2151.

Beare, P.A., Unsworth, N., Andoh, M., Voth, D.E., Omsland, A., Gilk, S.D., et al. (2009). Comparative Genomics Reveal Extensive Transposon-Mediated Genomic Plasticity and Diversity among Potential Effector Proteins within the Genus Coxiella. *Infect. Immun.*, 77, 642-656.

Bessière, D., Lacroix, C., Campagne, S., Ecochard, V., Guillet, V., Mourey, L., et al. (2008). Structure-function analysis of the THAP zinc finger of THAP1, a large C2CH DNA-binding module linked to Rb/E2F pathways. *J. Biol. Chem*, 283, 4352-4363.

Betrán, E., Quezada-Díaz, J.E., Ruiz, A., Santos, M. & Fontdevila, A. (1995). The evolutionary history of *Drosophila buzzatii*. XXXII. Linkage disequilibrium between allozymes and chromosome inversions in two colonizing populations. *Heredity*, 74 ( Pt 2), 188-199.

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.

Bingham, P.M., Kidwell, M.G. & Rubin, G.M. (1982). The molecular basis of P-M hybrid dysgenesis: the role of the P-element, a P-strain-specific transposon family. *Cell*, 29, 995-1004.

Bingham, P.M. & Zachar, Z. (1985). Evidence that two mutations, wDZL and z1, affecting synapsis-dependent genetic behavior of *white* are transcriptional regulatory mutations. *Cell*, 40, 819-825.

Boake, C.R.B. (2005). Sexual Selection and Speciation in Hawaiian Drosophila. *Behav Genet*, 35, 297-303.

Bourne, H.R., Sanders, D.A. & McCormick, F. (1991). The GTPase superfamily: conserved structure and molecular mechanism. *Nature*, 349, 117-127.

Bowen, N.J., Jordan, I.K., Epstein, J.A., Wood, V. & Levin, H.L. (2003). Retrotransposons and Their Recognition of pol II Promoters: A Comprehensive Survey of the Transposable Elements From the Complete Genome Sequence of Schizosaccharomyces pombe. *Genome Research*, 13, 1984 -1997.

Britten, R.J. & Kohne, D.E. (1968). Repeated sequences in DNA. Hundreds of thousands of

copies of DNA sequences have been incorporated into the genomes of higher organisms. *Science*, 161, 529-540.

Brookfield, J.F., Montgomery, E. & Langley, C.H. (1984). Apparent absence of transposable elements related to the *P-elements* of D. melanogaster in other species of Drosophila. *Nature*, 310, 330-332.

Brookfield, J.F.Y. (1982). Interspersed repetitive DNA sequences are unlikely to be parasitic. *Journal of Theoretical Biology*, 94, 281-299.

Brunet, F., Giraud, T., Godin, F. & Capy, P. (2002). Do Deletions of Mos1 -Like Elements Occur Randomly in the Drosophilidae Family? *Journal of Molecular Evolution*, 54, 227-234.

Cáceres, M., National Institutes of Health Intramural Sequencing Center Comparative Sequencing Program, Sullivan, R.T. & Thomas, J.W. (2007). A recurrent inversion on the eutherian X chromosome. *Proceedings of the National Academy of Sciences*, 104, 18571 -18576.

Cáceres, M., Puig, M. & Ruiz, A. (2001). Molecular Characterization of Two Natural Hotspots in the *Drosophila buzzatii* Genome Induced by Transposon Insertions. *Genome Research*, 11, 1353-1364.

Cáceres, M., Ranz, J.M., Barbadilla, A., Long, M. & Ruiz, A. (1999). Generation of a Widespread *Drosophila* Inversion by a Transposable Element. *Science*, 285, 415-418.

Caletka, B.C. & McAllister, B.F. (2004). A genealogical view of chromosomal evolution and species delimitation in the *Drosophila* virilis species subgroup. *Molecular Phylogenetics and Evolution*, 33, 664-670.

Campagne, S., Saurel, O., Gervais, V. & Milon, A. (2010). Structural determinants of specific DNA-recognition by the THAP zinc finger. *Nucleic Acids Res*, 38, 3466-3476.

Cappello, J. (1985). Sequence of Dictyostelium DIRS-1: An apparent retrotransposon with inverted terminal repeats and an internal circle junction sequence. *Cell*, 43, 105-115.

Capy, P. (1998). *Dynamics and evolution of transposable elements*. Springer.

Capy, P., Vitalis, R., Langin, T., Higuet, D. & Bazin, C. (1996). Relationships between transposable elements based upon the integrase-transposase domains: Is there a common ancestor? *Journal of Molecular Evolution*, 42, 359-368.

Casacuberta, E. & Pardue, M.-L. (2005). *HeT-A* and *TART*, two *Drosophila* retrotransposons with a bona fide role in chromosome structure for more than 60 million years. *Cytogenetic and Genome Research*, 110, 152-159.

Casals, F., Caceres, M., Manfrin, M.H., Gonzalez, J. & Ruiz, A. (2005). Molecular Characterization and Chromosomal Distribution of Galileo, Kepler and Newton, Three

*Foldback* Transposable Elements of the *Drosophila buzzatii* Species Complex. *Genetics*, 169, 2047-2059.

Casals, F., Caceres, 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.

Castresana, J. (2000). Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. *Molecular Biology and Evolution*, 17, 540 -552.

Chao, K. & Miller, W. (1995). Linear-space algorithms that build local alignments from fragments. *Algorithmica*, 13, 106-134.

Charlesworth, B. & Charlesworth, D. (1983). The Population Dynamics of Transposable Elements. *Genetics Research*, 42, 1-27.

Charlesworth, B. & Langley, C.H. (1991). Population genetics of Transposable Elements in Drosophila. In: *Evolution at the molecular level* (eds. Selander, R.K., Clark, A.G. & Whittam, T.S.). Sinauer Associates, pp. 150-176.

Charlesworth, B., Lapid, A. & Canada, D. (1992). The distribution of transposable elements within and between chromosomes in a population of *Drosophila* melanogaster. II. Inferences on the nature of selection against elements. *Genet. Res*, 60, 115-130.

Cheng, C., Tsuchimoto, S., Ohtsubo, H. & Ohtsubo, E. (2000). Tnr8, a *foldback* transposable element from rice. *Genes Genet. Syst*, 75, 327-333.

Clark, A.G., Eisen, M.B., Smith, D.R., Bergman, C.M., Oliver, B., Markow, T.A., et al. (2007). Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature*, 450, 203-18.

