Analysis of *Drosophila buzzatii* transposable elements

Doctoral Thesis

Nuria Rius Camps

Departament de Genètica i de Microbiologìa,
Universitat Autonoma de Barcelona,
Bellaterra (Barcelona), Spain

Memòria presentada per la Llicenciada en Biologia Nuria Rius Camps per a optar al grau de Doctora en Genètica.

Nuria Rius Camps

Bellaterra, a 23 de novembre de 2015

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

CERTIFICA que Nuria Rius Camps ha dut a terme sota la seva direcció el treball de recerca realitzat al Departament de Genètica i Microbiologia de la Facultat de Biociències de la Universitat Autònoma de Barcelona que ha portat a l'elaboració d'aquesta Tesi Doctoral titulada "Analysis of Drosophila buzzatii transposable elements".

I perquè consti als efectes oportuns, signa el present certificat a Bellaterra, a 23 de novembre de 2015

Alfredo Ruiz Panadero

I tell you all this because it's worth recognizing that there is no such thing as an overnight success. You will do well to cultivate the resources in yourself that bring you happiness outside of success or failure. The truth is, most of us discover where we are headed when we arrive. At that time, we turn around and say, yes, this is obviously where I was going all along. It's a good idea to try to enjoy the scenery on the detours, because you'll probably take a few.

(Bill Watterson)

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

BAC	Bacterial Artificial Chromosome
BDGP	Berkeley Drosophila Genome Project
BLAST	Basic Local Alignment Search Tool
BSC	Barcelona Supercomputing Center
CNAG	Spanish Centro Nacional de Análisis Genómico
DINE-1	Drosophila Interspersed Element-1
DIRS	Dictyostelium Intermediate Repeat Sequence
DNA	Deoxyribonucleic Acid
ERVK	Endogenous Retrovirus-K
HT	Horizontal transfer
LINE	Long Interspersed Repetitive Elements
LTR	Long Terminal Repeats
MITE	Miniature Inverted-Repeat TE
NCBI	National Center for Biotechnology Information
NGS	Next-Generation Sequencing
ORF	Open Reading Frame
PE	Paired End
piRNA	piwi-interacting RNA
PLE	Penelope-like Element
RNA	Ribonucleic Acid
SDS	Sodium Dodecyl Culfate
SINE	Short Interspersed Repetitive Elements
TE	Transposable Element
THAP	Thanatos-associated Protein
TIR	Terminal Inverted Repeat
TSD	Target Site Duplication
UAB	Universitat Autònoma de Barcelona
UTA	University of Texas at Arlington

ABSTRACT

Transposable genetic elements are genetic units able to insert themselves in other regions of the genomes they inhabit, and are present in almost all eukaryotes analyzed. The interest of transposable element analysis, it is not only because its consideration as intragenomic parasites. Transposable elements are an enormous source of variability for the genomes of their hosts, and are therefore key to understanding its evolution. In this work we addressed the analysis of *Drosophila buzzatii* transposable elements from two different approaches, the detailed study of one family of transposable elements and global analysis of all elements present in the genome. The study of chromosomal inversions in *D. buzzatii* led to the description of the non-autonomous transposable element, *BuT5*, which was later found to cause polymorphic chromosomal inversions in *D. mojavensis* and *D. uniseta*. In this work we have characterized the transposable element *BuT5* and we have described its master element. *BuT5* is found in 38 species of the group of species *D. repleta*. The autonomous element that mobilizes *BuT5* is a *P* element, we described three partial copies in the sequenced genome of *D. mojavensis* and a complete copy in *D. buzzatii*. The full-length and putatively active copy has 3386 base pairs and encodes a transposase of 822 residues in seven exons. Moreover we have annotated, classified and compared the transposable elements present in the genomes of two strains of *D. buzzatii*, st-1 and j-19, recently sequenced with next-generation sequencing technology, and in the *D. mojavensis*, the phylogenetically closest species sequenced, in this case with Sanger technology. Transposable elements make up for 8.43%, the 4.15% and 15.35% of the assemblies of the genomes of *D. buzzatii* st-1, j-19 and it D . mojavensis respectively. Additionally, we have detected a bias in the transposable elements content of genomes sequenced using next-generation sequencing technology, compared with the content in genomes sequenced with Sanger technology. We have developed a method based on the coverage that allowed us to correct this bias in the genome of *D. buzzatii* st-1 and have more realistic estimates of the content in transposable elements. Using this method we have determined that the transposable element content in *D. buzzatii* st-1 is between 10.85% and 11.16%. Additionally, the estimates allowed us to infer that the Helitrons order has undergone multiple cycles of activity and that the superfamily Gypsy and BelPao have recently been active in *D. buzzatii*.

RESUMEN

Los elementos transponibles son unidades genéticas capaces de insertarse en otras regiones de los genomas en los que habitan y están presentes en casi todas las especies eucariotas estudiadas. El interés del análisis de los elementos transponibles no se debe únicamente a su consideración de parásitos intragenómicos. Los elementos transponibles suponen una enorme fuente de variabilidad para los genomas de sus hospedadores, y son por lo tanto claves para comprender su evolución. En este trabajo hemos abordado el análisis de los elementos transponibles de *Drosophila buzzatii* desde dos enfoques distintos, el estudio detallado de una única familia de elementos transponibles y el análisis global de todos los elementos presentes en el genoma. El estudio de inversiones cromosómicas en *D. buzzatii* llevó a la descripción del elemento transponible no autónomo, *BuT5*, que posteriormente se descubrió como elemento causante de inversiones polimórficas en *D. mojavensis* y *D. uniseta*. En este trabajo hemos caracterizado el elemento transponible *BuT5* y hemos descrito su elemento maestro. *BuT5* se encuentra en 38 especies del grupo de especies de *D. repleta*. El elemento autónomo que moviliza a *BuT5* es un elemento *P*, del que hemos descrito 3 copias parciales en el genoma secuenciado de *D. mojavensis* y una copia completa en *D. buzzatii*. La copia completa y putativamente activa tiene 3386 pares de bases y codifica una transposasa de 822 residuos en siete exones. Por otra parte hemos anotado, clasificado y comparado los elementos transponibles presentes en los genomas de dos cepas de *D. buzzatii*, st-1 y j-19, secuenciadas recientemente con tecnología de nueva generación, y en el de *D. mojavensis*, la especie filogenéticamente más cercana secuenciada, en este caso mediante tecnología Sanger. Los elementos transponibles representan el 8.43%, el 4.15% y el 15.35% de los ensamblajes de los genomas de *D. buzzatii* st-1, j-19 y *D. mojavensis* respectivamente. Adicionalmente hemos detectado un sesgo en el contenido de elementos transponibles de los genomas secuenciados mediante tecnología de nueva generación, comparado con el contenido en los genomas secuenciados con tecnología Sanger. Hemos desarrollado un método basado en la cobertura que nos ha permitido corregir este sesgo en el genoma de *D. buzzatii* st-1 y contar con estimas mas realistas del contenido en elementos transponibles. Así hemos determinado que el contenido en elementos transponibles en *D. buzzatii* st-1 es de entre el 10.85% y el 11.16% del genoma. Adicionalmente las estimas nos han

permitido inferir que el orden de los Helitrones ha experimentado múltiples ciclos de actividad y que las superfamilias Gypsy y BelPao han sido recientemente activas en *D. buzzatii*.

1

INTRODUCTION

1.1 TRANSPOSABLE ELEMENTS

Transposable elements (TEs) are genetic units able to make copies of themselves that insert elsewhere within a host genome. They are almost ubiquitous; all eukaryotic genomes sequenced to date, except for *Plasmodium falciparum* (Gardner et al., 2002), have TE sequences within them. Moreover, TEs are capable of spreading within genomes, populations or species (Feschotte and Pritham, 2007; Rebollo et al., 2012).

The work of Barbara McClintock in chromosome breakage in maize during the 1940s and 1950s lead her to the discovery of mutable genes that could change its position within or between chromosomes. She named them "controlling elements" for their potential to regulate gene expression in precise ways. Furthermore, her findings challenged the concept that genes were static units arranged linearly in chromosomes (McClintock, 1983), an idea that, thanks to linkage maps, was just becoming to be accepted by the scientific community. In 1950, McClintock proposed that these mutable *loci* were responsible for variegation not only in maize but also in Drosophila (McClintock, 1950). McClintock's ideas take on greater significance if we consider that they roughly coincided in time with Watson and Crick's double-helical model for the structure of DNA (Watson and Crick, 1953). However, it was not until the 1970s that her findings were confirmed in other organisms and the implications of the mobile nature of vastly widespread genetic entities were recognized (Fedoroff, 2012). McClintock was credited with several honors including the Nobel Prize in Physiology or Medicine in 1983 (McClintock, 1983).

Reassociation kinetics experiments, performed during the late 1960s and 1970s, showed that middle-repetitive sequences made up a significant part of most species genomes (Britten and Kohne, 1968). Interspersed repeats, a fraction of the middle-repetitive sequences, were corroborated to occupy different *loci* in different strains and several researchers termed them mobile elements and other

names no longer in use like nomadic DNA (Young, 1979). TE abundance and persistence in a wide range of species in the absence of an evident beneficial role at the level of individual organism made them to be considered as selfish and junk DNA. The term of selfish DNA, defined as a sequence capable of rising its numbers without making a specific contribution to the phenotype, was used in 1980 by Orgel and Crick (Orgel and Crick, 1980), and specifically applied to mobile DNA by Doolittle and Sapienza (Doolittle and Sapienza, 1980). TEs were considered henceforth parasites of genomes that remained in them because of a replicative advantage over the host sequences (Hardman, 1986).

However, the current opinion on TEs may be changing again. Even though TEs are known for their deleterious effects interrupting host sequences, the examples of TEs exapted by their hosts to play a cellular function are more common than it was anticipated. TEs bear regulatory sequences that can multiply and spread across a genome, conferring the ability to lay the groundwork for regulatory networks (Casacuberta and González, 2013; Feschotte, 2008; Hua-Van et al., 2011; Kidwell and Lisch, 2001; Rebollo et al., 2012).

1.1.1 TRANSPOSABLE ELEMENT CLASSIFICATION

The discovery of new transposable elements and the similarities and differences between them made necessary to develop a classification system. The first classification system was proposed by Finnegan in 1989, and distinguished two classes of TEs, depending on the transposition intermediate (Finnegan, 1989). Class I, or retrotransposons, transposed via an RNA intermediate, while class II elements used a DNA intermediate. Over the following years many groups of TEs were described and placed within Finnegan's classes. These groups, superfamilies and families responded to a common origin, inferred from their sequence, structural, and transposition mechanism similarities.

In 2007 after the release of new genomes and the foreseeable increase in the number of sequenced genomes that would follow, two review articles were published with updated TE classification systems (Jurka et al., 2007; Wicker et al., 2007). Wicker and collaborators published a more comprehensive classification system made to help non-TE-experts to annotate new genomes and maintain TE classification coherence (Figure 1). They used Finnegan's class denomination, and names that were already in use in the TE community, like families and superfamilies. TEs were classified in six hierarchical levels: class, subclass, order, superfamily, family and subfamily.

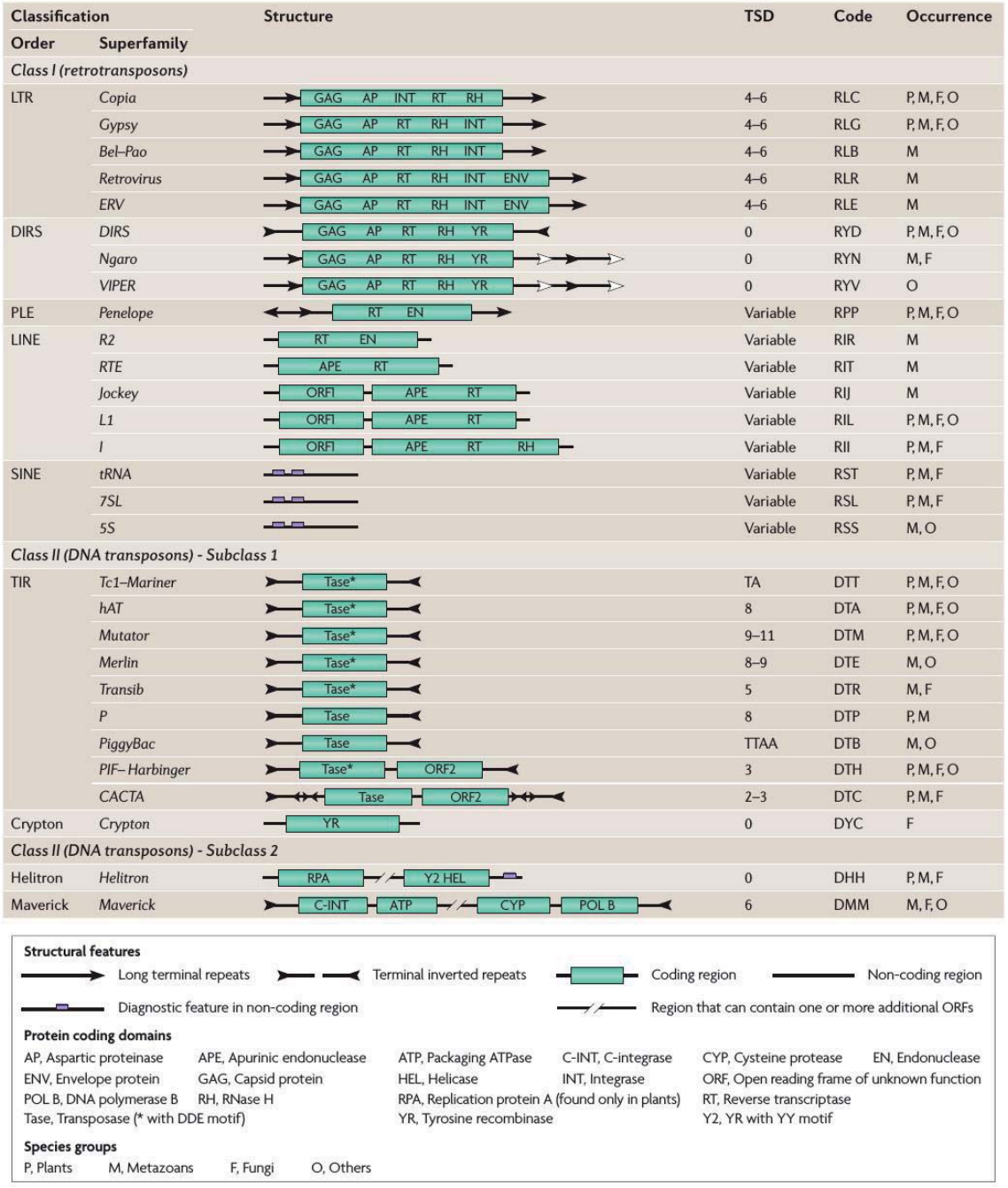

Figure 1.: Proposed TE classification. Taken from Wicker et al. (2007).

In Wicker's classification, like in Finnegan's system, classes distinguish between the presence or absence of an RNA transposition intermediate, Class I and Class II, respectively. Within a class, subclasses divide elements that remain in

their position and transpose a copy of themselves, and those elements that leave the donor site to insert elsewhere in the genome. Class I only has one subclass, as all TEs remain in the donor site and are transcribed into an RNA intermediate which is then retro-transcribed by a TE-encoded reverse transcriptase. Class I, and subclass 1, are divided in five orders, according to major differences in the insertion mechanisms (LTR, DIRS, PLE, LINE, and SINE). Class II is divided in two subclasses. Elements in subclass 1, which comprehends two orders, TIR elements and Crypton, require the cleavage of both DNA strands to transpose. In addition, elements in subclass 2, comprising two orders, Helitron and Polinton (or Maverick), require the displacement of one strand. It is important to note that this classification based on the presence or absence of RNA intermediates, or the number of strands cut during transposition does not necessarily imply phylogenetic relationship.

The TIR elements order, classically known as cut-and-paste transposons, is characterized by the presence of Terminal Inverted Repeats (TIRs) at their terminal ends. Besides this repeats of variable length between superfamilies, other inner repeats can be present. Most of TIR elements encode a single gene with transpose activity; PiF-Harbinger and CACTA superfamilies encode a second ORF (DeMarco et al., 2006; Wicker et al., 2003). An extensive analysis of TIR transposases determined the evolutionary relationships among superfamilies and revealed that all of them have a DDE/D motif (two aspartic acid (D) residues and a glutamic acid (E) residue or a third D) (Yuan and Wessler, 2011).

Orders are divided into superfamilies, which are formed by TEs with similarities at the protein level and in structural features like target site duplication (TSD) presence and size. According to Wicker and collaborators (Wicker et al., 2007), in 2007, nine superfamilies belonged to the TIR order, while 13 were proposed by Jurka and collaborators in 2007. However, this number is growing fast and some authors consider that the TIR order comprises 17 to 19 superfamilies (Bao et al., 2009; Yuan and Wessler, 2011). This figure is expected to keep growing as new genomes are being sequenced and annotated. Families, the next hierarchical level, comprehend TEs that have high similarity at the protein level and also similarity at the nucleotide level in the coding and terminal regions. The subfamily division differentiates phylogenetically close clades within a family, or in some cases autonomous and non-autonomous members of a family.

Complete canonical TEs encode the elements necessary to transpose into another genomic location, hence being autonomous elements and displaying the traits needed to be recognized by the transposition machinery. This does not apply to SINEs (Short Interspersed Repetitive Elements), which are not deletion derivatives of other elements, are naturally non-autonomous and rely on LINEs

(Long Interspersed Repetitive Elements) to transpose. During the "life" of a mobile element it can suffer point mutations and small insertions and deletions, which partially or completely remove the protein domains. These defective or deleted copies consequently become non-autonomous elements, however, and as long as they keep the features recognized by the transposition machinery will be active if an autonomous copy is present in the genome (Wicker et al., 2007).

Within non-autonomous elements there is a special group, called MITEs (Miniature Inverted-Repeat TEs), which stands out for their capacity to reach much higher numbers than the canonical elements relying on autonomous copies to transpose. Tourist and Stowaway, the first MITEs described (Bureau and Wessler, 1992, 1994), where found in plants as 100 to 500 bp sequences, with a similar insertion preference, and structural similarities, although they had no significant sequence similarities to known TEs. The term MITE was created to avoid a starting confusion between these new, short, abundant, and orphan elements and SINEs. Over the years MITEs have been found in more species and have been linked to TIR transposons, revealing complex relationships among different elements. MITEs most likely are deletion derivatives of full-length elements, although other mechanisms may be involved in their formation (Wallau et al., 2014). They have conserved terminal regions, especially the TIRs, and do not have coding capacity (Feschotte et al., 2002). The MITEs found in Drosophila (Rossato et al., 2014; de Freitas Ortiz et al., 2010; Holyoake and Kidwell, 2003) are somewhat longer and have a lower copy number than those found in plants.

1.1.2 TES IN THEIR HOST GENOMES

The "life cycle" of a TE has been divided in three states, invasion, maturity, and senescence (Figure 2). When a TE invades a genome it starts a proliferation or dynamic replication phase with new insertions and occasional mutations that yield some inactive copies. The second phase, maturity is characterized by the increase of copy inactivation due to mutations. During this phase the number of new insertion matches the number of inactive copies. Finally, the degradation or senescence arrives when there are no active copies able to transpose. This phase can last for millions of years while the inactive copies can be lost from the population, deleted or remain in the host genome until the remnants accumulate enough mutations to become unrecognizable (Kidwell and Lisch, 2001).

Nevertheless, that previous view was a simplification and the relationship between TEs and their hosts is far more complex (Le Rouzic et al., 2007). TEs are not just parasites that increase their activity until their disappearance of the

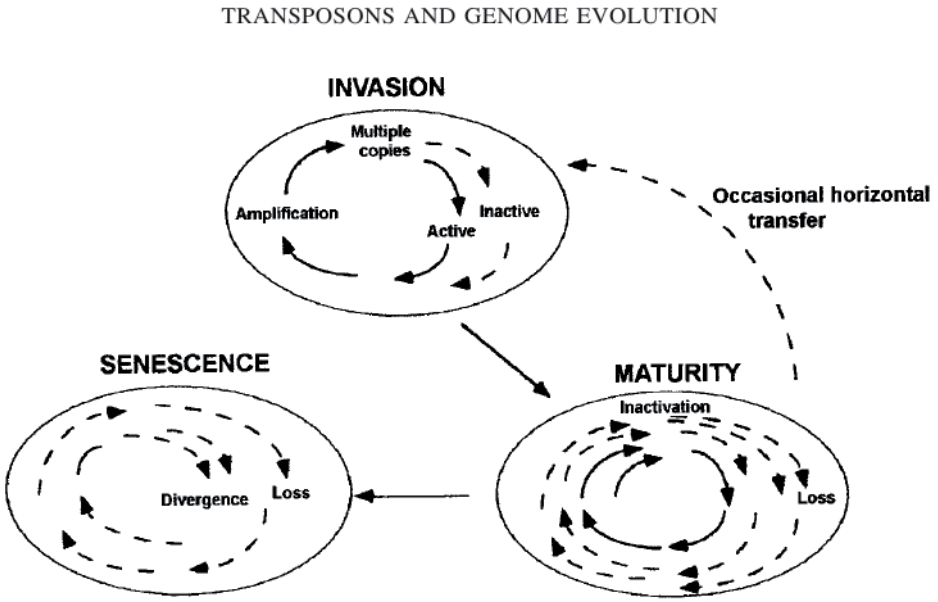

Figure 2.: General features of the life cycle of a Class II transposable element. Taken from Kidwell and Lisch (2001).

genomes they inhabit. Contrary to that, we now know that TEs can experience several waves of activity, as genome conditions change (Ray et al., 2008; Yang and Barbash, 2008), and that they can contribute in multiple ways to the host genome evolution (Casacuberta and González, 2013).

Even though TEs are not just parasites their presence and activity can be detrimental for the host fitness. As a result, organisms have developed several pathways to repress TE activity, such as the piwi-interacting RNA or piRNA mechanism. This pathway is enriched in animal gonads, including Drosophila, were transposition is more sensitive. piRNAs are non-coding RNAs processed from single-stranded (ss) RNA into pieces of 24 to 35 nucleotides in length. The majority of piRNAs are translated from piRNA clusters or active transposons, but are antisense to transposon transcripts, being able to pair with them. piRNA and PIWI proteins form piRNA-induced silencing complexes (piRISCs) and once double stranded (ds) RNA is formed with TE transcripts those are lead to their destruction preventing transposition (Siomi et al., 2011; Hirakata and Siomi, 2015).

However, TEs are not linked to a unique host lineage and its fate. Horizontal gene transfer is a phenomenon by which sequences are transmitted from one species to another not via vertical or parental transfer, and TEs by its mo-

bile nature seem especially prone to horizontal transfer (HT) (Schaack et al., 2010; Wallau et al., 2012). Since the classical description in 1990 of a HT event that let a *P element* from *D. willistoni*cross the species barrier and invade all *D. melanogaster* population in approximately 50 years (Daniels et al., 1990) more cases have been published, and HT does not seem an exceptional event (Bartolomé et al., 2009; Wallau et al., 2012).