Clark, A.G., Gibson, G., Kaufman, T.C., McAllister, B., Myers, E.W. & O'Grady, P.M. (2003). Proposal for *Drosophila* as a Model System for Comparative Genomics. *http://flybase.org/static_pages/news/wpapers.html*.

Clark, J.B. & Kidwell, M.G. (1997). A phylogenetic perspective on P transposable element evolution in Drosophila. *Proc. Natl. Acad. Sci. U.S.A*, 94, 11428-11433.

Clouaire, T., Roussigne, M., Ecochard, V., Mathe, C., Amalric, F. & Girard, J.-P. (2005). The THAP domain of THAP1 is a large C2CH module with zinc-dependent sequence-specific DNA-binding activity. *Proceedings of the National Academy of Sciences of the United States of America*, 102, 6907-6912.

Collins, J., Forbes, E. & Anderson, P. (1989). The Tc3 Family of Transposable Genetic Elements in Caenorhabditis elegans. *Genetics*, 121, 47 -55.

Collins, M. & Rubin, G.M. (1983). High-frequency precise excision of the *Drosophila foldback* transposable element. *Nature*, 303, 259-260.

Collins, M. & Rubin, G.M. (1984). Structure of chromosomal rearrangements induced by the *FB* transposable element in Drosophila. *Nature*, 308, 323-327.

Cordaux, R. (2009). Gene conversion maintains nonfunctional transposable elements in an obligate mutualistic endosymbiont. *Mol. Biol. Evol*, 26, 1679-1682.

Cordaux, R., Udit, S., Batzer, M.A. & Feschotte, C. (2006). Birth of a chimeric primate gene by capture of the transposase gene from a mobile element, 103, 8101-8106.

Craddock, E. (2000). Speciation processes in the adaptive radiation of Hawaiianplants and animals. *Evolutionary Biology*, 31, 1-53.

Craig, N.L., Craigie, R., Gellert, M. & Lambowitz , Alan M. (Eds.). (2002). *Mobile DNA II*. ASM Press.

Cui, Z., Geurts, A.M., Liu, G., Kaufman, C.D. & Hackett, P.B. (2002). Structure-Function Analysis of the Inverted Terminal Repeats of the Sleeping Beauty Transposon. *Journal of Molecular Biology*, 318, 1221-1235.

Daniels, S.B., Clark, S.H., Kidwell, M.G. & Chovnick, A. (1987). Genetic transformation of *Drosophila* melanogaster with an autonomous P-element: phenotypic and molecular analyses of long-established transformed lines. *Genetics*, 115, 711-723.

Daniels, S.B., Peterson, K.R., Strausbaugh, L.D., Kidwell, M.G. & Chovnick, A. (1990). Evidence for horizontal transmission of the P transposable element between *Drosophila* species. *Genetics*, 124, 339-355.

Daniels, S.B. & Strausbaugh, L.D. (1986). The distribution of *P-element* sequences in Drosophila: the willistoni and saltans species groups. *J. Mol. Evol*, 23, 138-148.

Daskalova, S.M., Scott, N.W. & Elliott, M.C. (2005). Folbos, a new *foldback* element in rice. *Genes Genet. Syst*, 80, 141-145.

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.

Doolittle, W.F. & Sapienza, C. (1980). Selfish genes, the phenotype paradigm and genome evolution. *Nature*, 284, 601-603.

Doudna, J.A. & Cech, T.R. (2002). The chemical repertoire of natural ribozymes. *Nature*, 418, 222-228.

Dray, T. & Gloor, G.B. (1997). Homology Requirements for Targeting Heterologous Sequences During P-Induced Gap Repair in *Drosophila* melanogaster. *Genetics*, 147, 689 -699.

*Drosophila* 12 Genomes Consortium, Clark, A.G., Eisen, M.B., Smith, D.R., Bergman, C.M.,

Oliver, B., et al. (2007). Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature*, 450, 203-18.

Drummond, A., Ashton, B., Cheung, M., Heled, J., Kearse, M., Moir, R., et al. (2010). *Geneious Pro*. Geneious. Biomatters Ltd.

Drummond, A., Ho, S.Y.W., Phillips, M.J. & Rambaut, A. (2006). Relaxed Phylogenetics and Dating with Confidence. *PLoS Biol*, 4, e88.

Drummond, A. & Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, 7, 214.

Durando, C.M., Baker, R.H., Etges, W.J., Heed, W.B., Wasserman, M. & DeSalle, R. (2000). Phylogenetic Analysis of the *repleta* Species Group of the Genus *Drosophila* Using Multiple Sources of Characters. *Molecular Phylogenetics and Evolution*, 16, 296-307.

Dynan, W.S. & Yoo, S. (1998). Interaction of Ku protein and DNA-dependent protein kinase catalytic subunit with nucleic acids. *Nucleic Acids Research*, 26, 1551 -1559.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucl. Acids Res.*, 32, 1792-1797.

Edgar, R.C. & Myers, E.W. (2005). PILER: identification and classification of genomic repeats. *Bioinformatics*, 21, i152-i158.

Edwards, K.A., Doescher, L.T., Kaneshiro, K.Y. & Yamamoto, D. (2007). A Database of Wing Diversity in the Hawaiian Drosophila. *PLoS ONE*, 2, e487.

Engels, W.R. (1979). Hybrid dysgenesis in *Drosophila* melanogaster: rules of inheritance of female sterility. *Genet. Res.*, 33, 219.

Engels, W.R. (1996). *P-elements* in Drosophila. *Curr. Top. Microbiol. Immunol*, 204, 103-123.

Engels, W.R., Johnson-Schlitz, D.M., Eggleston, W.B. & Sved, J. (1990). High-frequency *P-element* loss in *Drosophila* is homolog dependent. *Cell*, 62, 515-525.

Engels, W.R. & Preston, C.R. (1984). Formation of chromosome rearrangements by P factors in Drosophila. *Genetics*, 107, 657-678.

Etges, W., Johnson, W., Duncan, G., Huckins, G. & Heed, W. (1999). Ecological genetics of cactophilic Drosophila. In: *Ecology of Sonoran Desert Plants and Plant Communities* (ed. Robichaux, R.). University of Arizona Press, Tucson, pp. 164-214.

Fernandez Iriarte, P.J., Norry, F.M. & Hasson, E.R. (2003). Chromosomal inversions effect body size and shape in different breeding resources in *Drosophila buzzatii*. *Heredity*, 91, 51-59.

Feschotte, C., Jiang, N. & Wessler, S.R. (2002). Plant transposable elements: where genetics meets genomics. *Nat Rev Genet*, 3, 329-341.

Feschotte, C. & Pritham, E.J. (2007). DNA Transposons and the Evolution of Eukaryotic

Genomes. *Annu. Rev. Genet.*, 41, 331-68.

Feschotte, C., Swamy, L. & Wessler, S.R. (2003). Genome-Wide Analysis of mariner-Like Transposable Elements in Rice Reveals Complex Relationships With Stowaway Miniature Inverted Repeat Transposable Elements (MITEs). *Genetics*, 163, 747 -758.

Finnegan, D.J. (1989). Eukaryotic transposable elements and genome evolution. *Trends Genet*, 5, 103-107.

Fiston-Lavier, A.-S., Anxolabehere, D. & Quesneville, H. (2007). A model of segmental duplication formation in *Drosophila* melanogaster. *Genome Research*, 17, 1458 -1470.

Fontdevila, A., Ruiz, A., Alonso, G. & Ocaña, J. (1981). Evolutionary History of *Drosophila buzzatii*. I. Natural Chromosomal Polymorphism in Colonized Populations of the Old World. *Evolution*, 35, 148-157.

Fontdevila, A., Ruiz, A., Ocaña, J. & Alonso, G. (1982). Evolutionary History of *Drosophila buzzatii*. II. How Much Has Chromosomal Polymorphism Changed in Colonization? *Evolution*, 36, 843-851.

Formosa, T. & Alberts, B.M. (1986). DNA synthesis dependent on genetic recombination: characterization of a reaction catalyzed by purified bacteriophage T4 proteins. *Cell*, 47, 793-806.

Franz, G. & Savakis, C. (1991). Minos, a new transposable element from *Drosophila* hydei, is a member of the Tc1-like family of transposons. *Nucleic Acids Res*, 19, 6646.

Gardner, M.J., Hall, N., Fung, E., *White*, O., Berriman, M., Hyman, R.W., et al. (2002). Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature*, 419, 498-511.

Gloor, G., Nassif, N., Johnson-Schlitz, D., Preston, C. & Engels, W. (1991). Targeted gene replacement in *Drosophila* via P-element-induced gap repair. *Science*, 253, 1110-1117.

Gonzalez, J. & Petrov, D. (2009). MITEs--The Ultimate Parasites. *Science*, 325, 1352-1353.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W. & Gascuel, O. (2010). New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology*, 59, 307 -321.

Guindon, S. & Gascuel, O. (2003). A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. *Systematic Biology*, 52, 696 -704.

Hagemann, S. & Pinsker, W. (2001). *Drosophila* P Transposons in the Human Genome? *Molecular Biology and Evolution*, 18, 1979 -1982.

Hammer, S.E., Strehl, S. & Hagemann, S. (2005). Homologs of *Drosophila* P Transposons Were Mobile in Zebrafish but Have Been Domesticated in a Common Ancestor of Chicken and Human. *Molecular Biology and Evolution*, 22, 833 -844.

Hankeln, T. & Schmidt, E.R. (1990). New *foldback* transposable element T*FB*1 found in histone genes of the midge Chironomus thummi. *J. Mol. Biol*, 215, 477-482.

Han, Y. & Wessler, S.R. (2010). MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Research*.

Harden, N. & Ashburner, M. (1990). Characterization of the *FB*-NOF Transposable Element of *Drosophila* melanogaster. *Genetics*, 126, 387-400.

Hartl, D.L., Lohe, A.R. & Lozovskaya, E.R. (1997). Modern thoughts on an ancyent marinere: function, evolution, regulation. *Annu. Rev. Genet*, 31, 337-358.

Hasegawa, M., Kishino, H. & Yano, T.-aki. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol*, 22, 160-174.

Hasson, E., Naveira, H. & Fontdevila, A. (1992). The breeding sites of Argentiniancactophilic species of the *Drosophila mulleri* complex. *Rev. Chilena de Historia Natural*, 65, 319-326.

Hasson, E., Rodriguez, C., Fanara, J.J., Naveira, H., Reig, O.A. & Fontdevila, A. (1995). Macrogeographic patterns in the inversion polymorphisms of *Drosophila buzzatii* in New World populations. *Journal of Evolutionary Biology*, 8, 369-384.

Hasson, E., Vilardi, J.C., Naveira, H., Fanara, J.J., Rodriguez, C., Reig, O.A., et al. (1991). The evolutionary history of *Drosophila buzzatii*. XVI. Fitness component analysis in an original natural population from Argentina. *Journal of Evolutionary Biology*, 4, 209-225.

Hastings, P.J. (1988). Recombination in the eukaryotic nucleus. *BioEssays*, 9, 61-64.

Haynes, K.A., Caudy, A.A., Collins, L. & Elgin, S.C.R. (2006). Element 1360 and RNAi Components Contribute to HP1-Dependent Silencing of a Pericentric Reporter. *Current Biology*, 16, 2222-2227.

Hickey, D.A. (1982). Selfish DNA: a sexually-transmitted nuclear parasite. *Genetics*, 101, 519 -531.

Hickman, A.B., Chandler, M. & Dyda, F. (2010). Integrating prokaryotes and eukaryotes: DNA transposases in light of structure. *Critical Reviews in Biochemistry and Molecular Biology*, 45, 50-69.

Hickman, A.B., Perez, Z.N., Zhou, L., Musingarimi, P., Ghirlando, R., Hinshaw, J.E., et al. (2005). Molecular architecture of a eukaryotic DNA transposase. *Nat Struct Mol Biol*, 12, 715-721.

Hoffman-Lieberman, B.D., Lieberman, D. & Cohen, S.N. (1989). TU elements and puppy sequences. In: *Mobile DNA* (eds. Berg, D.E. & Howe, M.M.). American Society for Microbiology, Washington DC, pp. 593-617.

Houck, M.A., Clark, J.B., Peterson, K.R. & Kidwell, M.G. (1991). Possible horizontal transfer of *Drosophila* genes by the mite Proctolaelaps regalis. *Science*, 253, 1125-1128.

Hsia, A.P. & Schnable, P.S. (1996). DNA sequence analyses support the role of interrupted gap repair in the origin of internal deletions of the maize transposon, MuDR. *Genetics*, 142, 603-618.

Hua-Van, A., Le Rouzic, A., Maisonhaute, C. & Capy, P. (2005). Abundance, distribution and dynamics of retrotransposable elements and transposons: similarities and differences. *Cytogenet Genome Res*, 110, 426-440.