Probably to understand the interactions between TEs and their hosts it is important to consider the following four factors. First, TEs are an incredibly large source of variability, ranging from small scale mutations, to large rearrangements or epigenetic changes that TEs can leave behind even after their excision or loss. Second, TEs and the mutations they induce are subject to natural selection, that would tend to remove deleterious insertions, and neutral or advantageous may be maintained in the populations. Additionally, as any other genome component, TEs are affected by other evolutionary forces like genetic drift or migration. Third, TEs can increase their copy number independently of the genome, which allows TEs to evolve independently of the fate of the genome to a certain degree. Lastly, TEs are not merely genome parasites; both actors have contributed to their mutual transformation becoming an essential part to understand their evolution (Hua-Van et al., 2011).

1.1.3 THE *P ELEMENT*

The *P element* is one of the best studied eukaryotic transposable elements. It was discovered in *D. melanogaster*as the cause of hybrid dysgenesis, which involved high rate of sterility, mutations, chromosomal abnormalities, rearrangements, and male recombination. Males carrying *P elements* (P for paternally contributing strains) that mated with females lacking autonomous *P elements* resulted in progeny with genetic instability not observed in the reciprocal crosses. The molecular analysis led to the isolation and cloning of *P elements* (Bingham et al., 1981; Rubin et al., 1982), which were later used as vectors for germ line gene transfer in Drosophila (Rio, 2002).

D. melanogaster strains founded with individuals collected before the mid-1960s in America and 1974 in the URSS, were devoid of *P elements* , while they were present in younger strains. *P elements* were absent in the rest of the studied species of the melanogaster subgroup and an invasion of *D. melanogaster* population was proposed (Anxolabéhère et al., 1988). In 1990, Daniels and collaborators showed that *D. willistoniP element* was the closest relative to the *D. melanogaster* el-

ement, only differing in one nucleotide out of 2.9 kb, proving evidence for the horizontal transference of a TE between eukaryotes (Daniels et al., 1990).

As I mentioned before, *P elements* colonized all *D. melanogaster* populations in historical times (approximately 50 years ago). It is in this species where most of the *P element* traits were studied. The canonical *D. melanogaster P element* is 2.9 kb long, has 31-bp terminal inverted repeats (TIRs) and 11-bp internal inverted repeats located about 100 bp from the ends. *P elements* generate an 8-bp target site duplication (TSD) upon insertion. Autonomous copies have a four-exon gene that encodes a 751-aa transposase (Rio, 2002).

Active *P element* transposase is only expressed in the germ line cells of *D. melanogaster*, where the splicing of the three introns occurs, and the mRNA is translated into an 87 kDa protein. Thus, new insertions may be passed to offspring. In somatic cells and in part of the germ line, the third intron (IVS3) is retained and the mRNA, with a premature stop codon, is translated into a 66 kDa protein that acts as a repressor of transposition, named a type I repressor. Other truncated variants of the transposases have been described to also act as repressors in *D. melanogaster* (KP/type II repressors). Different patterns of splicing, producing both putatively active transposase and repressors, have also been described in other species like *D. bifasciata*, *D. helvetica* , and *Scaptomyza pallida* (Haring et al., 1998; Pinsker et al., 2001).

1.2 DROSOPHILA AS A MODEL ORGANISM

D. melanogaster was first used as a genetic model by Thomas Morgan in 1908, who studied the inheritance of mutations. Many traits made *D. melanogaster* a suitable species for studding genetics in the first place, it needs little space even for large cultures, its maintenance cost is low, it has a short generation time (10 days at room temperature), high fecundity (100 eggs per day), and it is easy to manipulate once anesthesiated. Morgan's work, led him and his students to the understanding of major biology breakthroughs. The analysis of multiple mutant flies were key to the modern interpretation of Mendelism, the linear disposition of genes, or dosage compensation, all discovered in a small fly and with repercussions in all species' research (Green, 2010).

However, beyond the species adequacy to become a model species in the first place, the knowledge and resources that had been built upon over a century of studies had an important in role in making *D. melanogaster* the excellent model to study eukaryotes genetics that it is today (Matthews et al., 2005). In the 1970s and 1980s tools like balancer chromosomes, or banding techniques in the gi-

ant polytene chromosomes guaranteed *D. melanogaster* a privileged place among model organisms. Additionally, these advances helped to classify phylogenetically a great deal of the more than 2000 species that form the *Drosophila genus*, particularly in the *Sophophora* subgenus, where *D. melanogaster* belongs, but also in the *Drosophila* subgenus. The knowledge of the phylogenetic relationships within the genus, allowed to these species to become a model system in which to study species evolution.

1.2.1 *D. BUZZATII* AND THE *D. REPLETA* SPECIES GROUP

D. buzzatii is a *Drosophila* subgenus species originally from South America, which feeds on decaying cladodes of Opuntia cacti and on the rotting stems of some columnar cacti (Hasson et al., 1992). Since Argentinian ports opened in the mid-1800s (Wasserman, 1992), *D. buzzatii* has spread through four of the six major biogeographical regions, South America, the South of Europe, North and Equatorial Africa, and Australia becoming a sub-cosmopolitan species (David and Tsacas, 1981).

Within the subgenus *Drosophila*, *D. buzzatii* belongs to the repleta group, mulleri subgroup and to the buzzatii complex (Ruiz and Wasserman, 1993). The polytene chromosome banding pattern analysis performed on repleta group species revealed more than 296 inversions, some of them shared between close species, helping to depict their phylogenetic relationships (Wasserman, 1992).

The research on *D. buzzatii* chromosomal rearrangements led to the discovery of the first natural chromosomal inversion caused by a TE (Cáceres et al., 1999). The recombination between two copies of the element Galileo, a TIR transposon of the P superfamily (Marzo et al., 2008), generated the 2j inversion in chromosome 2. Subsequently, Galileo was found to be the cause of two additional polymorphic inversions in *D. buzzatii* (Casals et al., 2003; Delprat et al., 2009). Inversion breakpoints were secondarily colonized by other TEs in part because of the reduced recombination in these regions (Cáceres et al., 2001; Delprat et al., 2009). Other fixed inversions within the repleta group had been proven to be caused by another TE. The inversions 2s of *D. mojavensis* (Guillén and Ruiz, 2012) and $2x^3$ of *D. uniseta* (Prada, 2010) were both caused by the non-autonomous element *BuT5*. This element does not encode a transposase and has remained unclassified for more than a decade. Consequently, the analysis of TEs in *D. buzzatii* and the species from repleta group have gained interest.

1.3 GENOMICS

In 2000 the genome of Drosophila melanogaster was published (Adams et al., 2000) in a joint effort to assess the viability of sequencing a complex eukaryotic genome before scaling up to the human genome and after the release of *Caenorhabditis elegans* (C. elegans Sequencing Consortium, 1998). At the same time, the genome provided an excellent resource, not just to learn to unravel the mysteries of a genome, but to contribute to the research in a useful model organism (Adams et al., 2000).

1.3.1 GENOMICS IN DROSOPHILA

D. melanogaster has become the gold standard for all the genomes sequenced after it. The first assembly was made with Whole Genome Shotgun (WGS) strategy (Myers et al., 2000), sequencing plasmid and Bacterial Artificial Chromosomes (BAC) paired-ends. That was a bold strategy at the time for a complex genome, instead of the clone-based more time-consuming approach. The first release of *D. melanogaster* genome combined that first assembly with a second draft genome with clone-based strategy and using 825 P1 and Bacterial Artificial Chromosomes (BAC) clones sequenced with Sanger technology (Adams et al., 2000).

Subsequent releases corrected the order and orientation of some scaffolds, closed gaps in the sequence, improved low quality regions, like the Y chromosome, and extended the assembly at the telomeric and centromeric ends of the chromosomes (Ashburner and Bergman, 2005; Celniker et al., 2002; Hoskins et al., 2015). In a similar manner, the functional annotation has become one of the more accurate among eukaryotes genomes in collaboration with the FlyBase team (Drysdale et al., 2005; Matthews et al., 2015). Additionally, the Drosophila Heterochomatin Genome Project has contributed to take *D. melanogaster* genome to a higher level (Hoskins et al., 2002, 2007). The annotation of transposable elements in the reference genome has not been left behind, improving after every release, including new TEs, or refining the small copy and TE nest annotations (Bergman et al., 2006; Kaminker et al., 2002).

In 2005, the genome of a second Drosophila species, *D. pseudoobscura*, was published, allowing comparative analysis between the two Drosophila species (Richards et al., 2005). Seven years after the publication of the first release of *D. melanogaster* genome, the drosophilist community took another leap into the comparative genomic era with the publication of the genomes of ten new Droso-

phila species and the comparative genomic studies between the twelve species (Drosophila 12 Genomes Consortium, 2007). These 12 genomes (*D. melanogaster*, *D. simulans*, *D. sechellia*, *D. erecta*, *D. yakuba*, *D. pseudoobscura*, *D. persimilis*, *D. willistoni*, *D. virilis*, *D. mojavensis* and *D. grimshawi*) were all sequenced with WGS and Sanger technology, although there were differences in the depth of coverage of each species.

The Drosophila 12 Genomes Consortium (2007) reported a preliminary analysis of the mobile fraction of the 12 Drosophila species. Even though several analysis of particular TE families and their presence in the 12 genomes have been published (Casola et al., 2007; de Freitas Ortiz and Loreto, 2009; Marzo et al., 2008), there are few comparative studies beyond those contained in the first publication (Feschotte et al., 2009).

In the last years, the development of Next-Generation Sequencing (NGS) techniques have drastically reduced the cost of sequencing, allowing small research groups to sequence the genomes of non-model organisms to answer particular questions. This revolution has impacted the *Drosophila genus*, with the sequencing of 16 new Drosophila genomes. Eight of those genomes, all belonging to the melanogaster group, have been jointly sequenced: *D. biarmipes*, *D. bipectinata*, *D. elegans*, *D. eugracilis*, *D. ficusphila*, *D. kikkawai*, *D. rhopaloa*, and *D. takahashii* (Chen et al., 2014). Two species, *D. albomicans* (Zhou et al., 2012) and *D. miranda* (Zhou and Bachtrog, 2012) have been sequenced to shed light into neo sex and B chromosome evolution. The genome of *D. suzukii*, has been sequenced by two independent groups because of the economical impact of this species as a fruit pest (Chiu et al., 2013; Ometto et al., 2013). Two strains of *D. americana*, H5 and W11, (Fonseca et al., 2013) and two strains of *D. buzzatii* , st-1 (Guillén et al., 2015) (see Appendix C) and j-19 (Rius et al submitted; see Section 3.2) have been sequenced to perform comparative analysis. Finally, the resequencing of *D. simulans* genome (Hu et al., 2013), previously sequenced by the Drosophila 12 Genomes Consortium Drosophila 12 Genomes Consortium (2007), was done to amend quality issues with the first assembly allowing lineage divergence studies. These 16 genomes have been sequenced with a combination of NGS (Illumina and/or 454 technologies) and in some cases with the addition of some Sanger sequences.

1.4 TE ANNOTATION AND CLASSIFICATION IN SEQUENCED GENOMES

Multiple challenges have arisen after the wave of new genomes recently sequenced. The efforts of many research groups were behind the sequencing and annotation of the first genomes. The more recent ones, on the other hand, are usually carried out by smaller groups without expertise in all the fields involved. The availability of ready-to-use TE annotation software is crucial for these smaller groups, as it is the software needed for the rest of the genome assembly and annotation process. Several authors have published reviews classifying and benchmarking the myriad of TE annotation programs available, providing a guide to chose them according to the knowledge of the genome and its repetitive fraction (Bergman and Quesneville, 2007; Lerat, 2010; Saha et al., 2008).

The annotation and classification of TEs in eukaryotic genomes requires a degree of automation to accomplish a vast and meticulous task, while at the same time manual curation is highly desirable. The *D. melanogaster* TE annotation had the advantage of an extensive TE collection maintained by FlyBase and as I mentioned above the effort of many TE experts (Bergman et al., 2006; Kaminker et al., 2002).

To analyze the TE content on the 12 genomes, as the previous knowledge of the repeat content in each of them was different, more complex strategies had to be developed. Six different combinations of TE detection methods and libraries were applied. The libraries were either previously built libraries or collections of sequences harvested from each genome. PILER (Edgar and Myers, 2005) and ReAs (Li et al., 2005) were used to build the *de novo* libraries, the last one using the unassembled reads. The already built libraries were: TE collection of Berkeley Drosophila Genome Project (BDGP), a library made with PILER scanning the 12 genomes plus the *Anopheles gambiae* genome, and the Repbase Update library (Jurka et al., 2005) without the Drosophila repeats. To annotate the TE fraction, the TE detection programs, RepeatMasker (Smit et al., 1996), BLASTER-tx, and RepeatRunner (`http://www.yandell-lab.org/repeat_runner/index.html`) were fed with these libraries and the scaffolds longer than 200 kb, the TE detection software CompTE, which do not require a library, was also used in each genome. All these strategies yielded six different results (Figure 3) and two of them (BLASTER-tx + PILER and RepeatMasker + ReAS) were finally averaged to obtain a unique figure (Drosophila 12 Genomes Consortium, 2007).

The analysis of the repetitive fraction of the genomes sequenced in the last years has not been as extensive. The two genomes with a more detailed TE

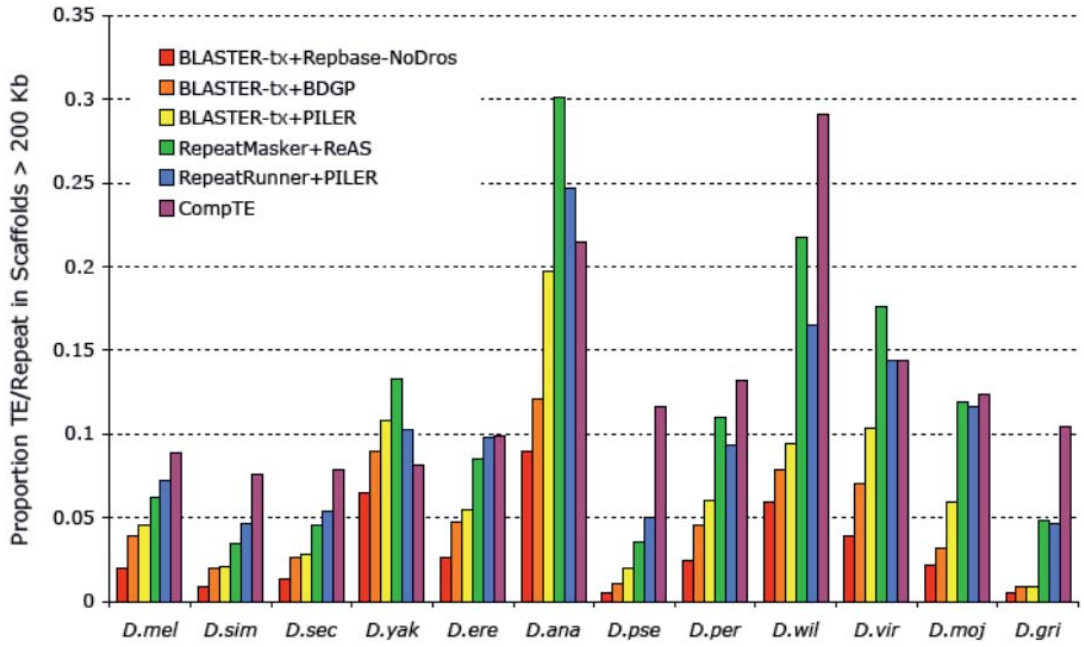

Figure 3.: Repeat and TE content of the 12 Drosophila genomes. Fraction of each genome covered by repeats based on different methods of repeat and TE annotation. Taken from Drosophila 12 Genomes Consortium (2007).

analysis are probably the two genomes of *D. suzukii* (Chiu et al., 2013; Ometto et al., 2013). Ometto and collaborators simply used RepeatMasker and Repbase library to analyze all the *D. suzukii* scaffolds. They also applied the same method and library to the rest of genomes sequenced at the publication (Figure 4). Chiu and collaborators used a less automated strategy based on BLAST (Altschul et al., 1997) searches using TEs detected in *D. melanogaster* reference genome and two TEs from *D. suzukii*.

For the *de novo* annotation of the 12 Drosophila genomes, as I mentioned above, two broad strategies were independently used; homology-based searches, that rely on libraries of already described elements from the studied species or close ones, and *de novo* strategies, that scan the genome looking for TE-like structures and repetitiveness. Nevertheless, better results are achieved if both strategies are combined, using TE detection software, like RepeatMasker, with an enhanced library, containing already known repeats, like the ones in Repbase Update, but also a custom collection made with sequences from the genome studied (Buisine et al., 2008).

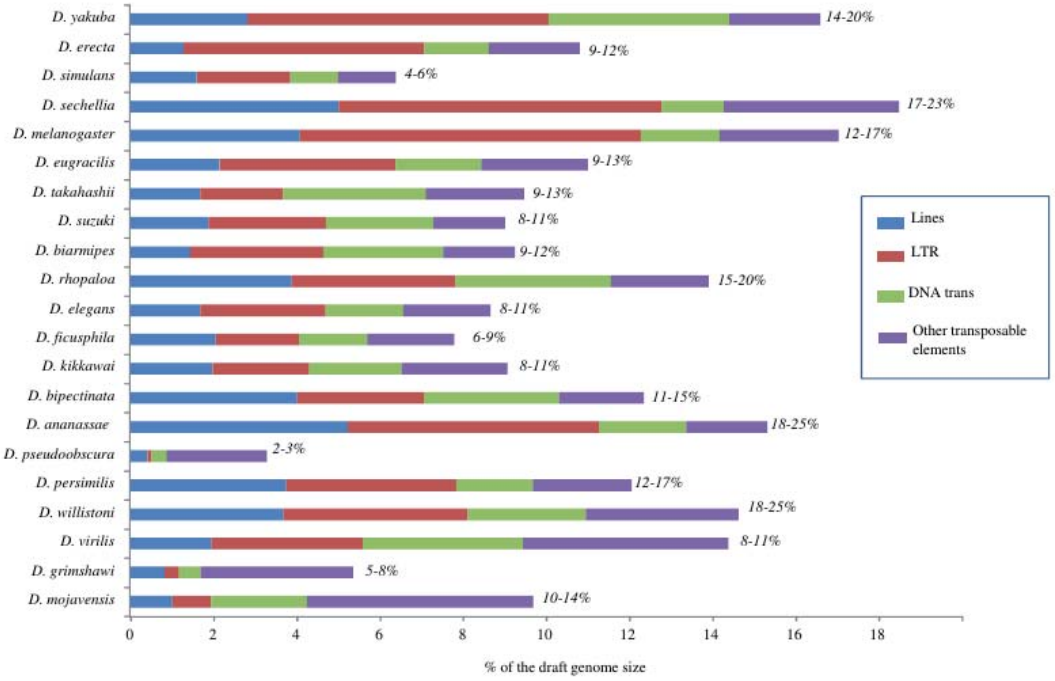

Figure 4.: Repeated elements in Drosophila sequenced genomes. Taken from Ometto et al. (2013).

Annotation pipelines like REPET (Flutre et al., 2011), Repclass (Feschotte et al., 2009), and RepeatModeler (Smit and Hubley, 2008) are able to build these custom libraries using programs that look for repetitive patterns in the genome, create groups with those repeated sequences and classify them within TE families based on homology and structural features.

2

OBJECTIVES

BuT5 was initially described as a secondary colonizer element in the proximal breakpoint of the *2j D. buzzatii* polimorphic inversion, caused by ectopic recombination between two copies of the transposon *Galileo*. *BuT5* was tentatively classified as a class II element, and named along with four other D. buzzatii transposons (*BuT1*, *BuT2*, *BuT3*, and *BuT4*), all except *BuT5* belonging to the hAT superfamily. More *BuT5* copies were found in the *D. buzzatii* polimorphic inversions $2q^7$ and $2z^3$, also the result of recombination between *Galileo* copies. However, neither of the new *BuT5* copies helped with its classification. Hidridization analyses revealed *BuT5* high abundance in different *D. buzzatii* strains. However, when the recombination between *BuT5* copies was found to be the cause of the fixed inversions *2s* in *D. mojavensis* and $2q^3$ in *D. uniseta BuT5* classification gain interest. At the same time, the project to sequence *D. buzzatii* st-1 genome lead to the opportunity to analyze the whole TE content of the species where two class II TEs causing chromosomal inversions, were described. During that process, the sequencing of the genome of another *D. buzzatii* strain, j-19, offered the opportunity to compare the TE content of both strains.

The objectives of this thesis are briefly described below

1. To study the distribution of *BuT5* in the *D. repleta* using both bioinformatic and experimental methods.
2. To isolate a copy of the autonomous element that mobilized *BuT5*.
3. To classify *BuT5* and its master TE.
4. To identify and classify the transposable elements presents in *D. buzzatii genome*.
5. To estimate the abundance of *D. buzzatii* transposable elements and compare it to that in other genomes, in particular *D. mojavensis*, the phylogenetically closest species with a sequenced genome.

6. To analyze the transposable element distribution in *D. buzzatii* among chromosomes and within chromosomal regions.

3

RESULTS

In Section 3.1, I describe the work done to characterize *BuT5*, a MITE, and its master element, the *P* element, in several Drosophila species.

In Section 3.2, I present the analysis of the transposable elements in *D. buzzatii* sequenced genomes.

3.1 A DIVERGENT *P ELEMENT* AND ITS ASSOCIATED MITE, *BUT5*

3.1.1 A DIVERGENT *P ELEMENT* AND ITS ASSOCIATED MITE, *BUT5*, GENERATE CHROMOSOMAL INVERSIONS AND ARE WIDESPREAD WITHIN THE *DROSOPHILA REPLETA* SPECIES GROUP

This Section is composed by the research article entitled "A divergent *P element* and its associated MITE, *BuT5*, generate chromosomal inversions and are widespread within the *Drosophila repleta* species group" published in the journal Genome Biology and Evolution on 2013.

A Divergent *P* Element and Its Associated MITE, *BuT5*, Generate Chromosomal Inversions and Are Widespread within the *Drosophila repleta* Species Group

Nuria Rius, Alejandra Delprat, and Alfredo Ruiz*

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

*Corresponding author: E-mail: Alfredo.Ruiz@uab.cat; Alfredo.Ruiz@uab.es.

Accepted: May 12, 2013

Data deposition: This project has been deposited at GenBank under the accession numbers KC690049–KC690135.

Abstract

The transposon *BuT5* caused two chromosomal inversions fixed in two *Drosophila* species of the *repleta* group, *D. mojavensis* and *D. uniseta*. *BuT5* copies are approximately 1-kb long, lack any coding capacity, and do not resemble any other transposable element (TE). Because of its elusive features, *BuT5* has remained unclassified to date. To fully characterize *BuT5,* we carried out bioinformatic similarity searches in available sequenced genomes, including 21 *Drosophila* species. Significant hits were only recovered for *D. mojavensis* genome, where 48 copies were retrieved, 22 of them approximately 1-kb long. Polymerase chain reaction (PCR) and dot blot analyses on 54 *Drosophila* species showed that *BuT5* is homogeneous in size and has a widespread distribution within the *repleta* group. Thus, *BuT5* can be considered as a miniature inverted-repeat TE. A detailed analysis of the *BuT5* hits in *D. mojavensis* revealed three partial copies of a transposon with ends very similar to *BuT5* and a *P*-element-like transposase-encoding region in between. A putatively autonomous copy of this *P* element was isolated by PCR from *D. buzzatii.* This copy is 3,386-bp long and possesses a seven-exon gene coding for an 822-aa transposase. Exon–intron boundaries were confirmed by reverse transcriptase-PCR experiments. A phylogenetic tree built with insect *P* superfamily transposases showed that the *D. buzzatii P* element belongs to an early diverging lineage within the *P*-element family. This divergent *P* element is likely the master transposon mobilizing *BuT5*. The *BuT5*/*P* element partnership probably dates back approximately 16 Ma and is the ultimate responsible for the generation of the two chromosomal inversions in the *Drosophila repleta* species group.