Ivics, Hackett, P.B., Plasterk, R.H. & Izsvák, Z. (1997). Molecular Reconstruction of Sleeping Beauty, a Tc1-like Transposon from Fish, and Its Transposition in Human Cells. *Cell*, 91, 501-510.

Jurka, Kapitonov, V.V., Kohany, O. & Jurka, M.V. (2007). Repetitive sequences in complex genomes: structure and evolution. *Annu Rev Genomics Hum Genet*, 8, 241-259.

Kahlon, A.S., Hice, R.H., O'Brochta, D.A. & Atkinson, P.W. (2011). DNA binding activities of the Herves transposase from the mosquito Anopheles gambiae. *Mobile DNA*, 2, 9.

Kaminker, J.S., Bergman, C.M., Kronmiller, B., Carlson, J., Svirskas, R., Patel, S., et al. (2002). The transposable elements of the *Drosophila* melanogaster euchromatin: a genomics perspective. *Genome Biol*, 3, RESEARCH0084.

Kapitonov, V.V. & Jurka, J. (2003). Molecular paleontology of transposable elements in the *Drosophila* melanogaster genome. *Proceedings of the National Academy of Sciences of the United States of America*, 100, 6569 -6574.

Katoh, K., Misawa, K., Kuma, K. & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, 30, 3059 -3066.

Kaufman, P.D. & Rio, D.C. (1992). *P-element* transposition in vitro proceeds by a cut-and-paste mechanism and uses GTP as a cofactor. *Cell*, 69, 27-39.

Kazazian, H.H. (2004). Mobile Elements: Drivers of Genome Evolution. *Science*, 303, 1626 -1632.

Kholodilov, N.G., Bolshakov, V.N., Blinov, V.M., Solovyov, V.V. & Zhimulev, I.F. (1988). Intercalary heterochromatin in Drosophila. III. Homology between DNA sequences from the Y chromosome, bases of polytene chromosome limbs, and chromosome 4 of D. melanogaster. *Chromosoma*, 97, 247-253.

Kidwell, M.G. (1977). Reciprocal differences in female recombination associated with hybrid dysgenesis in *Drosophila* melanogaster. *Genet. Res*, 30, 77-88.

Kidwell, M.G. (1985). Hybrid dysgenesis in *Drosophila* melanogaster: nature and inheritance of

*P-element* regulation. *Genetics*, 111, 337-350.

Kidwell, M.G. (1992). Horizontal transfer. *Current Opinion in Genetics & Development*, 2, 868-873.

Kidwell, M.G. (2002). Transposable elements and the evolution of genome size in eukaryotes. *Genetica*, 115, 49-63.

Kidwell, M.G., Kidwell, J.F. & Nei, M. (1973). A case of high rate of spontaneous mutation affecting viability in *Drosophila* melanogaster. *Genetics*, 75, 133-153.

Kidwell, M.G., Kidwell, J.F. & Sved, J.A. (1977). Hybrid dysgenesis in *Drosophila* melanogaster: a syndrome of aberrant traits including mutation, sterility and male recombination. *Genetics*, 86, 813-833.

Kidwell, M.G. & Lisch, D.R. (2002). Transposable Elements as sources of genomic variation. In: *Mobile DNA II* (eds. Craig, N., Craigie, R., Gellert, M. & Lambowitz , Alan M.). ASM Press, Washington DC, pp. 485-515.

Kidwell, M.G. & Novy, J.B. (1979). Hybrid dysgenesis in *Drosophila* melanogaster: sterility resulting from gonadal dysgenesis in the P-M system. *Genetics*, 92, 1127-1140.

Kimbacher, S., Gerstl, I., Velimirov, B. & Hagemann, S. (2009). *Drosophila* P transposons of the urochordata Ciona intestinalis. *Mol Genet Genomics*, 282, 165-172.

Kondo, T., Inagaki, S., Yasuda, K. & Kageyama, Y. (2006). Rapid construction of *Drosophila* RNAi transgenes using pRISE, a P-element-mediated transformation vector exploiting an in vitro recombination system. *Genes Genet. Syst*, 81, 129-134.

Kramerov, D.A. & Vassetzky, N.S. (2005). Short Retroposons in Eukaryotic Genomes. In: *A Survey of Cell Biology*. Academic Press, pp. 165-221.

Kumar, A. & Bennetzen, J.L. (1999). Plant retrotransposons. *Annu. Rev. Genet*, 33, 479-532.

Kurenova, E.V., Leĭbovich, B.A., Bass, I.A., Bebikhov, D.V., Pavlova, M.N. & Danilevskaia, O.N. (1990). [Hoppel-family of mobile elements of *Drosophila* melanogaster, flanked by short inverted repeats and having preferential localization in the heterochromatin regions of the genome]. *Genetika*, 26, 1701-1712.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409, 860-921.

Landschulz, W.H., Johnson, P.F. & McKnight, S.L. (1988). The leucine zipper: a hypothetical structure common to a new class of DNA binding proteins. *Science*, 240, 1759-1764.

Langley, C.H., Montgomery, E., Hudson, R., Kaplan, N. & Charlesworth, B. (1988). On the Role of Unequal Exchange in the Containment of Transposable Element Copy Number. *Genetics Research*, 52, 223-235.

Lansman, R.A., Stacey, S.N., Grigliatti, T.A. & Brock, H.W. (1985). Sequences homologous to

the P mobile element of *Drosophila* melanogaster are widely distributed in the subgenus *Sophophora*. *Nature*, 318, 561-563.

Lee, C.C., Beall, E.L. & Rio, D.C. (1998). DNA binding by the KP repressor protein inhibits *P-element* transposase activity in vitro. *EMBO J*, 17, 4166-4174.

Lee, J., Han, K., Meyer, T.J., Kim, H.-S. & Batzer, M.A. (2008). Chromosomal Inversions between Human and Chimpanzee Lineages Caused by Retrotransposons. *PLoS ONE*, 3, e4047.

Lee, M., Gippert, G., Soman, K., Case, D. & Wright, P. (1989). Three-dimensional solution structure of a single zinc finger DNA-binding domain. *Science*, 245, 635-637.

Lee, S.H., Clark, J.B. & Kidwell, M.G. (1999). A P-element-homologous sequence in the house fly, Musca domestica. *Insect Mol. Biol*, 8, 491-500.

Leonardo, T.E. & Nuzhdin, S.V. (2002). Intracellular Battlegrounds: Conflict and Cooperation Between Transposable Elements. *Genetics Research*, 80, 155-161.

Liao, G.C., Rehm, E.J. & Rubin, G.M. (2000). Insertion site preferences of the P transposable element in *Drosophila* melanogaster. *Proc. Natl. Acad. Sci. U.S.A*, 97, 3347-3351.

Lim, J.K. & Simmons, M.J. (1994). Gross chromosome rearrangements mediated by transposable elements in *Drosophila* melanogaster. *BioEssays*, 16, 269-275.

van der Linde, K., Houle, D., Spicer, G.S. & Steppan, S.J. (2010). A supermatrix-based molecular phylogeny of the family Drosophilidae. *Genet Res (Camb)*, 92, 25-38.

Ling, A. & Cordaux, R. (2010). Insertion Sequence Inversions Mediated by Ectopic Recombination between Terminal Inverted Repeats. *PLoS ONE*, 5, e15654.

Linheiro, R.S. & Bergman, C.M. (2008). Testing the palindromic target site model for DNA transposon insertion using the *Drosophila* melanogaster P-element. *Nucl. Acids Res.*, 36, 6199-6208.

Li, W.-H. (1997). *Molecular evolution*. Sinauer Associates.

Loreto, E.L.S., Carareto, C.M.A. & Capy, P. (2008). Revisiting horizontal transfer of transposable elements in Drosophila. *Heredity*, 100, 545-554.

Lupas, A., Van Dyke, M. & Stock, J. (1991). Predicting coiled coils from protein sequences. *Science*, 252, 1162-1164.

Machado, C.A., Matzkin, L.M., Reed, L.K. & Markow, T.A. (2007). Multilocus nuclear sequences reveal intra- and interspecific relationships among chromosomally polymorphic species of cactophilic Drosophila. *Molecular Ecology*, 16, 3009-3024.

Marchler-Bauer, A., Anderson, J.B., Cherukuri, P.F., DeWeese-Scott, C., Geer, L.Y., Gwadz, M., et al. (2005). CDD: a Conserved Domain Database for protein classification. *Nucleic Acids Res*, 33, D192-196.

Markow, T.A. & O'Grady, P.M. (2006). *Drosophila: a guide to species identification and use*. Academic Press.

Markow, T.A. & O'Grady, P.M. (2007). *Drosophila* Biology in the Genomic Age. *Genetics*, 177, 1269 -1276.

Marquez, C.P. & Pritham, E.J. (2010). Phantom, a New Subclass of Mutator DNA Transposons Found in Insect Viruses and Widely Distributed in Animals. *Genetics*, genetics.110.116673.

Marzo, M., Liu, D., Ruiz, A. & Chalmers, R. (2011). DNA-binding properties of THAP-containing *Galileo* transposase. *In preparation*.

Marzo, M., Puig, M. & Ruiz, A. (2008). The *Foldback*-like element *Galileo* belongs to the P superfamily of DNA transposons and is widespread within the *Drosophila* genus. *Proc Natl Acad Sci U S A*, 105, 2957-62.

Matzkin, L.M. & Markow, T.A. (2009). Transcriptional Regulation of Metabolism Associated With the Increased Desiccation Resistance of the Cactophilic *Drosophila* mojavensis. *Genetics*, 182, 1279 -1288.

McClintock, B. (1950). The origin and behavior of mutable loci in maize. *Proc. Natl. Acad. Sci. U.S.A*, 36, 344-355.

McClintock, B. (1951). Chromosome organization and genic expression. *Cold Spring Harb. Symp. Quant. Biol*, 16, 13-47.

Miskey, C., Izsvák, Z., Plasterk, R.H. & Ivics, Z. (2003). The Frog Prince: a reconstructed transposon from Rana pipiens with high transpositional activity in vertebrate cells. *Nucleic Acids Research*, 31, 6873 -6881.

Miskey, C., Papp, B., Mates, L., Sinzelle, L., Keller, H., Izsvak, Z., et al. (2007). The Ancient mariner Sails Again: Transposition of the Human Hsmar1 Element by a Reconstructed Transposase and Activities of the SETMAR Protein on Transposon Ends. *Mol. Cell. Biol.*, 27, 4589-4600.

Misra, S., Buratowski, R.M., Ohkawa, T. & Rio, D.C. (1993). Cytotype Control of *Drosophila* melanogaster *P-element* Transposition: Genomic Position Determines Maternal Repression. *Genetics*, 135, 785 -800.

Montgomery, E., Charlesworth, B. & Langley, C.H. (1987). A test for the role of natural selection in the stabilization of transposable element copy number in a population of *Drosophila* melanogaster. *Genet. Res*, 49, 31-41.

Morgante, M. (2006). Plant genome organisation and diversity: the year of the junk! *Curr. Opin. Biotechnol*, 17, 168-173.

Moschetti, R., Chlamydas, S., Massimiliano Marsano, R. & Caizzi, R. (2008). Conserved motifs

and dynamic aspects of the terminal inverted repeat organization within Bari -like transposons. *Molecular Genetics and Genomics*, 279, 451-461.

Moschetti, R., Marsano, R.M., Caggese, C., Caizzi, R. & Barsanti, P. (2004). *FB* elements can promote exon shuffling: a promoter-less *white* allele can be reactivated by *FB* mediated transposition in *Drosophila* melanogaster. *Molecular Genetics and Genomics*, 271, 394-401.

Newman, T. & Trask, B.J. (2003). Complex Evolution of 7E Olfactory Receptor Genes in Segmental Duplications. *Genome Research*, 13, 781 -793.

Nowotny, M. (2009). Retroviral integrase superfamily: the structural perspective. *EMBO Rep*, 10, 144-151.

O'Grady, P.M. & Markow, T.A. (2009). Phylogenetic taxonomy in Drosophila: Problems and prospects. *fly*, 3, 10-14.

O'Hare, K. & Rubin, G.M. (1983). Structures of P transposable elements and their sites of insertion and excision in the *Drosophila* melanogaster genome. *Cell*, 34, 25-35.

Oliveira, D.C.S.G., Almeida, F.C., O'Grady, P., Etges, W., Armella, M.A. & DeSalle, R. (2011). A molecular phylogenetic hypothesis, divergence times, host use, and comments on the evolutionary history of the *Drosophila repleta* species group. *In preparation*.

Oliveira de Carvalho, M., Silva, J.C. & Loreto, E.L.S. (2004). Analyses of P-like transposable element sequences from the genome of Anopheles gambiae. *Insect Mol. Biol*, 13, 55-63.

Oliver, K.R. & Greene, W.K. (2009). Transposable elements: powerful facilitators of evolution. *Bioessays*, 31, 703-714.