Key words: transposon, MITE, inversions, *Drosophila*, transposase, expression.

Introduction

Transposable elements (TEs) are DNA sequences able to proliferate and move to multiple sites in the genome. As a consequence of their mobility, TEs are a source of variation in gene and genome structure as well as size and organization of genomes (Kidwell and Lisch 2002). Therefore, the study of TEs can shed light on their ability to impact the genomes they inhabit (Kazazian 2004; Jurka et al. 2007; Fedoroff 2012). TEs that mobilize via an RNA intermediate are classified within class I and those which transpose directly, leaving the donor site, or via a DNA intermediate, within class II (Wicker et al. 2007; see also Kapitonov and Jurka 2008). Class II, or DNA transposons, is divided in two subclasses and subclass 1 comprises two orders, terminal inverted repeat (TIR) and Crypton. Canonical (autonomous) TIR transposons have TIRs and contain usually one (less often two) gene encoding the transposase, the protein that catalyzes their mobilization via a cut-and-paste mechanism. The numerous TIR transposon families have been grouped into 9–19 superfamilies based not only on phylogenetic relationships inferred from the transposase but also on TIR and target site duplication (TSD) features (Jurka et al. 2005, 2007; Feschotte and Pritham 2007; Wicker et al. 2007; Kapitonov and Jurka 2008; Bao et al. 2009; Yuan and Wessler 2011). The *P* superfamily comprises three transposons: *P* element (O'Hare and Rubin 1983), *1360* (also known as *Hoppel* or *ProtoP*) (Kapitonov and Jurka 2003; Reiss et al. 2003), and *Galileo* (Marzo et al. 2008).

The *P* element is a TIR transposon first discovered in *Drosophila melanogaster* (Bingham et al. 1982; Rubin et al.

1982) as the cause of the odd phenomenon of P-M hybrid dysgenesis (Kidwell and Novy 1979). The *D. melanogaster P* element is not only one of the first eukaryotic TEs to be discovered and molecularly characterized but also one of the most thoroughly studied (Rio 1991, 2002; Kidwell 1994; Engels 1996; Pinsker et al. 2001). The canonical *P* element of *D. melanogaster* is 2.9 kb in length and has 31-bp TIRs and a gene with four exons that encodes a 751 residues transposase. It also contains 11-bp sub-TIRs that act as transpositional enhancers and generates 8-bp TSD upon insertion (Rio 2002).

P-like elements are known to exist in a broad range of taxa, including protozoans such as *Trichomona vaginalis* (Kapitonov and Jurka 2009), several Dipterans (Perkins and Howells 1992; Lee et al. 1999; Sarkar et al. 2003), urochordata such as *Ciona intestinalis* (Kimbacher et al. 2009), and vertebrates (Hammer et al. 2005). In addition, *P* element has been repeatedly domesticated to generate cellular genes (Quesneville et al. 2005). For instance, the human genome contains 12 THAP-domain containing genes, and one of them (THAP9) has been recently shown to encode an active *P*-element transposase (Majumdar et al. 2013). In *Drosophila*, *P* element is widespread within the *Sophophora* subgenus (Daniels et al. 1990; Hagemann et al. 1992, 1994, 1996a, 1996b; Clark and Kidwell 1997) but seems much more scarce in the *Drosophila* subgenus (Loreto et al. 2001, 2012). An almost complete copy was isolated from *D. mediopunctata* in the *tripunctata* species group (Loreto et al. 2001), whereas relatively short fragments have been amplified by polymerase chain reaction (PCR) in other species of the *tripunctata* and *cardini* species groups (Loreto et al. 2012). The *D. mediopunctata P* element is 96.5% identical to that of *D. melanogaster,* and it has been suggested that it is the result of a horizontal transfer event (Loreto et al. 2001).

Miniature inverted-repeat TEs (MITEs) are small nonautonomous class II elements of a few dozen to a few hundred base pairs and flanked by TIRs. Their high copy numbers, homogeneous size, and high similarity within MITE families distinguish them from the typical defective nonautonomous transposons, which are usually unique copies (Feschotte et al. 2002; Guermonprez et al. 2008). Although MITEs were discovered in plants (Bureau and Wessler 1992, 1994), they have been found in a variety of organisms, including *Drosophila* (Smit and Riggs 1996; Tu 2000; Holyoake and Kidwell 2003; de Freitas Ortiz et al. 2010). Some MITEs have been found to be internal deletion derivatives of its autonomous partners and are likely to be mobilized by them (Feschotte and Mouchès 2000; Zhang et al. 2001). In other cases, however, MITEs share their terminal sequences with canonical elements, but their internal sequence does not have similarity to the master copy. The origin of these MITEs is obscure; they are the result of either profound changes in the original transposon sequence or the recruitment of unrelated sequences. As nonautonomous elements, MITEs depend on transposases encoded by canonical elements, but surprisingly MITEs can achieve higher copy numbers than their master transposons. The amplification success of MITEs has been attributed to different causes such as their promiscuity binding a range of related transposases, the gain of transposition enhancers, and loss of repressors when compared with autonomous TEs (Yang et al. 2009).

BuT5 was first described in the proximal breakpoint of a naturally segregating *D. buzzatii* inversion and tentatively classified as a class II TE (Cáceres et al. 2001). The reported copy was 1,039-bp long with 3-bp TIRs and imperfect 17-bp sub-TIRs and no coding capacity. Subsequently, similar *BuT5* copies were observed at the breakpoints of two other polymorphic *D. buzzatii* inversions (Casals et al. 2003; Delprat et al. 2009). The three inversions were caused by ectopic recombination between copies of *Galileo*, a *P*-superfamily transposon (Marzo et al. 2008), and *BuT5* was a secondary colonizer of the inversion breakpoints. The secondary colonization of breakpoints in recent polymorphic inversions (Casals et al. 2003; Delprat et al. 2009) and the relatively high abundance of *BuT5* in different *D. buzzatii* strains (Casals et al. 2006) indicate current or recent transpositional activity of *BuT5* in *D. buzzatii*. Furthermore, recent works in our group have revealed that *BuT5* generated two recently fixed inversions in two *repleta* group species, *2s* in *D. mojavensis* (Guillén and Ruiz 2012), and $2x^3$ in *D. uniseta* (Prada 2010). In both cases, each breakpoint harbors a copy of *BuT5* and the exchanged TSDs between copies of the two breakpoints denote ectopic recombination as the generation mechanism (fig. 1). Therefore, *BuT5* has had a significant role in the chromosomal evolution of the *repleta* group.

Despite *BuT5* significance as a genome reshaping force and its recent transpositional activity, this element has not been classified to date. Consequently, its phylogenetic distribution, how it mobilizes or whether it is a MITE, a deletion derivative of a known DNA transposon, or a new type of TE is still unknown. To fill this gap, our objectives were to 1) study the interspecific distribution of *BuT5* using both bioinformatic and experimental methods; 2) isolate a copy of the autonomous element that mobilizes *BuT5*; and 3) classify *BuT5* and its master TE.

Materials and Methods

Bioinformatic Searches

BuT5 bioinformatic searches were carried out using BlastN (Altschul et al. 1997) against all National Center for Biotechnology Information (NCBI) available databases (December 2012). Searches were also made using CENSOR tool (Jurka et al. 1996) and TEs deposited in Repbase Update (Jurka et al. 2005). Default parameters were used in these searches. The first copy described *BuT5*_1 (Cáceres et al. 2001), from *D. buzzatii*, was used as a query. Basic Local

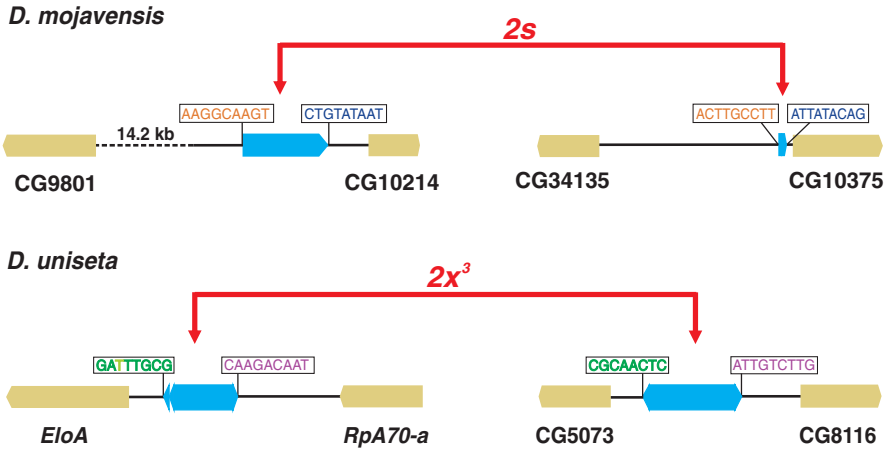

Fig. 1.—Molecular structure of breakpoint regions in two inversions generated by the transposon *BuT5*, *2s* in *Drosophila mojavensis* (Guillén and Ruiz 2012), and $2x^3$ in *D. uniseta* (Prada 2010). *BuT5* copies (blue rectangles) bounded by exchanged 8-bp or 9-bp TSD are found at the two breakpoints of each inversion, indicating ectopic recombination as the generating mechanism.

Alignment Search Tool (Blast) results in other species were also used as queries in their genomes, performing species-specific searches. Significance thresholds used to retrieve sequences for analysis were an *E* value $\leq 10^{-10}$ for results from species different of the query and $\leq 10^{-25}$ for results from the same species of the query.

Experimental Searches

Primers

BuT5 and *P*-element primers BL, BR, P3, and P13 were designed based on sequences obtained by bioinformatic searches (from *D. buzzatii* and *D. mojavensis*). PriFi (Fredslund et al. 2005) was used to find the best regions to place the primers on multiple alignments. The rest of the primers were designed based on *D. buzzatii P* element (BU-73 strain). All primers were designed with Primer Designer v.1.01 (Scientific and Educational Software) and produced by Sigma-Aldrich, Inc. Sequences of primers used are provided in supplementary table S1, Supplementary Material online.

BuT5 *Analysis in the* Repleta *Group*

To detect *BuT5,* 85 *Drosophila* DNA samples of 41 species (supplementary table S2, Supplementary Material online) were screened by PCR with primers BL and BR (supplementary table S1, Supplementary Material online). PCRs were carried out in a volume of 26.5 µl including 50–100 ng of DNA, 0.83 U of DNA Taq polymerase (Roche), 0.04 mM of each dNTP, and 0.33 µM of both primers. Amplification conditions were 94 °C for 4 min, 30 cycles at 94 °C for 30 s, 53 °C for 30 s, and 72 °C for 1 min followed by a final extension step at 72 °C for 7 min. These PCR products were purified with the NucleoSpin Extract II (Macherey-Nagel) and were cloned using either the pGEM-T Easy Vector Kit (Promega) or the StrataClone Kit (Agilent Technologies). Approximately four clones per sample were selected and PCR amplified with primers SP6 and T7 (pGEM-T) or T3 and T7 (StrataClone). The PCR products presenting different electrophoresis mobility were purified with the Nucleospin Extract II Kit and sequenced by Macrogen Inc (Seoul, Korea) using universal primers. We also analyzed by dot blot 35 DNA samples from 29 species, specified in supplementary table S2, Supplementary Material online. Denatured DNA (200 ng) was transferred onto a nylon membrane (Roche) by using a Bio Dot apparatus (Bio-Rad) according to manufacturer's specifications. The DNA was cross-linked by exposure to short-wavelength ultraviolet light. A *D. mojavensis BuT5* clone (G035_2) was used as probe. It was labeled by PCR with digoxigenin-11-dUTP (PCR DIG Labeling Mix, Roche). Final reaction volume was 50 µl, including 2.5 U of Taq DNA polymerase (Roche) and its buffer, 0.2 mM of dNTP labeling mixture, 0.5 µM of primers BL and BR, and 50–100 ng of linearized DNA. Membrane prehybridization was done in DIG Easy Hyb (Roche) and 50 ng/ml of denatured DNA, MB grade from fish sperm (Roche) at 37 °C during 1 h. Denatured probe (10 µl) was added into 3.5 ml of fresh DIG Easy Hyb, and the hybridization was performed at 37 °C for 16 h. Then two washes were done with 2× SSC and 0.1% sodium dodecyl sulphate (SDS) at room temperature and two with 0.5× SSC and 0.1% SDS at 45 °C. DIG Wash and Block Buffer Set (Roche) was used for washing and blocking incubations according to manufacturer's instructions, and detection was made with CDP-*Star* (Roche) also following the instructions. Membrane signals were quantified by Laboratori d'Anàlisi i

Fotodocumentació, d'Electroforesis, Autoradiografies i Luminiscència of the Universitat Autònoma de Barcelona with ChemiDoc XRS (BioRad) and Quantity ONE 4.7 software (BioRad).

P-*Element Sequence in* D. buzzatii

P-element amplifications, with primers BL + P13, P3 + BR, and P1 + P15 were performed with Expand Long Template in 50 μl including 50–100 ng of DNA, 1 U of Enzyme mix, 0.02 mM of each dNTP, and 0.20 μM of both primers. Amplification conditions were established following manufacturer's instructions. The BL + P13 2.8-kb band was identified and excised from a 1% agarose gel and cleaned up with the NucleoSpin Extract II Kit. The products of P3 + BR and P1 + P15 amplifications were directly cleaned up with the NucleoSpin Extract II Kit. PCR products were cloned with the StrataClone Kit (Agilent Technologies) following the manufacturer's instructions. DNA of three clones per cloning reaction was retrieved using the GeneJET Plasmid Miniprep Kit (Thermo Scientific) and finally was sequenced by Macrogen Inc (Seoul, Korea) using primers T3 and T7. The 5′- and 3′-ends of the *P* element were isolated by inverse PCR (iPCR) from *D. buzzatii* strain BU-73. Digestion (*Hind*III) and ligation were performed following Berkeley Drosophila Genome Project iPCR protocol (available from http://www.fruitfly.org/about/methods, last accessed June 2, 2013). PCR was carried out with primers InvL and InvR (supplementary table S1, Supplementary Material online) under conditions similar to those described earlier for BL + P13, P3 + BR, and P1 + P15 amplifications. A single band was identified by electrophoresis in a 1% agarose gel. The DNA was cleaned up with the NucleoSpin Extract II Kit, and cloned with the StrataClone Kit (Agilent Technologies). Minipreps of 22 clones were performed with GeneJET Plasmid Miniprep Kit (Thermo Scientific). The plasmids were used as template for PCRs with primers BL and InvR, and the clones that yielded PCR products of different length were sequenced using primers T3 and T7 by Macrogen Inc (Seoul, Korea).

Transposase Gene Exon–Intron Boundaries

Total RNA was extracted from *D. buzzatii* adult females of strain BU-73 (Berna, Argentina). Forty-five female heads and 90 ovaries were extracted in physiological solution, and RNA was obtained for each part with the High Pure RNA tissue kit (Roche) according to the manufacturer's instructions. Reverse transcriptase (RT)-PCRs were performed with the Transcriptor First-strand cDNA synthesis Kit (Roche) following the manufacturer's instructions. To favor amplification of *P* element over other transcripts, *P*-element-specific primers, P15 or P4 (supplementary table S3, Supplementary Material online), were used in two separate retrotranscription reactions. After obtaining the cDNA, five experiments, each one with two nested PCR reactions, were done to increase the amount of specific product. The combination of primers used for these PCR reactions is detailed in supplementary table S3, Supplementary Material online. These amplifications were performed in a volume of 100 μl and using 10 μl of a 1:10 dilution of the previous reaction as template, 2.5 U of DNA Taq polymerase (Roche), 0.02 mM of each dNTP, and 0.20 μM of both primers. Products of the second PCRs were cloned with the StrataClone Kit (Agilent Technologies). Screening analyses were done with Miniprep or PCR (using primers T3 and T7) on 3–47 clones per cloning reaction. Plasmids containing fragments with different electrophoretic mobility were recovered using the GeneJET Plasmid Miniprep Kit (Thermo Scientific) and sequenced by Macrogen Inc (Seoul, Korea) with T3 and T7 primers.

Sequence Analysis

Sequence analysis was performed with Geneious v5.1.3 (Biomatters Ltd.), and alignments were done with MUSCLE (Edgar 2004) through Geneious. Search for open reading frames (ORFs) with a minimum size of 100 bp was made with Geneious software. Predicted ORFs were subsequently used in BlastX searches against NCBI nonredundant protein sequences database. Gblocks (Castresana 2000) was used to select the conserved bocks of the alignment of *BuT5* sequences over 800 bp, keeping 87% of the original alignment length. To use less stringent condition, parameters were set as follows: minimum number of sequence for a flank position: 44, maximum number of contiguous nonconserved positions: 8, minimum length of a block: 5, and allowed gap position: "with half." MEGA 5 software (Tamura et al. 2011) was used to reconstruct *BuT5* phylogeny using maximum likelihood method and the best fit model according to jModelTest (Posada 2008), general time reversible model with a discrete gamma distribution (four discrete categories). Bootstrap test was performed with 1,000 replicates. The phylogeny of the *P* superfamily transposases was based on 31 putatively complete protein sequences from insects and the human THAP9 (NM_024672) and aligned with MUSCLE. *P*-like, *Galileo,* and *1360* sequences were taken from Repbase (Jurka et al. 2005) and Marzo et al. (2008). The alignment, with 1,192 positions, was used to conduct phylogenetic analyses with neighbor joining and maximum likelihood methods on MEGA 5. Bootstrap test was performed with 1,000 replicates.

P-element transposase gene introns were manually predicted using BlastX and NCBI Conserved Domains search. BlastX alignment of *D. buzzatii P*-element complete copy with *D. bifasciata* O-type *P*-element transposase (AAB31526, *E* value = 8e-91) revealed discontinuities coincident with stop codons and frameshift mutations. NCBI Conserved Domains search tool (Marchler-Bauer et al. 2011) provided information regarding which virtually translated residues were part of transposase domains. BlastN searches were also used to refine the first predictions by comparing the transposase generated with other *P*-element transposases.

Results

BuT5 Bioinformatic Searches

We carried out BlastN (Altschul et al. 1997) searches using as query *BuT5-1* (Cáceres et al. 2001) against all NCBI nucleotide databases (including 2,428 bacterial, 122 archaeal, and 426 eukaryotic genomes). We retrieved 36 previously published *D. buzzatii BuT5* sequences (supplementary table S4, Supplementary Material online) plus 48 new *BuT5* sequences from the genome of *D. mojavensis* (supplementary table S5, Supplementary Material online), a relative of *D. buzzatii* that belongs to the *repleta* group (*Drosophila* subgenus). No hits were significant in any of the other *Drosophila* genomes or the other genomes searched. In addition, no results were recovered from searches in Repbase Update (Jurka et al. 2005).

Only two other sequences from *D. buzzatii* had a size similar to that of *BuT5-1* (1,039 bp), the rest being fragments less than 800-bp long likely resulting from deletions. The three longest copies have 3-bp TIRs, imperfect (two mismatches) 17-bp sub-TIRs (fig. 2), and TSD 8-bp or 9-bp long (Cáceres et al. 2001; Casals et al. 2003; Delprat et al. 2009). Twenty-two out of the 49 *BuT5* copies retrieved from the *D. mojavensis* genome were over 800-bp long (mean ± standard deviation [SD] = 1,017.4 ± 23.2) and had a pairwise identity of 93.2%. Fifteen of them had both 3-bp TIRs, and 14 had 16-bp imperfect (two mismatches) sub-TIRs (fig. 2). Seventeen *D. mojavensis BuT5* copies were flanked by TSDs: 4 8-bp long and 13 9-bp long (one has two mismatches). The *BuT5* consensus sequence of *D. mojavensis*, built with the 22 longer copies, has 67.3% pairwise identity to the *BuT5* consensus sequence of *D. buzzatii*, built with the three longer copies previously isolated. However, the identity between *BuT5* consensuses of both species is higher at the terminal regions (fig. 2), where the first 65 bp shows 90.8% identity and the last 32 bp, 90.6%. This suggests that the size and the terminal features of *BuT5* are particularly conserved between *D. mojavensis* and *D. buzzatii* copies.

BuT5 Experimental Searches

A pair of degenerated primers, BR and BL, was designed to match *BuT5* ends, which are conserved between *D. buzzatii* and *D. mojavensis*, to increase the chances of successful interspecific amplification. PCR screening was done with 85 DNA samples of 41 species from the *Drosophila repleta* species group (supplementary table S2, Supplementary Material online). PCR products were cloned and sequenced and 86 clones from 26 species were confirmed as *BuT5* copies. However, as the primers were inside the element, some features such as TIRs or TSDs could not be retrieved from these copies. Sequences over 800 bp (61 from 19 species) had a mean size (±SD) of 959.2 bp (±46.8) that amounts to 1,014.2 bp if the unsequenced element ends are taken into account. To complement the PCR search, a dot blot analysis was carried out with 20 PCR-negative *repleta* group samples (15 species) plus samples from *D. nannoptera* and *D. wassermani*, two species in the cactophilic *nannoptera* species group (Pitnick and Heed 1994), and samples from *D. buzzatii* and 12 species with available genome sequences as controls (supplementary table S2, Supplementary Material online). Dot blot confirmed as negative three species of the *repleta* group (*D. hydei*, *D. nigrospiracula*, and *D. pegasa*) but yielded positive for the other 12 PCR-negative species. Results were also negative for the two species of the *nannoptera* group and for all species with sequenced genome except *D. mojavensis*.

In summary, *BuT5* was detected, either by PCR or dot blot, in 38 of the initial 41 species of *repleta* group, belonging to four of the six described subgroups (samples were not available for subgroups *fasciola* and *inca*) (fig. 3). *BuT5* is present in most lineages, including the most basal branch of the *repleta* group (*D. eremophila* and *D. mettleri*), estimated to

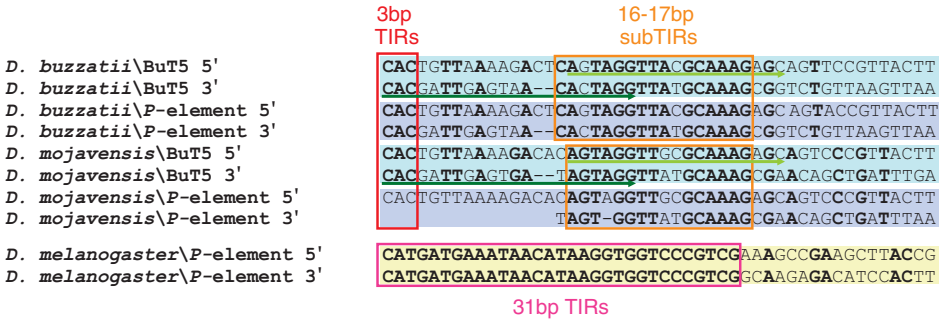

Fig. 2.—Alignment of the 5′- and 3′-terminal regions of *BuT5* and *P* element from *Drosophila buzzatti* and *D. mojavensis*. For comparison, the *D. melanogaster P*-element terminal sequences are included but not aligned. The red box indicates *BuT5* and *P*-element TIRs, the orange box *BuT5* and *P*-element sub-TIRs, and the pink box the *D. melanogaster P*-element TIRs. Green arrows indicate the primers BL (dark green) and BR (light green).

have shared their last common ancestor 16 Ma (Oliveira et al. 2012).