Oliver, K.R. & Greene, W.K. (2011). Mobile DNA and the TE-Thrust Hypothesis: Supporting Evidence from the Primates. *Mobile DNA*, 2, 8.

Orgel, L.E. & Crick, F.H. (1980). Selfish DNA: the ultimate parasite. *Nature*, 284, 604-607.

Ostertag, E.M. & Kazazian, H.H., Jr. (2001). Biology of mammalian L1 retrotransposons. *Annu. Rev. Genet*, 35, 501-538.

Pardue, M.-L. & DeBaryshe, P.G. (2011). Retrotransposons that maintain chromosome ends. *Proceedings of the National Academy of Sciences*.

Pardue, M.-L., Rashkova, S., Casacuberta, E., DeBaryshe, P.G., George, J.A. & Traverse, K.L. (2005). Two retrotransposons maintain telomeres in Drosophila. *Chromosome Research*, 13, 443-453.

Paro, R., Goldberg, M.L. & Gehring, W.J. (1983). Molecular analysis of large transposable elements carrying the *white* locus of *Drosophila* melanogaster. *EMBO J*, 2, 853-860.

Pavletich, N. & Pabo, C. (1991). Zinc finger-DNA recognition: crystal structure of a Zif268-

DNA complex at 2.1 A. *Science*, 252, 809-817.

Perkins, H.D. & Howells, A.J. (1992). Genomic sequences with homology to the *P-element* of *Drosophila* melanogaster occur in the blowfly Lucilia cuprina. *Proceedings of the National Academy of Sciences of the United States of America*, 89, 10753 -10757.

Petrov, D.A. (2002). DNA loss and evolution of genome size in Drosophila. *Genetica*, 115, 81-91.

Petrov, D.A., Aminetzach, Y.T., Davis, J.C., Bensasson, D. & Hirsh, A.E. (2003). Size Matters: Non-LTR Retrotransposable Elements and Ectopic Recombination in Drosophila. *Molecular Biology and Evolution*, 20, 880 -892.

Petrov, D.A., Fiston-Lavier, A.-S., Lipatov, M., Lenkov, K. & González, J. (2010). Population genomics of transposable elements in *Drosophila* melanogaster. *Molecular Biology and Evolution*.

Petrov, D.A. & Hartl, D.L. (1998). High rate of DNA loss in the *Drosophila* melanogaster and *Drosophila* virilis species groups. *Molecular Biology and Evolution*, 15, 293 -302.

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.

Pinsker, W., Haring, E., Hagemann, S. & Miller, W.J. (2001). The evolutionary life history of P transposons: from horizontal invaders to domesticated neogenes. *Chromosoma*, 110, 148-158.

Plasterk, R.H. (1991). The origin of footprints of the Tc1 transposon of Caenorhabditis elegans. *EMBO J.*, 10, 1919-1925.

Popadić, A. & Anderson, W.W. (1994). The history of a genetic system. *Proceedings of the National Academy of Sciences*, 91, 6819 -6823.

Posada, D. (2008). jModelTest: Phylogenetic Model Averaging. *Molecular Biology and Evolution*, 25, 1253 -1256.

Potter, S., Truett, M., Phillips, M. & Maher, A. (1980). Eucaryotic transposable genetic elements with inverted terminal repeats. *Cell*, 20, 639-647.

Powell, J.R. (1997). *Progress and prospects in evolutionary biology: the Drosophila model*. Oxford University Press.

Praefcke, G.J.K. & McMahon, H.T. (2004). The dynamin superfamily: universal membrane tubulation and fission molecules? *Nat Rev Mol Cell Biol*, 5, 133-147.

Quesneville, H., Nouaud, D. & Anxolabehere, D. (2005). Recurrent Recruitment of the THAP DNA-Binding Domain and Molecular Domestication of the P-Transposable Element. *Molecular Biology and Evolution*, 22, 741 -746.

Quesneville, H., Nouaud, D. & Anxolabehere, D. (2006). *P-elements* and MITE relatives in the

whole genome sequence of Anopheles gambiae. *BMC Genomics*, 7, 214.

Rambaut, A. (2006). *FigTree*.

Rasmuson-Lestander, A. & Ekström, K. (1996). Genetic and molecular analysis of a set of unstable *white* mutants in *Drosophila* melanogaster. *Genetica*, 98, 179-192.

Rebatchouk, D. & Narita, J.O. (1997). *Foldback* transposable elements in plants. *Plant Mol. Biol*, 34, 831-835.

Redder, P. & Garrett, R.A. (2006). Mutations and Rearrangements in the Genome of Sulfolobus solfataricus P2. *J. Bacteriol.*, 188, 4198-4206.

Reed, L.K., Nyboer, M. & Markow, T.A. (2007). Evolutionary relationships of *Drosophila* mojavensis geographic host races and their sister species *Drosophila* arizonae. *Mol. Ecol*, 16, 1007-1022.

Reiss, D., Quesneville, H., Nouaud, D., Andrieu, O. & Anxolabéhère, D. (2003). Hoppel, a P-like Element Without Introns: a *P-Element* Ancestral Structure or a Retrotranscription Derivative? *Molecular Biology and Evolution*, 20, 869 -879.

Richardson, J.M., Dawson, A., O'hagan, N., Taylor, P., Finnegan, D.J. & Walkinshaw, M.D. (2006). Mechanism of Mos1 transposition: insights from structural analysis. *EMBO J*, 25, 1324-1334.

Richards, S., Liu, Y., Bettencourt, B.R., Hradecky, P., Letovsky, S., Nielsen, R., et al. (2005). Comparative genome sequencing of *Drosophila* pseudoobscura: chromosomal, gene, and cis-element evolution. *Genome Res*, 15, 1-18.

Rio. (2002). P transposable elements in *Drosophila* melanogaster. In: *Mobile DNA II* (eds. Craig, N., Craigie, R., Gellert, M. & Lambowitz , Alan M.). ASM Press, Washington DC, pp. 485-515.

Rodriguez, C., Fanara, J.J. & Hasson, E. (1999). Inversion Polymorphism, Longevity, and Body Size in a Natural Population of *Drosophila buzzatii*. *Evolution*, 53, 612-620.

Roeder, G.S. (1983). Unequal crossing-over between yeast transposable elements. *MGG Molecular & General Genetics*, 190, 117-121.

Rong, Y.S. & Golic, K.G. (2000). Gene targeting by homologous recombination in Drosophila. *Science*, 288, 2013-2018.