Analysis of *BuT5* Sequences

As a result of the bioinformatic and experimental searches, we retrieved 86 *BuT5* sequences over 800-bp long from 19 species. These were used to build a phylogenetic tree using maximum likelihood methods (fig. 4). The *BuT5* sequence recovered from *D. nigricruria* was used to root the tree because this species is the most distant one and does not belong to the *mulleri, longicornis*, or *buzzatii* complexes. The *BuT5* phylogenetic tree (fig. 4) is broadly concordant with that of the host species (fig. 3) and mirrors the relationship between the *mulleri, longicornis,* and *buzzatii* complexes (Wasserman 1992; Ruiz and Wasserman 1993; Oliveira et al. 2005), yet sequences of the *longicornis* complex do not form a monophyletic cluster. Consequently, *BuT5* has been vertically transmitted, and no clear-cut evidence for horizontal transfer was found.

We estimated the age of the *BuT5* copies in *D. mojavensis* with the formula $t = K/r$ (Kapitonov and Jurka 1996), where K is the average divergence of the copies from their consensus sequence and r the neutral substitution rate (0.0111 substitutions per bp per Myr; Tamura 2004). For the 22 most complete *BuT5* copies isolated from the *D. mojavensis* genome, $K = 0.0267$ and $t = 2.4$ Myr. However, there is evidence for more recent transposition events. For a subset of five closely related copies, $K = 0.004$ and $t = 0.36$ Myr.

We searched putative ORFs in all *BuT5* copies by several methods. ORF longer than 100 bp showed no similarity to previously described proteins, corroborating that *BuT5* has no coding capacity (Cáceres et al. 2001). Furthermore, most *BuT5* copies had a size similar to that of the original *BuT5*-1 copy (~1 kb) or were smaller (partial copies). The TIR, lack of coding capacity, abundance, and homogeneous size of *BuT5* allow us to consider it tentatively as a MITE.

On the other hand, similarity searches with *BuT5* in the *D. mojavensis* genome revealed two nearby significant hits in scaffold_6541 spaced by approximately 3 kb. When the intervening sequence was explored using BlastX against protein databases, we found a significant similarity to the transposase of *P* element (O-type) from *D. bifasciata* (AAB31526.1, amino acid identity 47%, *E* value: 3e-91). The total sequence, including the terminal segments similar to *BuT5*, was 3,221-bp long. This sequence was used to search against the *D. mojavensis* genome with BlastN, finding two other sequences with similarity to the *P* element (supplementary table S6, Supplementary Material online). The consensus of the three copies has a size of 3,254 bp and shows similarity to the *BuT5* ends only (fig. 5). We hypothesized that this *P* element could represent the autonomous transposon family mobilizing *BuT5*.

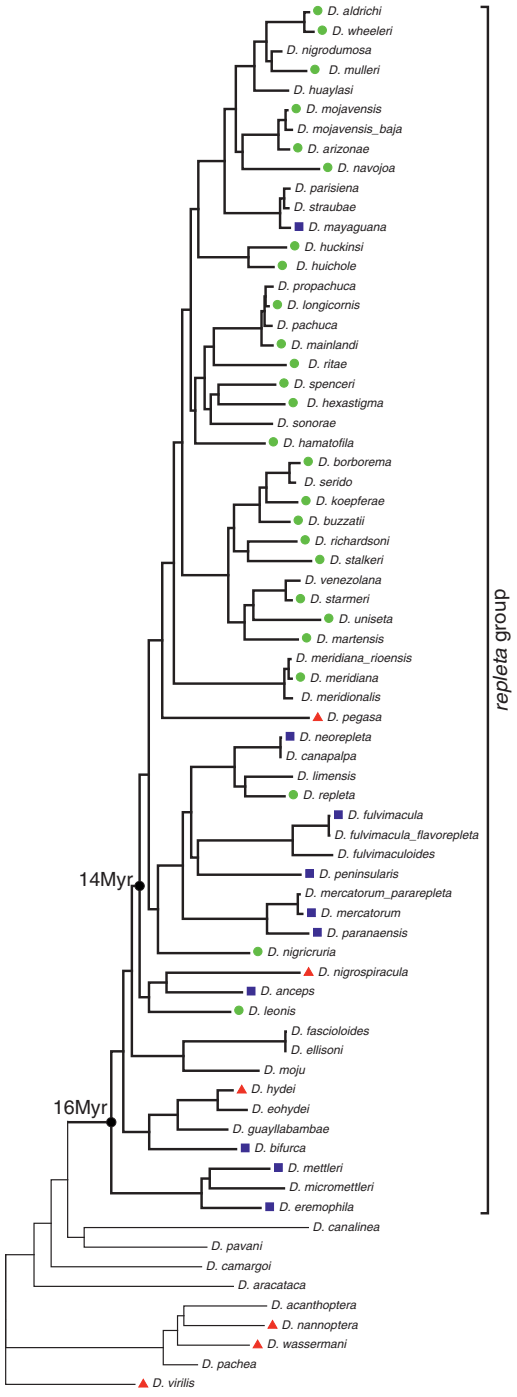

Fig. 3.—Distribution of the transposon *BuT5* plotted onto the *repleta* group phylogeny (taken from Oliveira et al. 2012). Green dots denote species with *BuT5* sequences recovered by PCR; blue squares and red triangles indicate positive and negative results for dot blot, respectively.

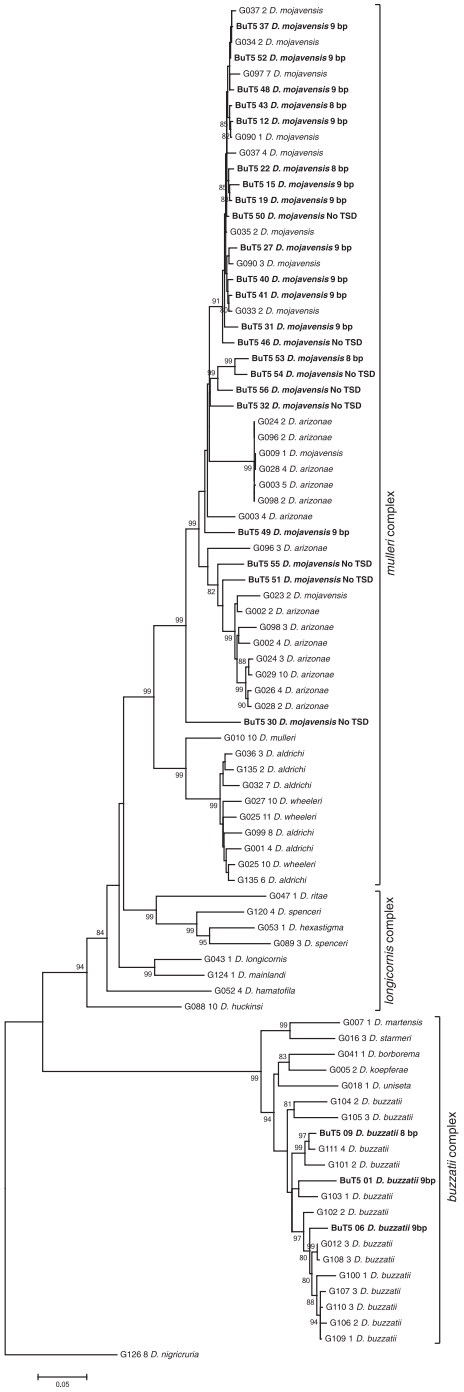

Fig. 4.—Maximum likelihood phylogenetic tree built with 86 *BuT5* sequences longer than 800 bp, recovered by PCR (sample code, clone number, and species name) or bioinformatic searches (boldface; copy number, species name and TSD length) from 19 species of the *repleta* group. Bootstrap values over 80 are shown at nodes.

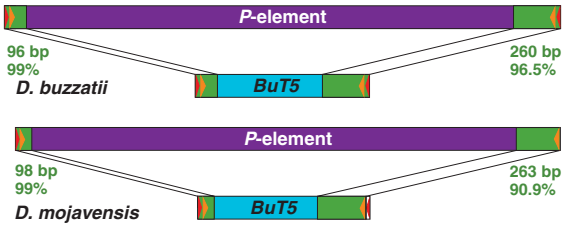

Fig. 5.—A comparison of *P* element and *BuT5* structures in *Drosophila buzzatii* and *D. mojavensis*. In both cases, the two elements show high similarity at both ends (green) but no similarity in the interior segment (purple or blue). Red and orange triangles denote the TIRs and sub-TIRs, respectively.

Isolation and Characterization of *P* Element in *D. buzzatii*

To test the hypothesis that the *P* element detected in *D. mojavensis* represents the master copy mobilizing *BuT5*, we searched for similar elements in 42 strains of *D. buzzatii* from different geographical origins, 7 strains of *D. mojavensis,* and 1 strain of *D. uniseta*. Two overlapping PCR reactions were carried out with primers BL + P13 and P3 + BR (fig. 6), because primers BL + BR placed at the ends of *BuT5* retrieve only copies of this MITE (see earlier). Only one *D. buzzatii* strain (BU-73 from Berna, Argentina) produced positive results for both reactions. The two sequences were 2,694-bp and 2,389-bp long and overlapped 1,755 bp. An additional PCR with primers P1 + P15 (fig. 6) generated a 2,583-bp product that covers the central part of the element. Finally, both ends of the *P* element were isolated by iPCR with primers InvL + InvR (fig. 6). Only one band, approximately 6.3-kb long, was observed. The retrieved sequences overlap 362 bp at the 5′-end and 637 bp at the 3′-end with the other *P*-element sequences and are 100% identical to them in the overlapping segments. The complete *P* element in *D. buzzatii* is 3,386-bp long and is flanked by 9-bp TSDs (CTAGTAGGT). This *P* element and *D. buzzatii BuT5* consensus sequence show 99% identity over the first 96 bp at the 5′-end and 96.5% over the last 260 bp at the 3′-end (fig. 5).

On the other hand, only one strain of *D. mojavensis* (G091) was positive for one of the reactions (BL + P13). We built a consensus sequence of 3,260 bp for *D. mojavensis P* element using the three copies from the sequenced genome (supplementary table S6, Supplementary Material online) and the 2,686-bp sequence recovered by PCR. This consensus sequence has 99% identity to *BuT5 D. mojavensis* consensus over 98 bp at the 5′-end and 90.9% over 263 bp at the 3′-end (fig. 5). The highly congruent observations in *D. buzzatii* and *D. mojavensis* support our hypothesis.

A BlastX search with the *D. buzzatii P* element against the NCBI protein database corroborated the similarity with *D. bifasciata* O-type *P*-element transposase (AAB31526, amino acid identity 34%, *E* value = 8e-91), but the BlastX alignment showed clear-cut discontinuities. A tentative

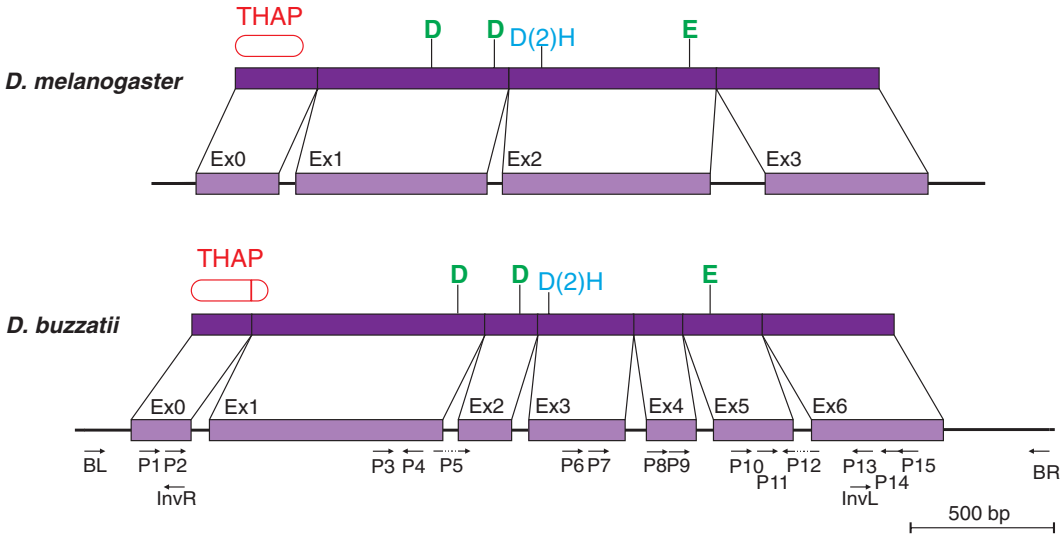

Fig. 6.—Comparison of the transposase gene in *Drosophila buzzatii P* element (bottom) with that of the *D. melanogaster* canonical *P* element (top). The exon–intron structure of the *D. buzzatii* gene was corroborated by RT-PCR experiments using the primers shown below the gene. The structure of the protein is also shown with the DNA-binding THAP domain (Roussigne et al. 2003; Clouaire et al. 2005) and the catalytic motives DDE and D(2)H (Yuan and Wessler 2011) highlighted in each case.

manual annotation using BlastX hits and the results of a search against NCBI Conserved Domain Database predicted a transposase-coding gene comprising eight exons and seven introns that encoded a protein of 761 residues. The subsequent expression experiments (see later) corroborated six of the seven introns (but with modifications) and rejected one of them (that keeps the reading frame and does not contain STOP codons). Therefore, the transposase gene of the *D. buzzatii P* element comprises seven exons and encodes an 822-aa protein.

Expression Analysis

RT-PCR experiments using diverse primers (fig. 6) were carried out to assess the expression of the *P* element and test the manual annotation of the transposase gene. Ovaries and female heads were used to extract total RNA. Then single-stranded cDNA was generated from the two RNA samples using one of two *P*-element-specific primers (P4 or P15). Finally, five amplification reactions were performed to amplify several TE fragments enclosing the predicted exon–intron junctions (supplementary table S3, Supplementary Material online). These PCRs were carried out in two consecutive rounds with the primers of the 2nd round designed within the product of the 1st round. PCR products were cloned, and a screening of the clones was performed through PCR or Miniprep. Those clones with a different electrophoretic mobility were sequenced (supplementary table S3, Supplementary Material online).

Positive expression results were produced in heads in four of the five amplification reactions. Two different products, with and without intron 1, were recovered from the first PCR with primers P2 + P4 (fig. 6). In the second PCR with primers P5 + P12, both located in exon–exon junctions, a product with introns 2, 3, 4, 5, and 6 spliced was produced (fig. 6). The third PCR with primers P7 + P12 gave rise to three differently spliced products, one keeping introns 4 and 5, one without introns 4 and 5, and one without intron 4 and a 3′ alternatively spliced intron 5. The fourth PCR with primers P9 + P12 was negative in heads (but positive in ovaries, see later). Finally, the fifth PCR with primers P11 + P14 gave a single product retaining intron 6.

In ovaries, positive expression results were produced in three of the five amplification reactions. A single product was obtained in the first PCR (P2 + P4) that kept intron 1. For the second PCR (P5 + P12), no products were obtained from ovarian samples. The third PCR (P7 + P12) did not produce any products. In the fourth PCR (P9 + P12), only one product with intron 5 present and intron 6 spliced was obtained. The fourth reaction (P11 + P14), in ovaries, gave rise to two different forms, one retaining exon 6 and the other without this last intron.

In other words, we have detected splicing of all six introns in heads (somatic tissue) but not in ovaries (somatic and germline cells), where only the splicing of the last intron was detected.

Phylogenetic Analysis

To place our *D. buzzatii P* element in the context of other *P* elements from *Drosophila* and other Dipterans, an

alignment was made with the nucleotide sequences of the *P*-element fragments generated by Clark and Kidwell (1997) and Loreto et al. (2012). The nucleotide identity between our *P* element and those previously isolated from *Drosophila*, Scaptomyza, *Lordiphosa,* and *Lucilia* genera is very low, and the alignment was not very reliable. When a phylogenetic tree (not shown) was generated with these sequences, the *D. buzzatii P*-element branch diverged early and was well separated from all other *P*-element sequences. Given the high nucleotide divergence observed, we turned to the protein phylogeny to compare our *P* element with all members of the *P* superfamily.

The complete 822-aa predicted transposase from the *P* element of *D. buzzatii* was used in a phylogenetic analysis that included 30 other *P*-like transposase sequences from insects (*Drosophila*, *Aedes*, *Anopheles*, *Acyrthosiphon,* and *Scaptomyza* genera), and the human THAP9 protein (Majumdar et al. 2013) was used as outgroup. The alignment of the 32 transposases is provided in supplementary figure S1, Supplementary Material online. Phylogenetic trees with the same topology were generated both with the neighbor joining method (fig. 7) and the maximum likelihood method (not shown). The tree presents three major monophyletic clades, grouping each of the families described in the *P* superfamily, namely *Galileo*, *1360,* and *P* element. The predicted transposase for *D. buzzatii* falls into the well-supported *P* family clade, unambiguously placing our TE within this family.

Discussion

A Divergent *P* Element in the *D. repleta* Species Group

We have identified and fully characterized a *P* element in *D. buzzatii*, a member of the *repleta* species group of the subgenus *Drosophila*. This element has also been found (but only partially characterized) in *D. mojavensis,* another member of the same species group. These two species diverged approximately 12 Ma (Oliveira et al. 2012) suggesting that this transposon is widespread within the *repleta* group (see later). The *P* element that we have characterized is very divergent from all the previously detected *P* elements in *Drosophila,* but phylogenetic analyses (fig. 7) indicate that it should be placed in this family rather than in the other two closely related families (*1360* and *Galileo*) of the *P* superfamily. This finding suggests that *P* elements may be more widespread within the *Drosophila* genus than previously thought and have gone undetected in previous bioinformatic or experimental searches due to their divergence from canonical *D. melanogaster P* element (Loreto et al. 2012).

The *D. buzzatii P* element is divergent from other *Drosophila P* elements in three regards: 1) nucleotide and protein sequence; 2) transposase-encoding gene structure; and 3) expression pattern.

The nucleotide sequence of *D. buzzatii P* element is quite dissimilar from that of *D. melanogaster P* element and all the previously described *P* elements in *Drosophila*. In DNA transposons, terminal regions bear important features for their mobilization and are accordingly evolutionarily conserved. *Drosophila melanogaster P* element has TIRs of 31-bp and sub-TIRs of 11-bp (that overlap the THAP-domain-binding sites). The TIRs of *P* elements in *D. bifasciata* (M and O types) and *Scaptomyza pallida* are 31- or 32-bp long, respectively (Hagemann et al. 1994). Similarly, the TIRs of *Anopheles gambiae* autonomous *P* element range from 27 to 31 bp (Quesneville et al. 2006). In contrast, the *D. buzzatii* and *D. mojavensis P* elements have 3-bp TIRs and 16–17 bp sub-TIRs (fig. 2). The distance between the beginning of the TIR and the end of the sub-TIR in *repleta* group *P* elements is 32 bp at the 5′-end and 30 bp at the 3′-end. Therefore, it seems that this segment is equivalent to the 31 bp of *D. melanogaster P*-element TIRs. We did not find an equivalent for the *D. melanogaster P*-element 11-bp sub-TIRs in the *repleta* group *P* element.

The *D. buzzatii P* element encodes a transposase with 822 residues. This transposase shows a 34.8% identity with that of the *D. melanogaster P* element and comparable identity values with other *Drosophila P*-element transposases. The highest identity seems to be that with the *Aedes aegypti* P-1 element transposase (36.9%). These data emphasize the high divergence between the *repleta P* element and all other described *P* elements. However, the *repleta P*-element transposase contains the typical protein domains of *P*-transposases, N-terminal DNA-binding domain, and C-terminal catalytic domain (fig. 6). The *D. melanogaster P*-element transposase contains a zinc-dependent DNA-binding domain evolutionarily conserved in an array of different cellular proteins and named THAP domain (Roussigne et al. 2003; Clouaire et al. 2005). It includes a metal-coordinating C2CH signature plus four other residues (P, W, F, and P) that are very conserved as well as eight residues that make contact with the double-helix major groove (M, Y, L, H, N, and Q) and the minor groove (R and R) (Sabogal et al. 2010). We searched in the *D. buzzatii P*-element transposase for domains using conserved domains search (Marchler-Bauer et al. 2011) and found a significant match (3.4E-21) with the THAP domain (PF05485) in the N-terminus (positions 1–91). The *D. buzzatii P*-element THAP domain contains the eight conserved residues in positions C3, C8, P23, W32, C51, H54, F55, and P82. However, only three of the eight residues that make contact with the double-helix seem to be present (M1, N47, and R72). This is not unexpected because there is variability in the residue composition of the THAP domain (Sabogal et al. 2010). The *D. buzzatii P*-element transposase contains also a putative catalytic domain in the C-terminus with the DDE triad and D(2)H motif (Yuan and Wessler 2011) conserved in positions D313/D386/E610 and D419(2)H422, respectively (fig. 6).

27

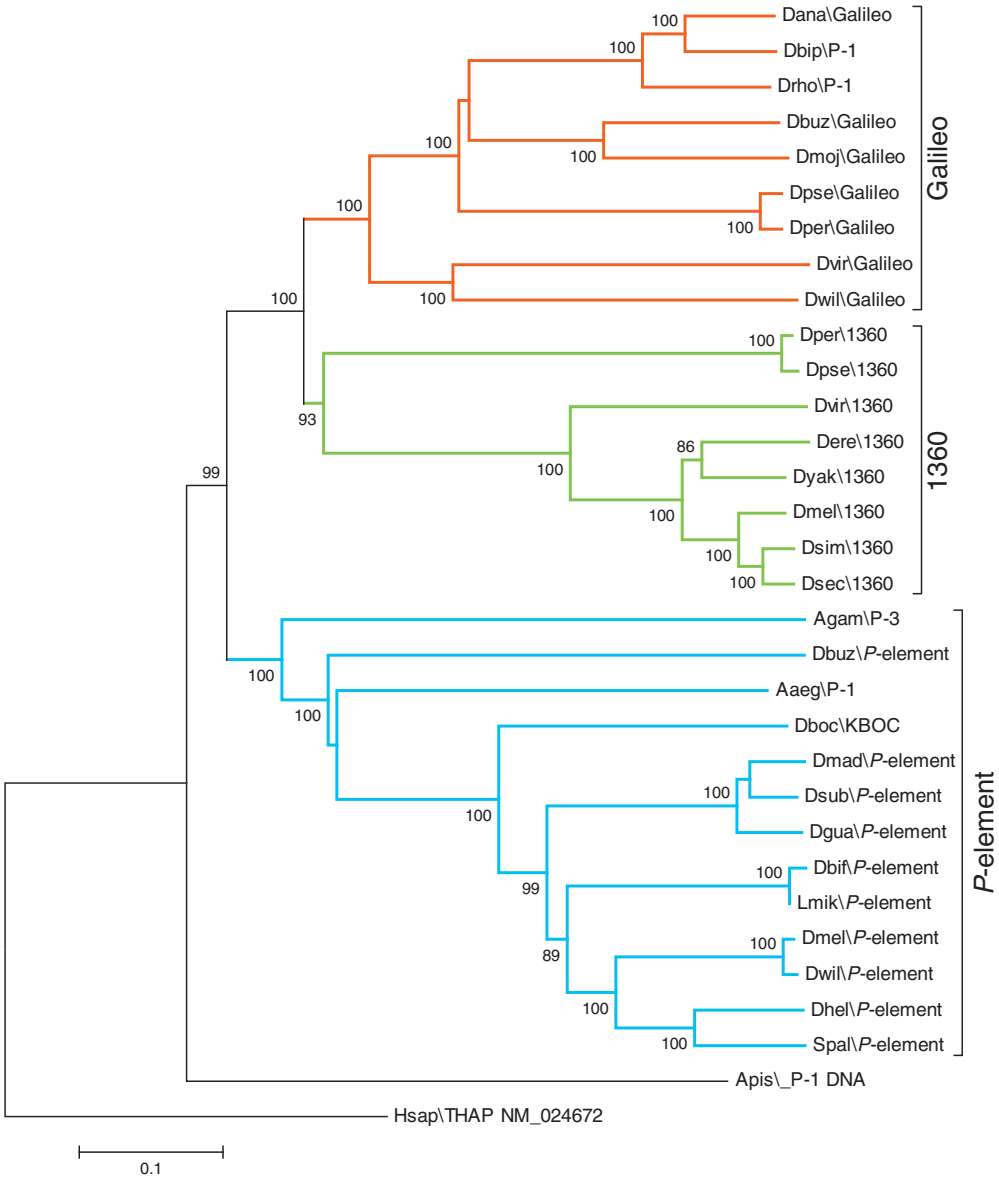

Fig. 7.—Neighbor joining phylogeny built with an alignment of 31 *P*-element transposases from insects and the human THAP9 protein as outgroup. *Galileo* family is highlighted in red, *1360* in green, and *P* element in blue. Bootstrap values over 80 are shown.

The THAP domain of *D. melanogaster P*-element transposase can bind to four experimentally verified binding sites within the *P*-element sequence, two in the 5′-end, and two in the 3′-end (table 1). Seemingly, the transposase THAP domain is able to recognize one binding site on each side, whereas the repressor THAP domain can recognize the four THAP-domain-binding sites (Kaufman et al. 1989; Lee et al. 1996). Because the *D. buzzatii* transposon belongs to the *P*-element family, and its transposase encodes a THAP domain, we searched for THAP-domain-binding sites in the *P*-element terminal regions similar to the consensus for THAP-domain-binding sites (Sabogal et al. 2010). We found in the *D. buzzatii P* element two sequences (one on each end) identical to two of the *D. melanogaster P*-element-binding sites, although they are placed in reverse orientation (table 1). We did not find any other putative THAP-domain-binding

28

Table 1
Transposase-Binding Sites in *P* Element

Start	THAP Domain Binding Site	End	TE	Region
D. melanogaster				
61	TAAGTGTA	54	*P* element	5′-end
2,859	TAAGTGGA	2,866	*P* element	3′-end
136	TAAGGGTT	129	*P* element	5′-end
2,763	TAAGGGTT	2,777	*P* element	3′-end
D. buzzatii				
90	TAAGTGTA	97	*P* element	5′-end
3,185	TAAGGGTT	3,178	*P* element	3′-end
90	TAAGTGTg	97	BuT5	5′-end
841	TAAGGGTT	834	BuT5	3′-end

Note.—Four experimentally verified naturally occurring binding sites for the *P*-element transposase in *Drosophila melanogaster* (Kaufman et al. 1989; Lee et al. 1996; Sabogal et al. 2010). Putative binding sites observed in *D. buzzatii P* element and *BuT5*. Each of the *D. buzzatii* sequences is identical to one of the *D. melanogaster*-binding sites except for a mismatch in the last nucleotide of the *BuT5* 5′-end binding site (lowercase). Coordinates are given in 5′→3′ orientation for *D. melanogaster P* element (O'Hare and Rubin 1983) and for *D. buzzatii P* element (this work) and *BuT5* (Cáceres et al. 2001).

sites, and therefore, if there are other binding sites in the *D. buzzatii P* element, they must be divergent from the *D. melanogaster* consensus.

The structure of the transposase encoding gene in the *repleta P* element is quite unusual, with seven exons (0–6) and six introns (1–6). The transposase-encoding gene in *D. melanogaster P* element has four exons (0–3) and three introns (IVS1-3). This structure seems conserved in other *Drosophila P* elements (Haring et al. 1998). *Anopheles gambiae P* elements seem to have three or four exons different from those in *D. melanogaster* (Quesneville et al. 2006). The locations of the *repleta P*-element introns along the transposase sequence are different from those of the *D. melanogaster P*-element introns (i.e., none of them coincide; see fig. 6) and also different from those of *A. gambiae*. Thus, they seem evolutionarily independent. Because *Galileo* and *1360*, the other two members of the *P* superfamily (fig. 7), do not contain introns (Marzo et al. 2008), it seems reasonable to assume that the ancestor of the *P* superfamily transposons was an intronless element that has acquired different introns along the several branches of the *P*-element phylogeny (fig. 7). Introns can be acquired via several different mechanisms although "intron duplication" has long been favored as the most likely source of new spliceosomal intron positions (Rodríguez-Trelles et al. 2006).

P-element expression has been extensively studied, and several regulatory mechanisms have been described (for reviews see Rio 1991; Castro and Carareto 2004). In *D. melanogaster*, transposase is only produced in the germline. In the somatic tissue, the *P*-element transcripts are not completely spliced, retaining the third intron (IVS3), which possess a termination codon in the first 9 bp. The resultant truncated protein of 66 kDa acts as a repressor of the transposase excision activity. This form was also found in other species (Haring et al. 1998) and proposed to occur irrespective of the *P*-element type and host species. However, this regulation mechanism, or a similar one based on the retention of specific introns, is not applicable to *D. buzzatii P* element. *Drosophila buzzatii P* elements not only have a different exonic structure but we have also detected splicing of all introns in the somatic tissue (head). Thus, seven-exon *P* element shows differences to the other *Drosophila P* elements also regarding regulation.

BuT5 Is a MITE Associated with the *P* Element

MITEs are short nonautonomous elements with TIRs, no coding capacity, and capable of reaching high copy numbers in plant genomes (Feschotte et al. 2002). Similar to nonautonomous elements, MITEs transpose using transposases encoded by autonomous elements. A distinctive feature of MITEs is their homogeneity in size and sequence, which differentiates them from typical nonautonomous elements, which are usually unique defective copies (Guermonprez et al. 2008). Moreover, some MITEs are not just deletion derivatives of complete transposons and unlike other nonautonomous TEs have internal sequences unrelated to their master copies (Feschotte et al. 2003). *BuT5* is a relatively short repetitive sequence without coding capacity; it has very short TIRs and is flanked by TSDs. Given this and the remarkably high identity between *BuT5* and the *repleta P* element at the 5′- and 3′-ends and the lack of similarity of the internal sequences (fig. 5), it is reasonable to consider *BuT5* as a MITE, probably mobilized by the *P*-element transposase.

The size of *BuT5* copies seems fairly homogeneous. In the genome of *D. mojavensis,* we retrieved 22 *BuT5* copies with a mean size of approximately 1 kb and a 93.2% pairwise identity. The rest of copies were smaller and likely to bear partial deletions. The size of *BuT5* in *D. buzzatii* is similar, as it is that of copies isolated from other species of the *repleta* group (see earlier). *BuT5* is also quite abundant. In *D. buzzatii, BuT5* was found (using in situ hybridization to polytene chromosomes) to be the most abundant of a set of seven transposons, with a basal density of 10^{-2} copies per genome and chromosomal band (Casals et al. 2006). *BuT5* is particularly abundant as secondary colonizer of the inversion breakpoints, indicating that it is (or has been until very recently) transpositionally active (Delprat et al. 2009). It is true that these copy numbers are not close to the thousands of MITEs detected in plant genomes (Feschotte et al. 2002). However, in *Drosophila,* MITEs are not as abundant as in plants (Holyoake and Kidwell 2003; Dias and Carareto 2011; Depra et al. 2012), possibly because the genome size is considerably smaller and autonomous TEs are not as abundant either (Bartolomé et al. 2002; Tenaillon et al. 2010).

The terminal sequences of *BuT5* and *P* element show high nucleotide identity (>90%), whereas their internal sequences do not have any similarity (figs. 2 and 5). Similarity between

terminal sequences of a transposon and a MITE has been previously found and seemingly indicates that the autonomous element is responsible for the MITE origin and amplification because these sequences are known to be important for transposition (Turcotte et al. 2001; Feschotte et al. 2002, 2003). Thus, we can hypothesize that the *P* element is the autonomous transposon responsible for *BuT5* mobilization. If *BuT5* were mobilized by the *P*-element transposase, we would expect that it contains THAP-domain-binding sites. We searched for such motifs in the three *D. buzzatii BuT5* longest sequences (copies 1, 6, and 9 in supplementary table S4, Supplementary Material online). All three contain in the 3′-end an identical sequence and in the 5′-end a nearly identical one (one mismatch) to one of the *D. melanogaster* THAP-domain-binding sites (table 1). Significantly, in the *D. buzzatii P* element, the putative 5′-THAP-domain-binding site is located in the limit of the segment conserved between *BuT5* and *P* element, whereas the 3′-binding site is embedded within the conserved segment (fig. 5). That is, the similarity of *BuT5* and *P* element in the 5′-end is lost precisely after the putative THAP-domain-binding site. These observations provide strong support for our hypothesis.

The size of the TSDs is a function of the transposase that catalyzes the element mobilization (Craig 2002) and is one of the key features that characterize each superfamily, yet several superfamilies have variable TSD size (Wicker et al. 2007). The *D. melanogaster P* element generates 8-bp TSDs (O'Hare and Rubin 1983) and the other two families of the *P* superfamily, *1360* (Kapitonov and Jurka 2003; Reiss et al. 2003), and *Galileo* (Cáceres et al. 1999), generate 7-bp TSDs. In *D. melanogaster*, the length of *P*-element TSDs is highly conserved; only 2% of TSDs from natural insertions were not 8-bp long (Liao et al. 2000; Linheiro and Bergman 2008, 2012). In contrast, a great number of *P*-element insertions have been studied, and the conservation of the TSDs and target site motives sequence (TSMs), although significant, is low when compared with consensuses for TSDs and TSMs of other TE families.

Two different lengths of *BuT5* TSDs (8 and 9 bp) were previously recovered from *D. buzzatii* (Cáceres et al. 2001; Delprat et al. 2009) and now have been detected flanking highly similar copies in the *D. mojavensis* genome (4 are 8 bp and 13 are 9 bp). The only *P*-element copy isolated from *D. buzzatii* is flanked by 9-bp TSDs (CTAGTAGGT). Although some superfamilies have variable TSD size, each element family usually has a single TSD size. However, there are exceptions. For instance, both the prokaryotic insertion sequence *ISRm3* and the maize element *Popin* show TSDs of 8 or 9-bp (Wheatcroft and Laberge 1991; Rhee et al. 2009). We consider the two sizes of TSDs are more likely the product of transposase flexibility in the staggered double-strand break (DSB) rather than the result of cross-mobilization events (Yang et al. 2009) that would imply the maintenance of *BuT5* and two transposases over at least 12 Myr. Because 9-bp TSDs are shared between *BuT5* and *D. buzzatii P* element, we consider these results as consistent with the notion that *P* element mobilizes *BuT5*.

Evidence has been provided supporting *BuT5* recent mobilization and the implication of *P*-element transposase in this process. The finding of a putatively complete *P*-element copy in *D. buzzatii* that is transcribed and spliced suggests *P*-element activity. However, an autonomous transposon will need to transpose in the germline, and we found splicing of all six introns in female heads but not in ovaries. This observation does not necessarily imply that *P* element is not active in *D. buzzatii* as transposition could be restricted to the male germline or to other developmental stages different from those we studied here. A more thorough expression analysis is required to solve this question.

BuT5 and the *P* Element, 16 Myr Partnership

We have found that the MITE *BuT5* is widespread in the *Drosophila repleta* species group (fig. 3) and has most likely been vertically transmitted (fig. 4). A recent and comprehensive analysis proposed that diversification of the main *repleta* group lineages occurred approximately 16 Ma (Oliveira et al. 2012). Therefore, *BuT5* would be at least 16 Myr old. In previous works, *BuT5* was detected as a secondary colonizer in very recent chromosomal inversion breakpoints (Delprat et al. 2009). Additionally, in this work, we have found *BuT5* copies from the *D. mojavensis* sequenced genome with noteworthy similarity (99.3%). Both findings reveal that *BuT5* has been recently active in two species that diverged approximately 12 Ma (Oliveira et al. 2012). Because *BuT5* is a nonautonomous element that cannot move by itself but requires the transposase of the *P* element, we can infer that the *P* element has been recently active in these two species, *D. buzzatii* and *D. mojavensis*. Similarly, the presence and conservation of *BuT5* ends in many species of the *repleta* group indicates a widespread distribution of the *P* element within this species group. Most likely, the partnership of *BuT5* and *P*-element traces back to at least 16 Ma.

BuT5 has been found to be the transposon directly responsible for inversions 2*s* in *D. mojavensis* (Guillén and Ruiz 2012) and $2x^3$ in *D. uniseta* (Prada 2010) that were generated by ectopic recombination between copies inserted in opposite orientation at two chromosomal sites (fig. 1). However, *BuT5* is a nonautonomous element that does not encode for a transposase and thus requires the *P*-element transposase for its mobilization. Thus, if *BuT5* is the main actor, *P* element must be considered as a necessary accomplice. *P* element is not only necessary for *BuT5* mobilization, that is, for the insertion of the two *BuT5* copies in their chromosomal sites, but it is also likely to have taken part in the ectopic recombination event. Ectopic recombination begins with the generation of a DSB followed by the DNA ends searching for homologous sequences for DSB repair. *P*-element transpose by a cut-and-paste mechanism

that involves the binding of the transposase to the element TIRs and the excision of the element generating a DSB at the donor site followed by the integration of the element into a different chromosomal site (Beall and Rio 1997; Tang et al. 2007). Hence, DSBs produced during normal or aberrant transposition events may provide the required initial step for ectopic recombination events.

The *P* element is well known for its potential to induce chromosomal rearrangements in the laboratory *D. melanogaster* (Berg et al. 1980; Engels and Preston 1981, 1984; Gray 2000). In contrast, no direct evidence has been found so far for the generation of natural chromosomal inversions by *P* elements. None of the eight inversions that are polymorphic in natural populations of *D. melanogaster* were seemingly generated by TEs (Corbett-Detig et al. 2012). In *D. willistoni*, a species with a rich inversion polymorphism, *P*-element hybridization sites in the polytene chromosomes often coincide with inversion breakpoints (Regner et al. 1996), but this provides only circumstantial evidence as TEs can be secondary colonizers of inversions breakpoints (Cáceres et al. 2001; Delprat et al. 2009). We have shown that *BuT5* and its master transposon *P* element generated two inversions recently fixed in *D. mojavensis* and *D. uniseta*, two species of the *repleta* group (fig. 1). Therefore, this is the first unequivocal demonstration of the role of *P* elements in *Drosophila* chromosomal evolution.

Supplementary Material

Supplementary tables S1–S6 and figure S1 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors are grateful to Cédric Feschotte, Josefa González, Yolanda Guillén, Mar Marzo, Marta Puig, and Rosemary Thwaite for critical reading of previous versions of the manuscript. They also thank UC San Diego Drosophila Species Stock Center, Deodoro Oliveira, and William Etges for sharing with them samples of *Drosophila* species. This work was supported by grants BFU2008-04988 and BFU2011-30476 from Ministerio de Ciencia e Innovación (Spain) to A.R. and by a PIF-UAB fellowship to N.R.

Literature Cited

Altschul SF, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25: 3389–3402.

Bao W, Jurka MG, Kapitonov VV, Jurka J. 2009. New superfamilies of eukaryotic DNA transposons and their internal divisions. Mol Biol Evol. 26:983–993.

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

Beall EL, Rio DC. 1997. *Drosophila P*-element transposase is a novel site-specific endonuclease. Genes Dev. 11:2137–2151.

Berg R, Engels W, Kreber R. 1980. Site-specific X-chromosome rearrangements from hybrid dysgenesis in *Drosophila melanogaster*. Science 210:427–429.

Bingham PM, Kidwell MG, Rubin GM. 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.

Bureau TE, Wessler SR. 1992. *Tourist*: a large family of small inverted repeat elements frequently associated with maize genes. Plant Cell 4:1283–1294.

Bureau TE, Wessler SR. 1994. *Stowaway*: a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants. Plant Cell 6:907–916.

Cáceres M, Puig M, Ruiz A. 2001. Molecular characterization of two natural hotspots in the *Drosophila buzzatii* genome induced by transposon insertions. Genome Res. 11:1353–1364.

Cáceres M, Ranz JM, Barbadilla A, Long M, Ruiz A. 1999. Generation of a widespread *Drosophila* inversion by a transposable element. Science 285:415–418.

Casals F, Cáceres M, Ruiz A. 2003. The foldback-like transposon *Galileo* is involved in the generation of two different natural chromosomal inversions of *Drosophila buzzatii*. Mol Biol Evol. 20:674–685.

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

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 17: 540–552.

Castro JP, Carareto CMA. 2004. *Drosophila melanogaster P* transposable elements: mechanisms of transposition and regulation. Genetica 121: 107–118.

Clark JB, Kidwell MG. 1997. A phylogenetic perspective on *P* transposable element evolution in *Drosophila*. Proc Natl Acad Sci U S A. 94: 11428–11433.

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 U S A. 102:6907–6912.

Corbett-Detig RB, Cardeno C, Langley CH. 2012. Sequence-based detection and breakpoint assembly of polymorphic inversions. Genetics 192: 131–137.

Craig NL. 2002. Mobile DNA: an introduction. In: Craig NL, Craigie R, Gellert M, Lambowitz AM, editors. Mobile DNA II. Washington: American Society for Microbiology Press. p. 3–11.

Daniels SB, Peterson KR, Strausbaugh LD, Kidwell MG, Chovnick A. 1990. Evidence for horizontal transmission of the *P* transposable element between *Drosophila* species. Genetics 124:339–355.

de Freitas Ortiz M, Lorenzatto KR, Corrêa BRS, Loreto ELS. 2010. *hAT* transposable elements and their derivatives: an analysis in the 12 *Drosophila* genomes. Genetica 138:649–655.

Delprat A, Negre B, Puig M, Ruiz A. 2009. The transposon *Galileo* generates natural chromosomal inversions in *Drosophila* by ectopic recombination. PloS One 4:7883.

Depra M, Ludwig A, Valente V, Loreto E. 2012. *Mar*, a MITE family of *hAT* transposons in *Drosophila*. Mob DNA. 3:13.

Dias ES, Carareto CMA. 2011. *msechBari*, a new MITE-like element in *Drosophila sechellia* related to the *Bari* transposon. Genome Res. 93:381–385.

Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113.

Engels WR. 1996. *P* elements in *Drosophila*. In: Saedler H, Gierl A, editors. Transposable elements. Berlin (Germany): Springer. p. 103–123.

Engels WR, Preston CR. 1981. Identifying *P* factors in *Drosophila* by means of chromosome breakage hotspots. Cell 26:421–428.

31

Engels WR, Preston CR. 1984. Formation of chromosome rearrangements by *P* factors in *Drosophila*. Genetics 107:657–678.
Fedoroff NV. 2012. Transposable elements, epigenetics, and genome evolution. Science 338:758–767.
Feschotte C, Mouchès C. 2000. Evidence that a family of miniature inverted-repeat transposable elements (MITEs) from the *Arabidopsis thaliana* genome has arisen from a *pogo*-like DNA transposon. Mol Biol Evol. 17:730–737.
Feschotte C, Pritham EJ. 2007. DNA transposons and the evolution of eukaryotic genomes. Annu Rev Genet. 41:331–368.
Feschotte C, Swamy L, Wessler SR. 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.
Feschotte C, Zhang X, Wessler SR. 2002. Miniature inverted-repeat transposable elements and their relantionship to established DNA transposons. In: Craig NL, Craigie R, Gellert M, Lambowitz AM, editors. Mobile DNA II. Washington: American Society for Microbiology Press. p. 1147–1158.
Fredslund J, Schauser L, Madsen LH, Sandal N, Stougaard J. 2005. PriFi: using a multiple alignment of related sequences to find primers for amplification of homologs. Nucleic Acids Res. 33:516–520.
Gray YHM. 2000. It takes two transposons to tango: transposable-element-mediated chromosomal rearrangements. Trends Genet. 16: 461–468.
Guermonprez H, Loot C, Casacuberta JM. 2008. Different strategies to persist: the *pogo*-like *Lemi1* transposon produces miniature inverted-repeat transposable elements or typical defective elements in different plant genomes. Genetics 180:83–92.
Guillén Y, Ruiz A. 2012. Gene alterations at *Drosophila* inversion breakpoints provide prima facie evidence for natural selection as an explanation for rapid chromosomal evolution. BMC Genomics 13:53.
Hagemann S, Haring E, Pinsker W. 1996a. A new *P* element subfamily from *Drosophila tristis*, *D. ambigua*, and *D. obscura*. Genome 39: 978–985.
Hagemann S, Haring E, Pinsker W. 1996b. Repeated horizontal transfer of *P* transposons between *Scaptomyza pallida* and *Drosophila bifasciata*. Genetica 98:43–51.
Hagemann S, Miller WJ, Pinsker W. 1992. Identification of a complete *P*-element in the genome of *Drosophila bifasciata*. Nucleic Acids Res. 20:409–413.
Hagemann S, Miller W, Pinsker W. 1994. Two distinct *P* element subfamilies in the genome of *Drosophila bifasciata*. Mol Gen Genet. 244: 168–175.
Hammer SE, 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. Mol Biol Evol. 22: 833–844.
Haring E, Hagemann S, Pinsker W. 1998. Transcription and splicing patterns of M- and O-type *P* elements in *Drosophila bifasciata*, *D. helvetica*, and *Scaptomyza pallida*. J Mol Evol. 46:542–551.
Holyoake AJ, Kidwell MG. 2003. *Vege* and *Mar*: two novel *hAT* MITE families from *Drosophila willistoni*. Mol Biol Evol. 20:163–167.
Jurka J, et al. 2005. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res. 110:462–467.
Jurka J, Kapitonov VV, Kohany O, Jurka MV. 2007. Repetitive sequences in complex genomes: structure and evolution. Annu Rev Genomics Hum Genet. 8:241–259.
Jurka J, Klonowski P, Dagman V, Pelton P. 1996. Censor—a program for identification and elimination of repetitive elements from DNA sequences. Comput Chem. 20:119–121.
Kapitonov VV, Jurka J. 1996. The age of Alu subfamilies. J Mol Evol. 42: 59–65.
Kapitonov VV, Jurka J. 2003. Molecular paleontology of transposable elements in the *Drosophila melanogaster* genome. Proc Natl Acad Sci U S A. 100:6569–6574.
Kapitonov VV, Jurka J. 2008. A universal classification of eukaryotic transposable elements implemented in Repbase. Nat Rev Genet. 9: 411–412.
Kapitonov VV, Jurka J. 2009. First examples of protozoan *P* DNA transposons. Repbase Rep. 9:2162.
Kaufman PD, Doll RF, Rio DC. 1989. *Drosophila P* element transposase recognizes internal *P* element DNA sequences. Cell 59:359–371.
Kazazian HH. 2004. Mobile elements: drivers of genome evolution. Science 303:1626–1632.
Kidwell MG. 1994. The evolutionary history of the *P* family of transposable elements. J Hered. 85:339–346.
Kidwell MG, Lisch DR. 2002. Transposable elements as sources of genomic variation. In: Craig NL, Craigie R, Gellert M, Lambowitz AM, editors. Mobile DNA II. Washington: American Society for Microbiology Press. p. 59–90.
Kidwell MG, Novy JB. 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.
Lee CC, Mul YM, Rio DC. 1996. The *Drosophila P*-element KP repressor protein dimerizes and interacts with multiple sites on *P*-element DNA. Mol Cell Biol. 16:5616–5622.
Lee SH, Clark JB, Kidwell MG. 1999. A *P* element-homologous sequence in the house fly, *Musca domestica*. Insect Mol Biol. 8:491–500.
Liao GC, Rehm EJ, Rubin GM. 2000. Insertion site preferences of the *P* transposable element in *Drosophila melanogaster*. Proc Natl Acad Sci U S A. 97:3347–3351.
Linheiro RS, Bergman CM. 2008. Testing the palindromic target site model for DNA transposon insertion using the *Drosophila melanogaster P*-element. Nucleic Acids Res. 36:6199–6208.
Linheiro RS, Bergman CM. 2012. Whole genome resequencing reveals natural target site preferences of transposable elements in *Drosophila melanogaster*. PloS One 7:e30008.
Loreto ELS, Zambra FMB, Ortiz MF, Robe LJ. 2012. New *Drosophila P*-like elements and reclassification of *Drosophila P*-elements subfamilies. Mol Genet Genomics. 287:531–540.
Loreto EL, Valente VL, Zaha A, Silva JC, Kidwell MG. 2001. *Drosophila mediopunctata P* elements: a new example of horizontal transfer. J Hered. 92:375–381.
Majumdar S, Singh A, Rio DC. 2013. The human THAP9 gene encodes an active *P*-element DNA transposase. Science 339:446–448.
Marchler-Bauer A, et al. 2011. CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Res. 39: D225–D229.
Marzo M, Puig M, Ruiz A. 2008. The *Foldback*-like element *Galileo* belongs to the *P* superfamily of DNA transposons and is widespread within the *Drosophila* genus. Proc Natl Acad Sci U S A. 105: 2957–2962.
O'Hare K, Rubin GM. 1983. Structures of *P* transposable elements and their sites of insertion and excision in the *Drosophila melanogaster* genome. Cell 34:25–35.
Oliveira DCSG, et al. 2012. Monophyly, divergence times, and evolution of host plant use inferred from a revised phylogeny of the *Drosophila repleta* species group. Mol Phylogenet Evol. 64: 533–544.
Oliveira DCSG, O'Grady PM, Etges WJ, Heed WB, DeSalle R. 2005. Molecular systematics and geographical distribution of the *Drosophila longicornis* species complex (Diptera: Drosophilidae). Zootaxa 1069:1–32.