Roussigne, M., Kossida, S., Lavigne, A.-C., Clouaire, T., Ecochard, V., Glories, A., et al. (2003). The THAP domain: a novel protein motif with similarity to the DNA-binding domain of *P-element* transposase. *Trends in Biochemical Sciences*, 28, 66-69.

Le Rouzic, A., Boutin, T.S. & Capy, P. (2007). Long-term evolution of transposable elements. *Proceedings of the National Academy of Sciences*, 104, 19375 -19380.

Le Rouzic, A. & Capy, P. (2005). The First Steps of Transposable Elements Invasion: Parasitic

Strategy vs. Genetic Drift. *Genetics*, 169, 1033-1043.

Le Rouzic, A. & Capy, P. (2006). Population Genetics Models of Competition Between Transposable Element Subfamilies. *Genetics*, 174, 785-793.

Le Rouzic, A. & Capy, P. (2009). Theoretical approaches to the dynamics of transposable elements in genomes, populations and species. In: *Transposons and the dynamic genome*, Genome dynamics and stability (eds. Lankenau, D.-H. & Volff, J.-N.). Springer, Heidelberg, pp. 1-19.

Rubin, E. & Levy, A.A. (1997). Abortive gap repair: underlying mechanism for Ds element formation. *Mol. Cell. Biol.*, 17, 6294-6302.

Rubin, G.M., Hazelrigg, T., Karess, R.E., Laski, F.A., Laverty, T., Levis, R., et al. (1985). Germ line specificity of *P-element* transposition and some novel patterns of expression of transduced copies of the *white* gene. *Cold Spring Harb. Symp. Quant. Biol*, 50, 329-335.

Rubin, G.M., Kidwell, M.G. & Bingham, P.M. (1982). The molecular basis of P-M hybrid dysgenesis: the nature of induced mutations. *Cell*, 29, 987-994.

Rubin, G.M. & Spradling, A.C. (1982). Genetic transformation of *Drosophila* with transposable element vectors. *Science*, 218, 348-353.

Rubin, G.M., Yandell, M.D., Wortman, J.R., Gabor Miklos, G.L., Nelson, C.R., Hariharan, I.K., et al. (2000). Comparative genomics of the eukaryotes. *Science*, 287, 2204-2215.

Ruiz, A., Fontdevila, A., Santos, M., Seoane, M. & Torroja, E. (1986). The Evolutionary History of *Drosophila buzzatii*. VIII. Evidence for Endocyclic Selection Acting on the Inversion Polymorphism in a Natural Population. *Evolution*, 40, 740-755.

Ruiz, A. & Heed, W.B. (1988). Host-Plant Specificity in the Cactophilic *Drosophila mulleri* Species Complex. *Journal of Animal Ecology*, 57, 237-249.

Ruiz, A., Heed, W.B. & Wasserman, M. (1990). Evolution of the mojavensis cluster of cactophilic *Drosophila* with descriptions of two new species. *J. Hered*, 81, 30-42.

Ruiz, A., Santos, M., Barbadilla, A., Quezada-Diaz, J.E., Hasson, E. & Fontdevila, A. (1991). Genetic Variance for Body Size in a Natural Population of *Drosophila buzzatii*. *Genetics*, 128, 739 -750.

Ruiz, A. & Wasserman, M. (1993). Evolutionary cytogenetics of the *Drosophila buzzatii* species complex. *Heredity*, 70, 582-596.

Russo, C., Takezaki, N. & Nei, M. (1995). Molecular phylogeny and divergence times of drosophilid species. *Mol Biol Evol*, 12, 391-404.

Ryder, E. & Russell, S. (2003). Transposable elements as tools for genomics and genetics in Drosophila. *Briefings in Functional Genomics and Proteomics*, 2, 57-71.

Sabogal, A., Lyubimov, A.Y., Corn, J.E., Berger, J.M. & Rio, D.C. (2010). THAP proteins target specific DNA sites through bipartite recognition of adjacent major and minor grooves. *Nat Struct Mol Biol*, 17, 117-123.

Sabogal, A. & Rio. (2010). A green fluorescent protein solubility screen in E. coli reveals domain boundaries of the GTP-binding domain in the *P-element* transposase. *Protein Sci*, 19, 2210-2218.

Sambrook, J., Fritsch, E.F. & Maniatis, T. (1989). *Molecular cloning: a laboratory manual*. Cold Spring Harbor Laboratory Press.

Sánchez-Gracia, A., Maside, X. & Charlesworth, B. (2005). High rate of horizontal transfer of transposable elements in Drosophila. *Trends in Genetics*, 21, 200-203.

Sarkar, A. (2003). *P-elements* are found in the genomes of nematoceran insects of the genus Anopheles. *Insect Biochemistry and Molecular Biology*, 33, 381-387.

Schaeffer, S.W., Bhutkar, A., McAllister, B.F., Matsuda, M., Matzkin, L.M., O'Grady, P.M., 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.

Schneuwly, S., Kuroiwa, A. & Gehring, W.J. (1987). Molecular analysis of the dominant homeotic Antennapedia phenotype. *EMBO J*, 6, 201-206.

Schwartz, A., Chan, D.C., Brown, L.G., Alagappan, R., Pettay, D., Disteche, C., et al. (1998). Reconstructing hominid Y evolution: X-homologous block, created by X-Y transposition, was disrupted by Yp inversion through LINE-LINE recombination. *Hum. Mol. Genet.*, 7, 1-11.

Shimodaira, H. (2002). An approximately unbiased test of phylogenetic tree selection. *Syst. Biol*, 51, 492-508.

Shimodaira, H. & Hasegawa, M. (2001). CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics*, 17, 1246 -1247.

Silva, J.C., Loreto, E.L. & Clark, J.B. (2004). Factors that affect the horizontal transfer of transposable elements. *Curr Issues Mol Biol*, 6, 57-71.

Simonelig, M. & Anxolabéhère, D. (1991). A *P-element* of Scaptomyza pallida is active in *Drosophila* melanogaster. *Proceedings of the National Academy of Sciences of the United States of America*, 88, 6102 -6106.

Singh, N., Larracuente, A., Sackton, T. & Clark, A. (2009). Comparative Genomics on the *Drosophila* Phylogenetic Tree. *ANNUAL REVIEW OF ECOLOGY EVOLUTION AND SYSTEMATICS*, 40, 459-480.

Sinzelle, L., Kapitonov, V.V., Grzela, D.P., Jursch, T., Jurka, J., Izsvák, Z., et al. (2008).

Transposition of a reconstructed Harbinger element in human cells and functional homology with two transposon-derived cellular genes. *Proceedings of the National Academy of Sciences*, 105, 4715 -4720.