Perkins HD, Howells AJ. 1992. Genomic sequences with homology to the *P* element of *Drosophila melanogaster* occur in the blowfly *Lucilia cuprina*. Proc Natl Acad Sci U S A. 89:10753–10757.
Pinsker W, Haring E, Hagemann S, Miller W. 2001. The evolutionary life history of *P* transposons: from horizontal invaders to domesticated neogenes. Chromosoma 110:148–158.
Pitnick S, Heed WB. 1994. New species of cactus-breeding *Drosophila* (Diptera: Drosophilidae) in the *nannoptera* species group. Ann Entomol Soc Am. 87:307–310.
Posada D. 2008. jModelTest: phylogenetic model averaging. Mol Biol Evol. 25:1253–1256.
Prada CF. 2010. Evolución cromosómica del cluster *Drosophila martensis*: origen de las inversiones y reutilización de los puntos de rotura [PhD thesis]. [Barcelona (Spain)]: Universitat Autònoma de Barcelona.
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.
Quesneville H, Nouaud D, Anxolabehere D. 2006. *P* elements and MITE relatives in the whole genome sequence of *Anopheles gambiae*. BMC Genomics 7:214.
Regner LP, Pereira MSO, Alonso CEV, Abdelhay E, Valente VLS. 1996. Genomic distribution of *P* elements in *Drosophila willistoni* and a search for their relationship with chromosomal inversions. J Hered. 87:191–198.
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? Mol Biol Evol. 20: 869–879.
Rhee Y, Lin H, Buell R, Childs K, Kaeppler S. 2009. A *c2* allele of maize identified in regenerant-derived progenyfrom tissue culture results from insertion of a novel transposon. Maydica 54:429–437.
Rio DC. 1991. Regulation of *Drosophila P* element transposition. Trends Genet. 7:282–287.
Rio DC. 2002. *P* transposable element in *Drosophila melanogaster*. In: Craig NL, Craigie R, Gellert M, Lambowitz AM, editors. Mobile DNA II. Washington: American Society for Microbiology Press. p. 484–518.
Rodríguez-Trelles F, Tarrío R, Ayala FJ. 2006. Origins and evolution of spliceosomal introns. Annu Rev Genet. 40:47–76.
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.
Rubin GM, Kidwell MG, Bingham PM. 1982. The molecular basis of P-M hybrid dysgenesis: the nature of induced mutations. Cell 29: 987–994.
Ruiz A, Wasserman M. 1993. Evolutionary cytogenetics of the *Drosophila buzzatii* species complex. Heredity 70:582–596.
Sabogal A, Lyubimov AY, Corn JE, Berger JM, Rio DC. 2010. THAP proteins target specific DNA sites through bipartite recognition of adjacent major and minor grooves. Nat Struct Mol Biol. 17:117–123.
Sarkar A, et al. 2003. *P* elements are found in the genomes of nematoceran insects of the genus *Anopheles*. Insect Biochem Mol Biol. 33: 381–387.
Smit AF, Riggs AD. 1996. Tiggers and DNA transposon fossils in the human genome. Proc Natl Acad Sci U S A. 93:1443–1448.
Tamura K, et al. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol. 28:2731–2739.
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 DC. 2007. Analysis of *P* element transposase protein-DNA interactions during the early stages of transposition. J Biol Chem. 282:29002–29012.
Tenaillon MI, Hollister JD, Gaut BS. 2010. A triptych of the evolution of plant transposable elements. Trends Plant Sci. 15:471–478.
Tu Z. 2000. Molecular and evolutionary analysis of two divergent subfamilies of a novel miniature inverted repeat transposable element in the yellow fever mosquito, *Aedes aegypti*. Mol Biol Evol. 17:1313–1325.
Turcotte K, Srinivasan S, Bureau T. 2001. Survey of transposable elements from rice genomic sequences. Plant J. 25:169–179.
Wasserman M. 1992. Cytological evolution of the *Drosophila repleta* species group. In: Krimbas CB, Powell JR, editors. *Drosophila* inversion polymorphism. Boca Raton (FL): CRC press. p. 455–552.
Wheatcroft R, Laberge S. 1991. Identification and nucleotide sequence of *Rhizobium meliloti* insertion sequence *ISRm3*: similarity between the putative transposase encoded by *ISRm3* and those encoded by *Staphylococcus aureus* IS256 and *Thiobacillus ferrooxidans* IST2. J Bacteriol. 173:2530–2538.
Wicker T, et al. 2007. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 8:973–982.
Yang G, Holligan-Nagel D, Feschotte C, Hancock C, Wessler S. 2009. Tuned for transposition: molecular determinants underlying the hyperactivity of a *Stowaway* MITE. Science 325:1391–1394.
Yuan Y-W, Wessler SR. 2011. The catalytic domain of all eukaryotic cut-and-paste transposase superfamilies. Proc Natl Acad Sci U S A. 108:7884–7889.
Zhang X, et al. 2001. *P* instability factor: an active maize transposon system associated with the amplification of *Tourist*-like MITEs and a new superfamily of transposases. Proc Natl Acad Sci U S A. 98: 12572–12577.

Associate editor: Esther Betran

3.1.2 SUPPLEMENTARY MATERIAL

Supplementary information is available at the Genome Biology and Evolution website (http://gbe.oxfordjournals.org/content/5/6/1127/suppl/DC1) and in the appedix section A.

3.2 EXPLORATION OF THE *D. BUZZATII* TRANSPOSABLE ELEMENT CONTENT

This Section contains the manuscript "Exploration of the *D. buzzatii* transposable element content suggests underestimation of repeats in Drosophila genomes" that has been recently completed and is presently under review.

Exploration of the *D. buzzatii* transposable element content suggests underestimation of repeats in Drosophila genomes

Nuria Rius[(1)], *Yolanda Guillén*[(1)], *Alejandra Delprat*[(1)], *Aurélie Kapusta*[(2)], *Cédric Feschotte*[(2)] *and Alfredo Ruiz*[(1)]

[(1)] Department de Genética i Microbiologia
Universitat Autònoma de Barcelona
Bellaterra Spain

[(2)] Department of Human Genetics
University of Utah School of Medicine
Salt Lake City, UT, USA

3.2.1 ABSTRACT

Background

Many new Drosophila genomes have been sequenced in recent years using new-generation sequencing platforms and assembly methods. Transposable elements (TEs), as repetitive sequences, are often misassembled, especially in the genomes sequenced with short reads. Consequently, the mobile fraction of many of the new genomes has not been analyzed in detail or compared with that of other genomes sequenced with different methods, which could shed light into the understanding of genome and TE evolution. Here we compare the TE content of three genomes *D. buzzatii* st-1, j-19, and *D. mojavensis*.

Results

We have sequenced a new *D. buzzatii* genome (j-19) that complements the *D. buzzatii* reference genome (st-1) already published, and compared their TE content with that of *D. mojavensis*. We found an underestimation of TE sequences in *Drosophila* genus NGS-genomes when compared to Sanger-genomes. To be able to compare genomes sequenced with different technologies, we developed a coverage-based method and applied it to *D. buzzatii* st-1 genome. Between 10.85 and 11.16% of *D. buzzatii* genome are made up by TEs, while TEs are a 15.35% of D. mojavensis genome. Helitrons are the most abundant order in both species.

Conclusions

TEs in *D. buzzatii* are less abundant than in *D. mojavensis*, as expected according to the genome size and TE content positive correlation. However, TEs alone does not explain the genome size difference. TEs accumulate in the dot chromosomes and proximal regions of *D. buzzatii* and *D. mojavensis* chromosomes. We also report a significantly higher TE density in *D. buzzatii* and *D. mojavensis* X chromosome, which is not expected under the current models. Our easy-to-use correction method allowed us to identify recently active families in *D. buzzatii* st-1 belonging to the LTR-retrotransposon superfamily Gypsy.

3.2.2 BACKGROUND

Transposable elements (TEs) are mobile DNA sequences present in virtually all the eukaryote genomes sequenced and are accountable for variable fractions of the genomes they inhabit. TEs are important not only because of their abundance but also because they are active components of the genomes, inducing structural rearrangements, inactivating or duplicating genes and adding or removing regulatory regions (Akagi et al., 2013).

There are two classes of TEs, those that mobilize via an RNA intermediate belong to class I and those which transpose directly, leaving the donor site, or via a DNA intermediate, to class II (Wicker et al., 2007; Kapitonov and Jurka, 2008). Further divisions in this classification comprehend orders that distinguish TEs with different insertion mechanisms, and superfamilies that are composed by TEs with similar domain structures and protein sequence similarity.

Progress in all aspects of genome sequencing and assembly has driven a revolution in the field. After *D. melanogaster* (Adams et al., 2000) and *D. pseudoobscura* (Richards et al., 2005) were sequenced, joint efforts provided the research community with ten new species genomes that allowed multiple species comparisons (Drosophila 12 Genomes Consortium, 2007). After those, six *de novo* genomes were published individually (Zhou et al., 2012; Zhou and Bachtrog, 2012; Ometto et al., 2013; Chiu et al., 2013; Fonseca et al., 2013; Guillén et al., 2015), and eight more were recently published (Chen et al., 2014).

The production of new genomes seems unstoppable and the comparisons and the knowledge drawn from them limitless. However, the information contained in some *de novo* draft genomes sequenced with Next-Generation Sequencing (NGS) is not fully accurate (Salzberg and Yorke, 2005; Narzisi and Mishra, 2011). TEs, because of their repetitive nature, are in the root of most of these problems causing misassemblies (Ricker et al., 2012; Salzberg et al., 2012). Hence, contextualization and comparison of the TE fraction of genomes sequenced and annotated separately is difficult and scarce. Advances in sequencing technology (English et al., 2012; Huddleston et al., 2014) and standardization in annotation methods (McCoy et al., 2014) arrive with the promise to solve this issue, but meanwhile, already sequenced genomes keep piling up.

In this article, we analyze in detail the TE content of the *D. buzzatii* reference (st-1) genome (Guillén et al., 2015), and compare it to that of a second *D. buzzatii* strain (j-19), described here, and that of *D. mojavensis*, another member of the *repleta* group (Drosophila 12 Genomes Consortium, 2007). We also compare the TE fraction in all available *Drosophila* genus genomes to test whether there are

differences between NGS and Sanger-sequenced genomes, propose a method to correct such differences, and apply it to *D. buzzatii* reference genome.

3.2.3 METHODS

Genomes

The genomes used in this work were all freely available online except the genome of *D. buzzatii* strain j-19, which is described here and it is available through `http://dbuz.uab.cat`.

Strain j-19 was isolated from flies collected in Ticucho (Argentina) using the balanced-lethal stock Antp/Δ^5 (Piccinali et al., 2007). Individuals of the j-19 strain are homozygous for the chromosome arrangement 2j (Cáceres et al., 2001). DNA was extracted from male and female adults using the sodium dodecyl sulfate (SDS) method (Milligan, 1998) or the method described by Piñol and colleagues (Piñol et al., 1988) for isolating high molecular weight DNA. Three Illumina HiSeq Paired End (PE) libraries were prepared and sequenced at CNAG (Centro Nacional de Análisis Genómico) with an insert size of 500 bp and a mean read length of 102 bp. SOAPdenovo (Li et al., 2010) version 1.05 was used to assembly the genome of j-19 strain. We fed the assembler with 251,719,776 filtered reads setting the assembler with kmer size k=31. The final assembly contains 10529 scaffolds over 3 kb (total size = 153,440,896 bp). The N50 index is 1666, and the N50 length 24268 bp, the N90 index is 6825, and the N90 length 5747 bp.

Publicly available genomes from *Drosophila* genus were downloaded from FlyBase (*D. ananassae* r1.3, *D. erecta* r1.3, *D. grimshawi* r1.3, *D. melanogaster* r6.05, *D. mojavensis* r1.3, *D. persimilis* r1.3, *D. pseudoobscura* r 3.2, *D. sechellia* r1.3, *D. simulans* r1.3 and r2.01 (Hu et al., 2013), *D. virilis* r1.2, *D. willistoni* r1.3, and *D. yakuba* r1.3 (Drosophila 12 Genomes Consortium, 2007)), NCBI (*D. albomicans* (Zhou et al., 2012), *D. biarmipes*, *D. bipectinata*, *D. elegans*, *D. eugracilis*, *D. ficusphila*, *D. kikkawai*, *D. miranda* (Zhou and Bachtrog, 2012), *D. rhopaloa*, *D. suzukii* (Chiu et al., 2013), and *D. takahashii* (Chen et al., 2014)) or project web sites (*D. americana* H5 (`http://cracs.fc.up.pt/~nf/dame/index.html`) (Fonseca et al., 2013) and *D. buzzatii* st-1 (`http://dbuz.uab.cat`) (Guillén et al., 2015)).

Transposable element library

We built a custom library to annotate and classify the mobile elements in *D. buzzatii* and *D. mojavensis* genomes. The library comprised already known repeats (FlyBase and Repbase) and *de novo* elements found in *D. buzzatii* st-1 genome (RepeatModeler and Repclass). FlyBase canonical set of TEs (`http://flybase.org/`) were blasted (Altschul et al., 1997) against an early assembly of the *D. buzzatii* st-1 genome. For each query, significant hits were manually inspected in order to recover the most complete copy. Repbase (Jurka et al., 2005) repeats from *Insecta* species were added to the library. RepeatModeler (version 1.0.4) (Smit and Hubley, 2008) was used with RepeatScout (Price et al., 2005) and Recon (Bao and Eddy, 2002) to identify repeats, and RMBlast engine and Repbase database to classify them. Repclass (Feschotte et al., 2009) was used to classify repeats identified by RepeatScout. Elements classified by Repclass without identity to previously identified repeats or being more complete were added to the library. Sequences classified as simple repeats, satellite or low complexity, were removed from the library. Additionally, a blast analysis was performed to filter non-TE related sequences. Sequences with significant hits (e-value blast $<$ 1e-25) with *D. mojavensis* coding sequences (cds) and at the same time with no significant similarity to repeats deposited in Repbase were removed.

Repeat annotation

To compare the three genomes of the two *Drosophila repleta* group species (*D. buzzatii* st-1, *D. buzzatii* j-19 and *D. mojavensis*), we masked them with RepeatMasker (Smit et al., 1996) (version 4.0.5) and RMBlast (version 2.2.27+) and the *D. buzzatii* custom library using the default options except for cut off (score value 250), nolow and norna. We used the RepeatMasker output files *.out to estimate the amount of nucleotides of each order and superfamily. We also used RepeatMasker, with cut off 250, nolow, and norna, to assess the TE content of the 26 available Drosophila genomes, from 25 species. To reduce library bias factor we used the RepBase *Insecta* library. The assembly size was used, in each case, to compute the percentage of transposable elements.

Chromosomal analysis

We analyzed the TE distribution along the chromosomes of *D. buzzatii* st-1 and *D. mojavensis*. We used the information of the previously mapped and oriented

scaffolds, the 158 N90 scaffolds (145 Mb) of *D. buzzatii* (Guillén et al., 2015), and the 11 N80 scaffolds (156 Mb) of *D. mojavensis* (Schaeffer et al., 2008). The scaffolds were broken down into 50 kb non-overlapping windows using bedtools (makewindows) and the TE nucleotides in each window were calculated using also bedtools (intersect). We plotted the TE density (TE bp/window length) for all windows, including those smaller than 50 kb from the tip of each scaffold, in the reported order.

To assess the TE-density in every chromosome, in the proximal regions and in the rest of the chromosome independently, another set of windows was made with the *D. buzzatii* and *D. mojavensis* mapped scaffolds previously mentioned. The most proximal 3 Mb of chromosomes X, 2, 3, 4 and 5 ($\sim$ 10% of the chromosome) were divided in 50 kb windows as well as the remaining $\sim$90% of the chromosomes, and the entire chromosome 6. Only whole windows (50 kb) were taken into account. For each chromosome and region, we computed the mean TE-density and standard deviation and plotted the TE-density window distribution. Additionally, differences among these distributions (whole chromosome, proximal and central+distal regions) were tested with the two samples Kolmogorov-Smirnov test.

Correction

We mapped the reads used in the genome pre-assembly of *D. buzzatii* st-1 (21924977 reads from 454, Illumina, and Sanger) (Guillén et al., 2015) with GS Reference Mapper (v2.9) (`http://454.com/products/analysis-software`) to the final *D. buzzatii* assembly using default options. GS Reference Mapper aligned 95.3% of the reads (20422434 reads), 20270 reads less than those used by gs-Assembler to build the pre-assembly. Every read base pair that mapped in a TE-annotated position was added up to know the coverage of that position. The corrected value for each TE order and superfamily is the sum of read base pairs annotated as part of an order or superfamily, divided by the average coverage of single copy genes. Single copy gene average coverage, 22.37x, was calculated with the same procedure used for TEs, but with 13657 single copy genes identified in *D. buzzatii* st-1 genome (Guillén et al., 2015).

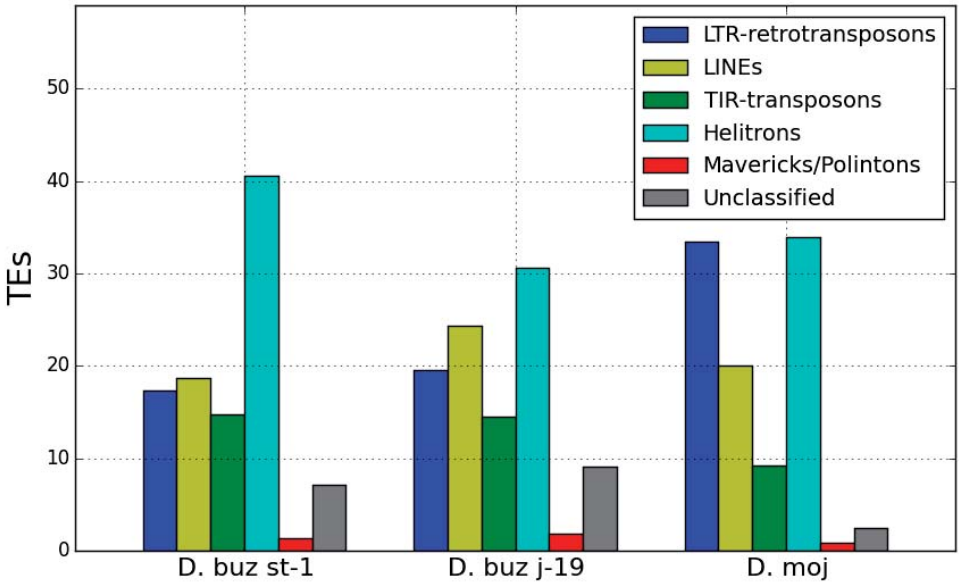

Figure 5.: Percentage of transposable element orders relative to the mobile fraction of the genomes of *D. buzzatii* st-1, j-19, and *D. mojavensis*.

3.2.4 RESULTS

TE content in D. buzzatii and D. mojavensis

In *D. buzzatii* st-1 TEs account for 8.43% of the genome, about twice the value of TEs in *D. buzzatii* j-19 (4.15%), but almost half of the value of *D. mojavensis* (15.35%). In order to make a fair comparison, we also considered only 3-kb or longer scaffolds for *D. mojavensis*, 2419 (187.4 Mb) out of 6841 scaffolds (193.8 Mb). However, the TE fraction in *D. mojavensis* genome is still higher (14.35%) than the fraction in both *D. buzzatii* strains. Henceforth, the complete *D. mojavensis* genome was used for the subsequent analyses.

Table 1.: Contribution of every order and superfamily (kb) to the *D. buzzatii* (st-1, st-1 after the correction and j-19) and *D. mojavensis* genomes [1].

Superfamily	***D. buz***			***D. moj***
	st-1	*st-1* corr.	*j-19*	
LTR Total	2366.44 (17.38 %)	4693.31 (26.03 %)	1243.43 (19.54 %)	9953.02 (33.46 %)
BelPao	435.35	1025.76	198.65	2255.95
Copia	309.80	522.62	162.75	718.71
ERVK	10.92	9.97	8.09	18.06
Gypsy	1610.37	3134.95	873.94	6960.30
LINE Total	2541.65 (18.66 %)	3401.72 (18.87 %)	1551.05 (24.37 %)	5977.29 (20.09 %)
CR1	396.35	761.48	117.39	947.96
I	74.63	136.15	20.19	110.53
Jockey	478.24	600.72	246.54	765.64
L1	6.71	6.01	6.70	8.08
L2	191.37	213.18	145.73	395.99
LOA	1.18	1.31	0.82	1.95
R1	1383.35	1663.22	1011.77	3721.30
R2	1.49	9.30	0.51	23.03
R4	1.57	0.80	0.70	1.37
RTE	6.76	9.55	0.69	1.43
TIR Total	2016.98 (14.81 %)	2476.88 (13.74 %)	919.50 (14.46 %)	2747.83 (9.24 %)
hAT	563.03	661.13	239.06	654.13
Mutator	21.00	16.32	16.14	22.73
Novosib	17.35	16.43	11.89	16.15
P	590.70	830.17	216.28	752.39
PIFHarbinger	3.81	9.71	2.21	7.82
PiggyBack	18.67	9.46	5.38	77.21
Tc1Mariner	407.93	507.35	186.38	534.42
TIR other	113.23	310.35	69.75	55.43
Transib	281.27	115.97	172.40	627.54
Helitron	5531.01 (40.61 %)	6331.89 (35.12 %)	1950.81 (30.65 %)	10083.94 (33.90 %)
Maverick	189.27 (1.39 %)	129.44 (0.72 %)	118.57 (1.86 %)	263.81 (0.89 %)
Others	0.24 (0 %)	0.11 (0 %)	0.67 (0 %)	0.19 (0 %)
Unknown	973.76 (7.15 %)	994.61 (5.52 %)	580.02 (9.11 %)	721.26 (2.42 %)
Total	13619.34	18027.96	6364.04	29747.33

[1] Order constributions, relative to the total TE fraction, are given in percentages

The contribution of the different orders, deffined by Wicker et al. (2007), to the total amount of TEs (Figure 5 and Table 1), is similar between the two *D. buzzatii* genomes (Helitrons, LINEs, LTR-retrotransposons, TIR-transposons, and Mavericks/Polintons), and differs from the one of *D. mojavensis*. Even though, there are some significant differences. Although Helitrons are the most abundant order in the three genomes, they are more abundant in *D. buzzatii* st-1 genome (40.61% of the TEs content) than in the other two genomes (30.65% in *D. buzzatii* j-19 and 33.90% in *D. mojavensis*). LTR-retrotransposons are the second most abundant order in *D. mojavensis* (33.46%), but not in *D. buzzatii* (17.38% in st-1 and 19.54% in j-19) where in both strains LINEs are the second order in genome contribution. TIR-transposons are more frequent in *D. buzzatii* genomes (14.81% in st-1 and 14.46% in j-19) than in *D. mojavensis* (9.24%), like the unclassified repeats that are more abundant in *D. buzzatii* (7.15% in st-1 and 9.11% in j-19) than in *D. mojavensis* (2.42%).

Chromosomal distribution

The TE distribution along *D. buzzatii* N90 mapped scaffolds and *D. mojavensis* N80 mapped scaffolds, see Figure 6, shows that TE density in all chromosomes rises when closer to the centromere. The results also show an increase in *D. buzzatii* and *D. mojavensis* X chromosomes when compared to the autosomes (Figure 6). The density of the main orders plotted individually (Supplementary Fig. 1, a-h) reveals the prevalence of Helitrons in *D. buzzatii* proximal regions, specially the 3 Mb closest to the centromere.

We compared the abundance of TEs annotated in *D. buzzatii* and *D. mojavensis*, specifically the distribution of TE density in 50 kb windows, for whole chromosomes (the N90 mapped scaffolds of *D. buzzatii* and the N80 mapped scaffolds of *D. mojavensis*), for proximal regions (3 Mb), and for central and distal regions (Table 2). It is important to note that only the largest scaffolds are being considered, and that 10 and 20% of *D. buzzatii* and *D. mojavensis* assemblies respectively, contained in the smallest and typically TE-enriched scaffolds, were discarded from this analysis. This explains the differences between the annotation of the whole assembly and the mean values of the mapped scaffolds. The smaller and TE-richer scaffolds are likely located in proximal regions, as the centromeric regions have the higher TE-density and more nested TEs. However, all recent TE insertions are susceptible to misassemblies and small scaffolds could be between mapped scaffolds.

Table 2.: TE fraction in *D. buzzatii* and *D. mojavensis* computed in 50 kb non-overlapping windows [1].

Chr	Species	Proximal		Cent+Dist		Total	
		TE (%)	N	TE (%)	N	TE (%)	N
X	*D. buzzatii*	16.13	57	7.44	505	8.32	562
	D. mojavensis	42.24	59	8.71	579	11.81	638
2	*D. buzzatii*	13.91	59	4.77	638	5.54	697
	D. mojavensis	38.68	60	5.11	622	8.06	682
3	*D. buzzatii*	12.96	58	4.12	522	5.01	580
	D. mojavensis	60.52	60	5.60	586	10.70	646
4	*D. buzzatii*	12.50	58	3.77	434	4.80	492
	D. mojavensis	39.24	60	4.31	486	8.14	546
5	*D. buzzatii*	14.98	58	4.06	462	5.87	520
	D. mojavensis	21.47	60	4.11	476	6.06	536
6	*D. buzzatii*	41.22	28	-	-	41.22	28
	D. mojavensis	50.65	60	14.22	8	46.30	68
Total	*D. buzzatii*	16.51	318	4.87	2561	5.86	2879
	D. mojavensis	42.13	359	5.68	2757	8.87	3116

[1] Proximal regions corresponds to the 3 most proximal Mb; Central+ Distal to the rest of the chromosome and Total to both parts. N stands for number of windows. Only mapped and oriented scaffolds are present, N90 for *D. buzzatii*, and N80 for *D. mojavensis*.

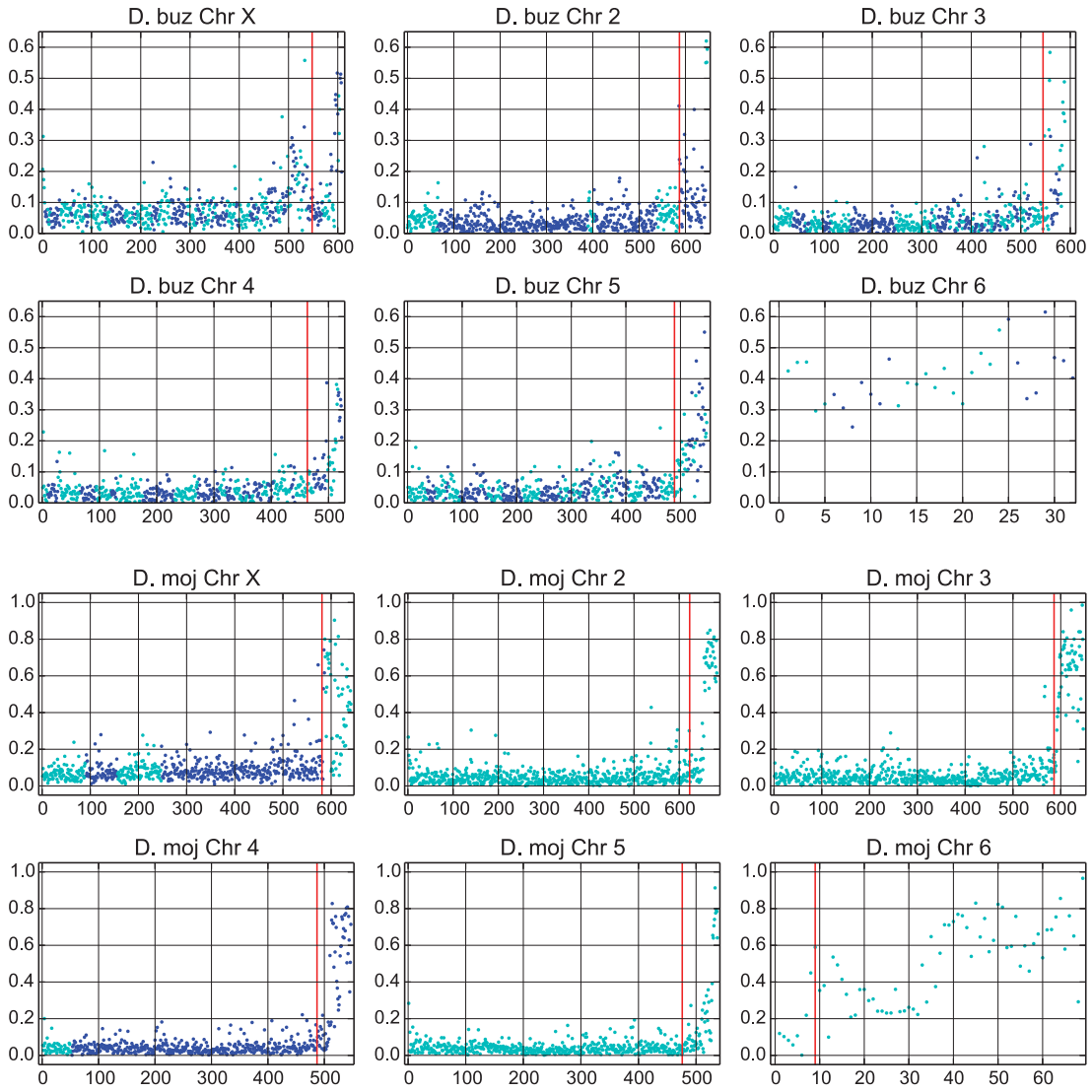

Figure 6.: Density of transposable elements in 50 kb non-overlapping windows, starting (left) from the telomere. Only mapped and oriented scaffolds are included, N90 for *D. buzzatii* st-1, and N80 for *D. mojavensis*. Telomere is on the right and centromere on the left. Changes in dot colors denote scaffold changes and the red lines mark the most proximal 3 Mb of each chromosome.

D. mojavensis chromosomes, as a whole, or any of their parts, have a higher TE fraction than *D. buzzatii* chromosomes. The biggest differences are in the proximal regions, and fade in the central and distal regions. Chromosome 6 (Muller element F) is the TE-richest chromosome in both species, 41.22% in *D. buzzatii*

and 46.30% in *D. mojavensis*. In *D. buzzatii*, 8.32% of chromosome X (Muller element A) sequence is made up by TEs, followed by the other chromosomes with values between 4.80% and 5.86%. In *D. mojavensis*, the X chromosome has 11.81% of TEs, chromosome 3 10.70% and the rest of the chromosomes have values comprised between 8.14% and 6.06%. *D. buzzatii* chromosomes 6 and X, when analyzed as a whole, are the only ones with TE density distributions significantly different (Two samples Kolmogorov-Smirnov test $p < 0.001$) from all other chromosomes, whereas in *D. mojavensis* are chromosomes 6, X, and 3 (Supplementary Tables 1-4). If we discard the 3 most proximal Mb and chromosome 6, chromosome X of both species is the only with significantly different TE density distribution from all the other chromosomes (Supplementary Tables 5-8). When the pericentromeric regions are compared, in *D. buzzatii* there are not significant differences among chromosomes, while among *D. mojavensis* proximal regions, chromosome 3 TE density is significantly different from the rest of the chromosomes (Supplementary Tables 9-12). Consequently, in both species chromosomes 6 and X display a significantly different TE distribution pattern from the rest of the chromosomes.

Impact of the sequencing method in Drosophila genus

Because the genomes of *D. mojavensis*, *D. buzzatii* st-1 and j-19 strains were sequenced with different platforms and assembly strategies (see Methods), the differences in TE content between these genomes could be related to the methodologies used. More specifically, the Sanger sequenced *D. mojavensis* genome (Drosophila 12 Genomes Consortium, 2007) shows a higher TE content than the *D. buzzatii* reference (st-1) genome sequenced with 454, Illumina and Sanger (Guillén et al., 2015), which itself has a higher TE content than the *D. buzzatii* j-19 genome sequenced only with Illumina. Therefore it seems that NGS yields a smaller repeat content than Sanger sequencing (Alkan et al., 2011).

In order to test this hypothesis, we widened our scope to include all the available genomes of *Drosophila* genus (Table 3). As in the cases of *D. mojavensis* and *D. buzzatii* there is a difference in the mobile fraction depending on the sequencing method. The mean of TE percentage in 12 genomes sequenced with Sanger technology is 19.31%, whereas that in 15 newly sequenced genomes chiefly with NGS is 10.98%. It is possible that the species sequenced with Sanger technology have *per se* more TEs than those sequenced with NGS, and sequencing or assembly methods do not influence the assemblies TE fraction. However, when species belonging to the same subgroup are compared, the Sanger-sequenced genomes

Table 3.: Percentage of TEs annotated with Repeat Masker and RepBase *Insecta* library on every available genomes of *Drosophila* genus.

Species	Subgenus	Group	Subgroup	Seq method	TEs
D. albomicans	Drosophila	immigrans	nasuta	NGS	2.73
D. buzzatii st-1	Drosophila	repleta	mulleri	NGS	5.99
D. buzzatii j-19	Drosophila	repleta	mulleri	NGS	2.40
D. mojavensis	Drosophila	repleta	mulleri	Sanger	16.14
D. americana	Drosophila	virilis	virilis	NGS	9.11
D. virilis	Drosophila	virilis	virilis	Sanger	17.51
D. grimshawi	Hawaian	grimshawi	grimshawi	Sanger	15.86
D. ananassae	Sophophora	melanogaster	ananassae	Sanger	30.33
D. bipectinata	Sophophora	melanogaster	ananassae	NGS	16.94
D. elegans	Sophophora	melanogaster	elegans	NGS	12.05
D. eugracilis	Sophophora	melanogaster	eugracilis	NGS	13.67
D. ficusphila	Sophophora	melanogaster	ficusphila	NGS	9.45
D. erecta	Sophophora	melanogaster	melanogaster	Sanger	14.41
D. melanogaster	Sophophora	melanogaster	melanogaster	Sanger	21.67
D. sechellia	Sophophora	melanogaster	melanogaster	Sanger	20.90
D. simulans	Sophophora	melanogaster	melanogaster	Sanger	11.85
D. simulans	Sophophora	melanogaster	melanogaster	NGS	8.44
D. yakuba	Sophophora	melanogaster	melanogaster	Sanger	21.98
D. kikkawai	Sophophora	melanogaster	montium	NGS	11.95
D. rhopaloa	Sophophora	melanogaster	rhopaloa	NGS	18.62
D. biarmipes	Sophophora	melanogaster	suzukii	NGS	14.48
D. suzukii	Sophophora	melanogaster	suzukii	NGS	18.70
D. takahashii	Sophophora	melanogaster	takahashii	NGS	14.68
D. miranda	Sophophora	obscura	obscura	NGS	5.47
D. persimilis	Sophophora	obscura	obscura	Sanger	23.97
D. pseudoobscura	Sophophora	obscura	obscura	Sanger	12.68
D. willistoni	Sophophora	willistoni	willistoni	Sanger	24.39

show a consistently higher percentage of TEs. The *mulleri* subgroup species, *D. buzzatii* and *D. mojavensis*, have different values than those yielded by our custom library but the pattern is the same. More examples (Table 3) are in *virilis*, *ananassae* or *obscura* subgroups, where the species sequenced with shorter reads have a lower percentage of mobile elements. Two genomes from the *virilis* subgroup have been sequenced, *D. virilis* with Sanger and *D. americana* with NGS, and have 17.51% and 9.11% of TEs respectively. *D. ananassae* sequenced with Sanger has 30.33% of TEs, *D. bipectinata* sequenced with NGS has 16.94%. Similarly, *D. persimilis* and *D. pseudoobscura*, sequenced with Sanger technology, have 23.91% and 12.68% respectively, whereas *D. miranda*, sequenced with NGS, has 5.47% of TEs in its genome. Moreover, the case of the same species sequenced by both technologies further supports the trend. *D. simulans* has been recently resequenced with NGS and old Sanger sequences to amend significant problems with the previous Sanger project. Our results show that the newly sequenced genome has 8.44% of TEs (6.85% according to Hu et al. (2013), the authors of the latter assembly) while the old assembly has 11.85%. Although various methodologies of repeat detection render various results, the use of the same procedure on Sanger and primarily NGS genomes gives consistently higher values of repeats in Sanger genomes. Hence, to accurately compare the results of *D. buzzatii* genome to other Sanger genomes like *D. mojavensis*, we thought it was necessary to correct our previous estimates of *D. buzzatii* st-1 TE fraction.

Correction of TE estimation by coverage

We found 403.3 Mb of reads, out of 3609 Mb, mapping to regions annotated as TEs in *D. buzzatii* st-1 assembly, corresponding to 11.16% of all reads mapped. After dividing this 403.3 Mb by the average gene coverage (22.37x) we got the corrected value of TEs of *D. buzzatii*, 18 Mb. Therefore there is a 1.32 fold underestimation (4.4 Mb) with respect to the 13.6 Mb initially annotated with RepeatMasker. If we keep considering the assembly size as the genome size, and assume the extra 4.4 Mb belong to the gaps within scaffolds (15 Mb) the initial estimate of TEs in the genome of 8.43% increases to 11.16%. On the other hand, if we add the 4.4 new Mb to the assembly size, we get 165.9 Mb genome size, and the TE fraction is 10.85%. We conclude that the TE fraction in *D. buzzatii* is between 10.85% and 11.16%.

Consequently, the orders and superfamilies with a higher correction factor are the ones with copies missing in the assembly. The results (Figure 7 and Table 1) show that LTR-retrotransposons are the most underestimated order by a factor

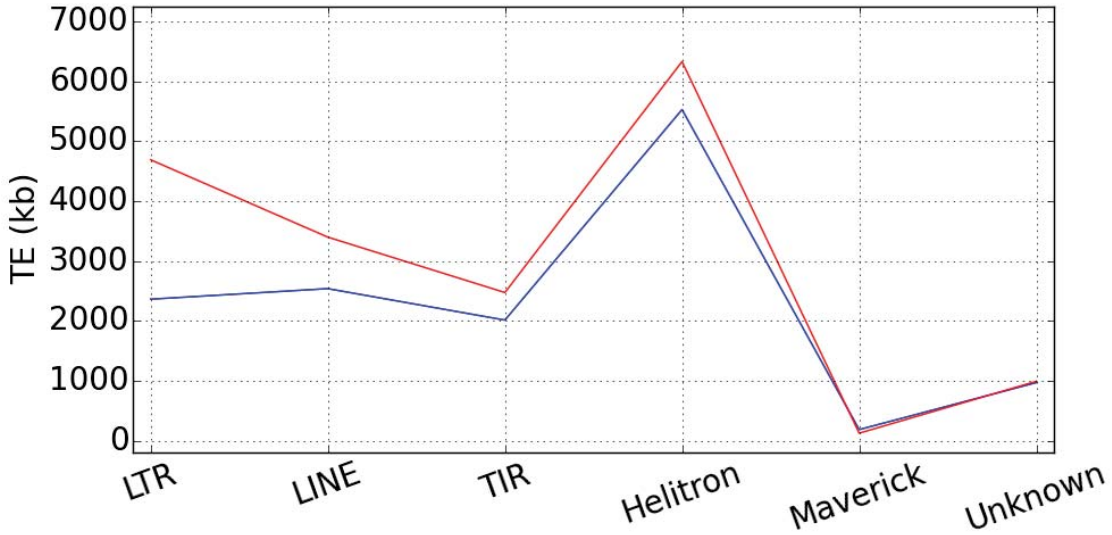

Figure 7.: Main order contribution (kb) to *D. buzzatii* genome, before (blue) and after (red) the coverage-based correction.

of 1.98. At the superfamily level (Figure 8), Gypsy and BelPao are the most underestimated, increasing after the correction by a factor of more than two fold. Consequently, both species TE profiles are more similar after the correction as *D. buzzatii* LTR-retrotransposons have now overtaken LINEs as the second most frequent order. LINEs are underrepresented in the genome annotation by 1.34 fold. The superfamilies CR1 and R1 increase 365 kb and 280 kb respectively after the correction. R2 superfamily represents a singular case, since it is not relevant in absolute value (1.5 kb annotated), but the correction factor is the highest of all superfamilies (6.24 fold) and after the correction 9.3 kb are found to belong to R2 superfamily. TIR-transposons are underestimated in the annotation by a 1.23 factor, with most superfamilies with a fair representation (correction factor close to one), but due to its large size, this small factor correction represent a substantial change in the base count. After the correction, P superfamily sequence has been increased in 239 kb (1.41 fold), Tc1/mariner cover 99 new kb (1.24 fold) and hAT 98 kb (1.17 fold). Helitrons are underestimated by a 1.15 factor, but like TIR-transposons, their abundance in the genome prior to the correction (5.5 annotated Mb) translates into a remarkable increase, 800 kb absent from the annotation. These superfamilies are likely to include highly similar insertions probably recently transposed.

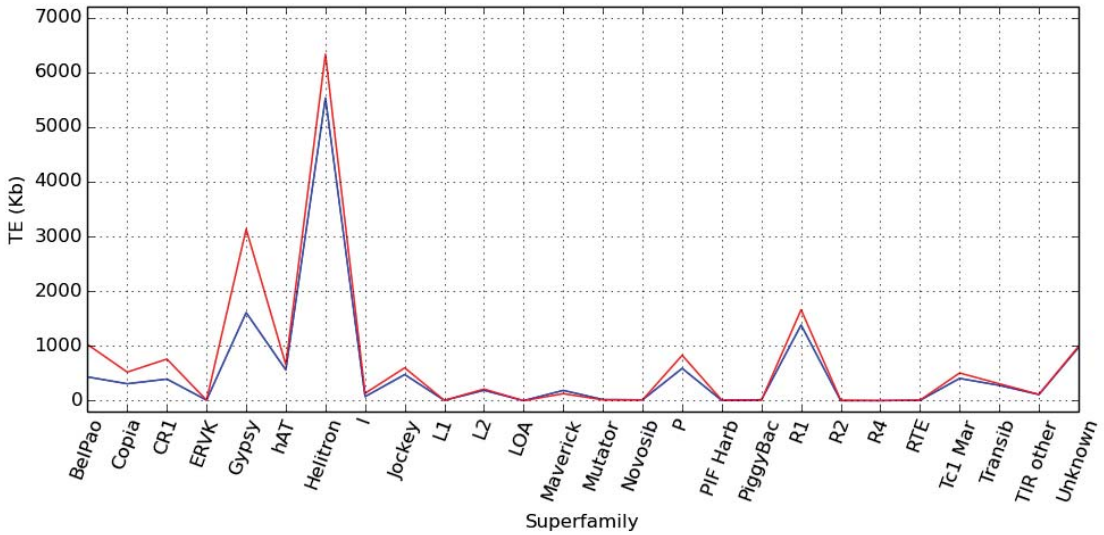

Figure 8.: Superfamilies contribution (kb) to *D. buzzatii* st-1 genome before (blue) and after (red) the coverage-based correction.

3.2.5 DISCUSSION

We have shown that *D. buzzatii* st-1 and j-19 genomes have a lower TE percentage than *D. mojavensis*. We have also reported that there is an underestimation of the mobile fraction of genomes sequenced with Next Generation Sequencing, possibly due to sequencing and assembly methods that affect *D. buzzatii* st-1 genome, and probably j-19 as well.