Spradling, A.C., Stern, D., Beaton, A., Rhem, E.J., Laverty, T., Mozden, N., et al. (1999). The Berkeley *Drosophila* Genome Project gene disruption project: Single *P-element* insertions mutating 25% of vital *Drosophila* genes. *Genetics*, 153, 135-177.

Spradling, A.C., Stern, D.M., Kiss, I., Roote, J., Laverty, T. & Rubin, G.M. (1995). Gene disruptions using P transposable elements: an integral component of the *Drosophila* genome project. *Proc. Natl. Acad. Sci. U.S.A*, 92, 10824-10830.

Spradling & Rubin, G.M. (1982). Transposition of cloned *P-elements* into *Drosophila* germ line chromosomes. *Science*, 218, 341-347.

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22, 2688-2690.

Talavera, G. & Castresana, J. (2007). Improvement of Phylogenies after Removing Divergent and Ambiguously Aligned Blocks from Protein Sequence Alignments. *Systematic Biology*, 56, 564 -577.

Tamura, K., Dudley, J., Nei, M. & Kumar, S. (2007). MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol. Biol. Evol*, 24, 1596-1599.

Tamura, K., Subramanian, S. & Kumar, S. (2004). Temporal Patterns of Fruit Fly (Drosophila) Evolution Revealed by Mutation Clocks. *Mol Biol Evol*, 21, 36-44.

Tang, M., Cecconi, C., Bustamante, C. & Rio, D.C. (2007). Analysis of *P-element* transposase protein-DNA interactions during the early stages of transposition. *J. Biol. Chem*, 282, 29002-29012.

Tang, M., Cecconi, C., Kim, H., Bustamante, C. & Rio, D.C. (2005). Guanosine triphosphate acts as a cofactor to promote assembly of initial *P-element* transposase-DNA synaptic complexes. *Genes Dev*, 19, 1422-1425.

Templeton, N.S. & Potter, S.S. (1989). Complete *foldback* transposable elements encode a novel protein found in *Drosophila* melanogaster. *EMBO J*, 8, 1887-1894.

The C. elegans Sequencing Consortium. (1998). Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology. *Science*, 282, 2012 -2018.

Truett, M.A., Jones, R.S. & Potter, S.S. (1981). Unusual structure of the *FB* family of transposable elements in Drosophila. *Cell*, 24, 753-763.

Venter, J.C., Adams, M.D., Sutton, G.G., Kerlavage, A.R., Smith, H.O. & Hunkapiller, M. (1998). Shotgun Sequencing of the Human Genome. *Science*, 280, 1540 -1542.

Villasante, A., Abad, J.P., Planelló, R., Méndez-Lago, M., Celniker, S.E. & de Pablos, B.

(2007). *Drosophila* telomeric retrotransposons derived from an ancestral element that was recruited to replace telomerase. *Genome Research*, 17, 1909 -1918.

Wasserman, M. (1982). Cytological evolution in the *Drosophila repleta* species group. In: *Ecological Genetics and Evolution. The Cactus-Yeast-Drosophila Model System* (eds. Barker, J.S.F. & Starmer, W.T.). Academic Press, Sydney, pp. 49-64.

Wasserman, M. (1992). Cytological evolution of the *Drosophila repleta* species group. In: *Drosophila Inversion Polymorphism* (eds. Krimbas, C.B. & Powell, J.R.). CRC Press, Boca Raton, Florida, pp. 455–552.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., et al. (2007). A unified classification system for eukaryotic transposable elements. *Nat Rev Genet*, 8, 973-982.

Yang, G., Nagel, D.H., Feschotte, C., Hancock, C.N. & Wessler, S.R. (2009). Tuned for Transposition: Molecular Determinants Underlying the Hyperactivity of a Stowaway MITE. *Science*, 325, 1391-1394.

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci*, 13, 555-556.

Yuan, Y.-W. & Wessler, S.R. (2011). The catalytic domain of all eukaryotic cut-and-paste transposase superfamilies. *Proceedings of the National Academy of Sciences*.

Zayed, H., Izsvak, Z., Walisko, O. & Ivics, Z. (2004). Development of Hyperactive Sleeping Beauty Transposon Vectors by Mutational Analysis. *Mol Ther*, 9, 292-304.

Zdobnov, E.M. & Apweiler, R. (2001). InterProScan--an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, 17, 847-848.

# Acknowledgements

En primer lugar, quiero agradecer a Alfredo Ruiz el que me diera la oportunidad de llevar a cabo este proyecto. En segon lloc, agrair a la Marta Puig el haver-me introduït al món de la experimental i la bioinformàtica "artesanal", i el suport del Grup de Genòmica, Bioinformàtica i Evolució. Per una altra part, agrair a les secretàries (la Julia, la Maite, l'Elena i la Conchi) i a les tècnics (la Montse i la Raquel) l'ajut en les diferents gestions i, fins i tot, en l'interminable comptatge de mosques.

Als companys de laboratori i als de la segona planta, sempre els tindré present per totes els moments de pausa, cafè i les activitats "extra-escolars" que hem dut a terme durant aquestos anys.

A very-big thank you to all the members of the Ronnie Chalmers' lab in the University of Nottingham, for their patience teaching me protein techniques: Azeem, Danxu, Belinda and Neill. And, of course, Corentin and all the people from the Box that were always up for some pints.

També m'agradaria nomenar al Dojo del Shihan Gallego i a tots els companys d'entrenament, que han fet el periode d'escriptura de la tesi fos més desafiant, "***mens sana in corpore sano***". OSU! Jo ara diria "si no duele no es una tesis..haha"!

I fer una molt especial menció al Víctor, gràcies per tot i per estar ahí.

I des d'aquí nomenar a la família i als amics de tota la vida, que m'han vist tan poquet durant aquests últims anys, però amb els qui sempre he pogut comptar tot i la distància.

Finally, I would like to thank the Thesis Dissertation Committee for accepting to be part of it: Dr Pierre Capy, Dr Mario Cáceres, Dra Elena Casacuberta, Dr Julio Rozas and Dra Josefa González.

This work was supported by a Formación de Personal Investigador doctoral fellowship (to Mar Marzo.) and Secretarıa de Estado de Universidades e Investigacion (Ministerio de Educacion y Ciencia, Spain) Grant BFU2005-022379 and BFU2008-04988 awarded to Alfredo Ruiz.

Moltes gràcies a tots els que heu estat ahí durant aquesta etapa! De tot cor! ;)