We have proposed a method based on read coverage to assess the magnitude of the bias, and used it to correct the *D. buzzatii* st-1 TE estimate. In *D. buzzatii* st-1 the correction revealed another 4.4 Mb of TEs and increased the TE percentage to 11%. Thus, although the TE content in *D. buzzatii* genome increased with the correction, it is still lower than that of *D. mojavensis* genome. Our methodology does not allow us to locate the TEs absent from the assembly. However, we consider it is important to describe the TEs present in the published assembly for several reasons. The differences while affecting particularly some orders and superfamilies have a small effect in others. Moreover, *D. buzzatii* uncorrected TE chromosomal distribution shows the same trends than those we observed in *D. mojavensis*. Finally, the published assembly should be analyzed and its limitations assessed in order to become a useful resource.

D. buzzatii and D. mojavensis TE content

Our results show that TEs in *D. buzzatii* genome are less abundant than in *D. mojavensis* genome, even after taking into account the bias correction. The size of the two genomes have been estimated by Feulgen Image Analysis Densitometry and the *D. buzzatii* genome estimates are between 21% (st-1) and 25% (j-19) smaller than those for *D. mojavensis*. Thus, our results agree with the well known positive correlation between genome size and transposable element fraction (Kidwell, 2002; Boulesteix et al., 2006; Feschotte and Pritham, 2007). However, the difference in TE content does not explain the difference in size between the two genomes. Interestingly, after the coverage-based correction, *D. buzzatii* st-1 and *D. mojavensis* have a TEs order composition relative to the TE fraction very similar, suggesting that the changes that lead to the differences affected every order in a uniform manner.

There are several non-mutually excluding explanations for the wide diversity in genome sizes and the forces driving its variation. The mutational explanation, ascribe part of such diversity to differences in insertion and deletion rates among species (Petrov et al., 2000; Gregory, 2004); other authors suggest that non-adaptative forces have diminished the efficiency of selection, explaining genomes expansions (Lynch, 2007); positive natural selection proposes that genome size constraints may be different depending of the lineage history (Charlesworth and Barton, 2004). And according to Charlesworth and Barton (2004), having a larger genome size may be advantageous, or at least not as strongly selected against, in some scenarios. Genome size has been reported to be negatively correlated with developmental rate, which is also negatively correlated with body size (Pagel and Johnstone, 1992; Wyngaard et al., 2005). Hence, species without a constrain on developmental time and favored by a larger body size may have accumulated more repetitive sequences than closer species with developmental time constraints.

This is possibly the case of *D. buzzatii*, which generally lay its eggs in rotting tissues of several Opuntia cacti, although it can occasionally use columnar cacti (Hasson et al., 1992; Ruiz et al., 2000; Oliveira et al., 2012); while *D. mojavensis* primarily uses larger rotting columnar or barrel cacti (*Stenocereus gummosus* and *Stenocereus thurberi*, and *Ferocactus cylindraceous*), except for the Santa Catalina Island population that uses Opuntia (Fellows and Heed, 1972; Heed and Mangan, 1986; Ruiz and Heed, 1988; Etges et al., 1999). In other words, *D. buzzatii* individuals mainly live in smaller cacti which dry faster, consequently a more ephemeral resource than those used by *D. mojavensis*. The selective pressure to

keep a faster development in *D. buzzatii*, or the relaxation of this pressure in *D. mojavensis* could be behind their different genome size and TE contribution.

Chromosomal distribution of TEs

TEs in *D. melanogaster* have been reported to accumulate in the proximal regions of the chromosomes, the transition between euchromatin and heterochromatin, where the recombination rate drops.The dot chromosome, which has a recombination rate considered null (Comeron et al., 2012), has the highest TE density of all chromosomes (Kaminker et al., 2002; Rizzon et al., 2002). Moreover, recent analyses of several *D. melanogaster* populations have found a negative correlation between recombination rate and TE population frequency (Petrov et al., 2011; Kofler et al., 2012).

TE dynamics has been extensively studied; however there is not a consensus about why some regions have a higher TE density. The transposition-selection balance model comprehends three non-mutually excluding hypotheses, which explain how TE insertions can be selected against: gene-disruption, deleterious TE-product expression, and ectopic recombination. According to the ectopic recombination hypothesis, the decrease in the recombination rates weakens the selection against TE insertions by reducing the crossing-over events between non-homologous TE copies (Comeron et al., 2012; Mackay et al., 2012). On the other hand, an alternative to ectopic recombination hypothesis, the transposition bursts model does not assume a constant transposition rate, instead it assumes that TEs undergo periods of high transposition activity. Although bursts are known to occur, ectopic recombination is so far the only explanation for the correlation between recombination rate and TE frequency (for review see (Barrón et al., 2014)).

Accumulation of specific transposable elements in *D. buzzatii* centromeric regions was previously noticed using in situ hybridization (Casals et al., 2005, 2006). Additionally *D. mojavensis* dot chromosome TE density is approximately 50%, higher than those of *D. melanogaster*, *D. erecta* and *D. grimshawi* (Leung et al., 2015). We are now reporting TE accumulations in the dot chromosomes and in the proximal regions of the rest of the chromosomes of *D. buzzatii* st-1 and *D. mojavensis*. The available linkage maps for *D. buzzatii* and *D. mojavensis* (Schafer et al., 1993; Staten et al., 2004) are not very detailed; even so, we can assume that like in *D. melanogaster* these regions have a reduced recombination rate.

The X chromosome poses a challenge when trying to explain its TE dynamics. Because the X has a higher recombination rate than the autosomes, and

mutations are directly exposed to selection in hemizygous males, deleterious insertions should be removed more efficiently in the X chromosome than in the autosomes. An early analysis of the *D. melanogaster* reference genome showed a reduced accumulation of TEs in the *D. melanogaster* X chromosome (Bartolomé et al., 2002). However, recent analyses have surveyed several *D. melanogaster* populations and have not found evidence of a lower TE presence in the X chromosome, and some have even reported a higher abundance (Cridland et al., 2013; Petrov et al., 2011; Kofler et al., 2012). Our observations show that in *D. buzzatii* and *D. mojavensis* the X chromosome has a significantly higher TE density than the autosomes, except for the dot. And this difference remains even when the most proximal 3 Mb are discarded. Interestingly, the increase is sustained throughout the whole length of chromosome X in both species (Fig. 2). The X higher TE density is observed not only in *D. buzzatii* but also in *D. mojavensis*. Consequently, the assembly problem, that could have more impact on chromosome X as using males and female flies implies a lower coverage, does not seem to explain our results. The argument that some families with an insertion preference for the X have recently suffered an expansion in *D. melanogaster* (Cridland et al., 2013) is interesting and may suggest that *D. buzzatii* and *D. mojavensis* TEs are actively transposing. However, there are possibly other factors, besides recombination, needed to understand the unpredicted TE abundance in the X chromosome.

TEs and NGS

Issues with the NGS genomes repeats have been reported before (Alkan et al., 2011) suggesting that stringent assembly strategies and shorter reads do not produce an accurate representation of the repeats in a specific *locus* but a consensus built with sequences from other *loci* (Natali et al., 2013). Although dealing with different technologies, it resembles the case of *D. melanogaster* Release 3 (Celniker et al., 2002), where after extensive experimental efforts, most of the repetitive sequences of the previous release were found to be composite sequences of the newly sequenced TEs. Consequently, comparing the mobile fraction of the two strains of *D. buzzatii* between them (st-1 sequenced with a mixture of Sanger, Illumina and 454 reads and j-19 sequenced solely with Illumina reads) and to *D. mojavensis* genome (sequenced with Sanger reads) raised questions about the reliability of such comparisons.

To find out if the sequencing technology, and potentially the assembly methods, implied major differences in TE annotation, we look at published genomes

and their analyses of TE fractions. Two dozens of genomes of different *Drosophila* genus species have been released since *D. melanogaster* reference genome. Nevertheless, the mobile fraction of most of the recently published genomes has not been analyzed or has only been analyzed superficially (Zhou et al., 2012; Zhou and Bachtrog, 2012; Ometto et al., 2013; Fonseca et al., 2013) yet there are some exceptions (Chiu et al., 2013). At least two analyses comparing some of these genomes in a uniform manner have been published (Drosophila 12 Genomes Consortium, 2007; Ometto et al., 2013) but they yielded very different values. The main reasons seem to be the use of different annotation methods and updates in the TE libraries. The discrepancies between estimations compelled us to analyze all the *Drosophila* genus genomes available simultaneously, in the most homogeneous way possible and trying to reduce the unavoidable bias of library specificity. The values differ from previous studies but the comparisons should be more consistent. We found that genomes sequenced with Sanger technology have a higher TE percentage than those sequenced mainly with Illumina and 454 technologies. Because the data is not phylogenetically independent it is possible that species sequenced with one technology have actually a higher TE fraction than the ones sequenced with the other. However, from all the species from the same subgroup, sequenced with different technologies, the ones sequenced with Sanger show the highest TE percentage, suggesting that there is indeed an impact from the sequencing technology.

Correction in D. buzzatii st-1

We mapped the reads used in the *D. buzzatii* st-1 pre-assembly to the final assembly, following the lead of several projects that used high quality reference genomes and re-sequenced data from different individuals to accurately identify TE insertions (Fiston-Lavier et al., 2011; Petrov et al., 2011; Kofler et al., 2012; Jiang et al., 2015). The mapping showed how some regions annotated as TE insertions had a TE coverage depth much higher than the surrounding regions. We also noticed that some gaps had TE annotations from the same family on each side, suggesting that the gap should be filled with TE sequence. In order to obtain a reliable estimate and account for the problems related to NGS (see above), we directly counted how many read nucleotides belonged to TEs. One could argue that some of those reads may belong to the heterochromatin, were casted aside during the pre-assembly, and have been aligned now to euchromatin repeats. However, GS Reference Mapper aligned 20270 reads less in this process than those used by GS Reference Assembler. After mapping and dividing by

gene average coverage, we pulled the data for every order and superfamily together.

Sequence similarity among TE family copies is related to its transpositional activity. TE families which have recently transposed will contain highly similar copies and will be the most affected by the assembly problems mentioned before. Therefore, our correction method is expected to have a higher impact on these families. Our results show that LTR-retrotransposons were the most affected order. Their recent activity and their double repetitive nature, as not only LTR-retrotransposon copies will generate similar reads, but the LTRs from a single copy can produce reads susceptible to be assembled together are likely explanations. Additionally, LTR-retrotransposons are the longest TEs in Drosophila genomes, thus suffering more than other orders the artificial fragmentation by identification software (Feschotte et al., 2009) and assembly problems due to reads that do not span the lenght of the insertions. *Osvaldo* and *Isis* elements, from the Gypsy superfamily, were reported to be active in *D. buzzatii* (Labrador and Fontdevila, 1994; García Guerreiro and Fontdevila, 2007), which agrees with our ours results as Gypsy is the LTR-retrotransposon superfamily with a higher correction rate. The LINEs superfamilies R1 and R2 are nested within ribosomal regions, typically poorly assembled, explaining their underestimation in *D. buzzatii* st-1 genome (Xiong et al., 1988; Jakubczak et al., 1992). Helitrons presence in insects have been known for over a decade and are remarkably abundant in *D. melanogaster* genome (Locke et al., 1999; Kapitonov and Jurka, 2003). Yang and Barbash (Yang and Barbash, 2008) carried out and extensive analysis of *DINE-1* on the firsts 12 Drosophila genomes sequenced. Their analyses revealed that *D. mojavensis* is the second in number of *DINE-1* copies, than those copies had probably undergone multiple rounds of transposition and silencing, and some had been recently transposed. Previous studies have already identified several families of Helitrons in *D. buzzatii* named ISBu (for Insertion Sequence of *D. buzzatii*) in chromosomal inversion breakpoints (Cáceres et al., 2001; Delprat et al., 2009). We have now detected that over 800 kb of Helitrons were incorrectly assembled in *D. buzzatii* st-1, suggesting that 12.65% of the Helitrons have been recently transposed, while 5531 kb of Helitrons are either sequenced in reads with other regions, that allowed the assembler to map them, or are not as similar to confound the assembler. Hence, like in *D. mojavensis*, Helitrons, the most abundant order in *D. buzzatii* st-1, also seem to have undergone several rounds of activity.

Our methods has drawbacks; the correction does not inform of where the repeats are in the genome, or their specific sequence, an information that may not be precise in a NGS genome (see above). However, it is a method easy to apply that provides more acurate estimates of the abundance of each order and

superfamily. Therefore, our strategy facilitates comparisons among the wealth of already sequenced genomes and deepens our understanding of genome evolution.

COMPETING INTERESTS

The authors declare that they have no competing interests.

AUTHOR'S CONTRIBUTIONS

NR designed, and carried out the transposable element analyses and drafted the manuscript. YG assembled *D. buzzatii* j-19 genome. AD extracted DNA for sequencing and contributed to the analyses design. AK helped with transposable element analyses. CF contributed to design the study. AR conceived of the study, participated in its design and coordination and helped to draft the final manuscript. All authors read and approved the final manuscript.

ACKNOWLEDGEMENTS

This work was supported by grants BFU2008-04988 and BFU2011-30476 from the Spanish Ministerio de Ciencia e Innovación to A.R., grant R01GM077582 to C.F from the National Institutes of Health, and by PIF-UAB fellowship to N.R. We want to thank Jordi Camps, Marta Gut and Ivo G Gut from the Spanish Centro Nacional de Análisis Genómico (CNAG) for their collaboration with sequencing of *D. buzzatii* j-19 and also to Valentí Moncunill and David Torrent from Barcelona Supercomputing Center (BSC) for their collaboration with the genome assembly.

3.2.6 SUPPLEMENTARY MATERIAL

Supplementary information is available in the Appendix Section B.

4

DISCUSSION

The two Result Sections, 3.1 and 3.2, take two different approaches to TE analysis in genomes. The case described in "A divergent *P element* and its associated MITE, *BuT5*, generate chromosomal inversions and are widespread within the *Drosophila repleta* species group" deals with the in-depth analysis of a previously unclassified element, and with the detailed process to understand how, when, and to where it moved. A different approach was taken for the second section, "Exploration of the *D. buzzatii* transposable element content suggests underestimation of repeats in Drosophila genomes" where a bioinformatic survey of the TE content in a sequenced genome was carried out. The widened scope needed to answer the question of this later section, inevitably results in the loss of the detailed analysis that was achieved on the first.

In the Results, both the published paper and the submitted manuscript include a section where our findings are discussed with some detail. Thus, in this section I will only extend the discussion on those points that were briefly mentioned or not covered before.

4.1 MITES IN *DROSOPHILA GENUS* GENOMES

MITEs (Miniature Inverted-Repeat TEs) are short non-autonomous TIR element-derivatives with some characteristic features that set them apart from standard defective class II elements (see Introduction). In Section 3.1 of the Results the non-coding TE, *BuT5*, is classified as a MITE, and its autonomous counterpart, the P-element, is identified and described. BuT5 gathers all the requisites to be called a MITE: presence of TIRs, short length, no evidence of protein-coding capacity, and high copy number (Feschotte et al., 2002). Nevertheless, it is worth mentioning that Drosophila MITEs seem to have some differences when compared to MITEs in other species (particularly plants). First, most of Drosophila genus MITEs described so far are around 700 to 1000 bp (Holyoake and Kid-

well, 2003; Fonseca et al., 2012; Rossato et al., 2014; Wallau et al., 2014), unlike MITEs in other species that are usually under or around 500 bp. Second, MITE copy number in Drosophila is also smaller than in other species, were hundreds of MITEs copies can be found in a single genome (Feschotte et al., 2002; Quesneville et al., 2006; Yang et al., 2013).

The firsts true MITEs discovered in Drosophila genus, *Vege* and *Mar*, found in *D. willistoni*, were both assigned to the hAT superfamiliy, although no hAT autonomous element was recovered (Holyoake and Kidwell, 2003). *Vege* is 884 bp long and 6 to 8 copies per genome were detected; *Mar* is 610 bp long and only one copy was retrieved. Holyoake and Kidwell (2003) argued that Drosophila has fewer TIR transposons than other species, which would explain its low number of MITEs. In *D. willistoni* sequenced genome, 93 new *Mar* copies, likely mobilized by the hAT Buster family, were found. Additionally, four partial copies of a Buster element were recovered from *D. willistoni* genome, and an undetermined number from *D. tropicalis* (Deprá et al., 2012). *BuT5* case resembles *Mar* and other MITE examples where few or no elements able to supply the transposase were found (Holyoake and Kidwell, 2003; Fonseca et al., 2012). This seems to be in the line of the analysis reporting that only 22% of *D. melanogaster* genome transposons are full-length (Bartolomé et al., 2002). In *D. mojavensis* genome only three partial copies of the autonomous element were found, compared to the 48 partial and complete copies of *BuT5*. It is then possible that the lower number of TIR transposons in Drosophila provided few occasions for derivatives to appear (Kidwell, 2002; Feschotte and Pritham, 2007). However, as these examples show the scarcity of available transposase may also be accountable for Drosophila low number of MITEs. Both factors can be the result of Drosophila low TE fraction in general. TE abundance is positively correlated with genome size (Kidwell, 2002). Hence, the genome size of Drosophila species, smaller than those of others organisms where MITEs are frequent, such as mosquitoes and plants, would explain MITEs paucity (Casacuberta and Santiago, 2003; Jiang et al., 2004; Boulesteix and Biémont, 2005; Goubert et al., 2015).

Nevertheless, the shortage of TIR transposons, either as raw material for MITEs formation or as transposase suppliers, is not necessarily the only cause of Drosophila genus MITE patterns. Mariner-related MITEs found recently in the new Drosophila sequenced genomes (Wallau et al., 2014) were able to shed some light on the copy number differences. Wallau and collaborators found 724 copies of Mariner-related MITEs, most of them between 900 and 1000 bp, except for four lineages or subfamilies with approximately 460-560 bp. Interestingly the most successful were the shortest ones, reaching 314 copies in *D. eugracilis* genome.

Nevertheless, the reasons why Drosophila genus MITEs are generally longer than other species MITEs remains unknown.

4.2 THE LIFESPAN OF A MITE

The transposition selection equilibrium, previously seen as a common mechanism by which TEs were maintained in a population (Kidwell and Lisch, 2001; Le Rouzic and Deceliere, 2005) is now regarded as an uncommon state (Le Rouzic et al., 2007). Population genetics models challenged the idea of long-term stable equilibrium, and cycles of transposition activity were proposed as a mechanism to explain persistent TE invasions, particularly for short non-autonomous elements as SINEs and MITEs (Le Rouzic and Capy, 2006).

The idea that MITEs, as compact elements, can be more transpositionally successful than the full-length elements that originated them has been previously discussed (Feschotte et al., 2002). Several reasons have been argued to explain the abundance of MITEs. MITEs shortness may help them to evade epigenetic silencing, to which would also contribute their preferential insertion close to genes. Elimination through recombination would be less effective on shorter sequences. Additionally, the proximity of the TIRs could increase their transposition chances (Yaakov et al., 2013). The capacity of MITEs to use the transposases of different TE families can also explain its success (Fattash et al., 2013; Yang et al., 2009). Regardless of these factors, there are not predictions of population genetic models or enough examples of long-term transmitted MITEs to know how long MITEs can survive in a lineage.

BuT5 is present in 38 species of the *repleta* group with its last common ancestor approximately 16 million years ago (Mya). Although horizontal transfer (HT) events involving P elements have been described in several occasions (Clark et al., 1994, 1995; Clark and Kidwell, 1997; Silva and Kidwell, 2000) there is no evidence of *BuT5* being horizontally transferred within the *repleta* group. According to the phylogenetic reconstruction ((Rius et al., 2013), Figure 4), *BuT5* has been likely vertically transmitted in three species complexes (*mulleri*, *longicornis*, and *buzzatii*) that shared a common ancestor 14 Mya (Oliveira et al., 2012).

The *P* elements present in *D. buzzatii* and in *D. mojavensis* seem the only possible elements capable to share its transposase with *BuT5*. We detected a putatively autonomous copy in *D. buzzatii*. *In vitro* transposition experiments to test whether the P element transposase is able to mobilize *BuT5* were not performed. Even though, there are strong evidences supporting the P element and *BuT5* partnership, such as the highly similar terminal regions and the conservation of

THAP domain binding sites in both the MITEs, and the autonomous elements of *D. buzzatii* and *D. mojavensis* (Section 3.1 Figure 2 and Table 1). Additionally, the recent transposition of *BuT5* is supported by its presence as secondary colonizer of young inversions in Drosophila (Casals et al., 2003; Delprat et al., 2009) and by the similarity of the copies in *D. mojavensis* genome (Section 3.1 Figure 4). Consequently, we found a partnership between an autonomous and a non-autonomous element maintained for at least 14 million years. There is only one example of the approximate lifespan of a MITE in Drosophila genus, the Mar element that has been probably active for 5.7 million years (Deprá et al., 2012). Becoming the *BuT5* and P element partnership the longest described in Drosophila.

4.3 THE IMPORTANCE OF THOROUGH AND DETAILED ANALYSIS IN THE GENOMIC ERA

Genome sequencing has radically changed genetics research. The possibility to study a particular aspect of the biology of a species, population, individual or tissue, viewing the whole genome, or genomes, at the greatest resolution, the nucleotide, has opened an enormous field of possibilities with countless advantages. However, this new game-changing technology has some drawbacks or limitations.

The case of *BuT5* presented in Section 3.1 is an example of how despite of the advantages of analyzing multiple genomes simultaneously, detailed manual analysis is still valuable to discover the genome TE diversity. The *P* element described there belongs undoubtedly to the P element superfamily and family, according to the transposase analysis (Results, Section 3.1, Figure 7). However, it is fairly divergent from the rest of *Drosophila* genus *P* elements, in nucleotide sequence, transposase structure, and expression. Not long before the publication of the article included in Section 3.1, an automated search for *P* elements in the 12 Drosophila genomes sequenced at the time missed to find the partial copies of the *P* element in *D. mojavensis* (Loreto et al., 2012). Indeed, the similarity to other *P* elements is restricted to the transposase, and it is only significant when both nucleotide sequences are translated (tBLASTx). However, it is not difficult to imagine that if *D. mojavensis P* element was undetected by a wide genome search, biased by the prior knowledge of how a *P* element should look like, there may exist other unnoticed elements in the sequenced genomes. This case is somehow similar to the discovery of *DINE-1* in *D. melanogaster* reference genome, where it is the most abundant TE family, and yet was detected only

after the first TE annotation was published (Kapitonov and Jurka, 2003). These two examples reveal the importance of manual and detailed analysis to unveil the TE diversity of new genomes that automated and wide-range strategies can overlook.

5

CONCLUSIONS

1. In *D. mojavensis* sequenced genome *BuT5* complete insertions are on average 1014 bp, have 3-bp TIRs, 16-bp subTIRs, and no coding capacity.
2. *BuT5* is present in 38 species of the *D. repleta* group.
3. In 19 species of *mulleri*, *longicornis*, and *buzzatii* complexes *BuT5* has been likely vertically transmitted.
4. Partial copies of the *P* element are present in *D. mojavensis*.
5. The putative complete copy of *P* element in *D. buzzatii* is 3386 bp long, has 3-bp TIRs, 17-bp subTIRs as is flanked by a 9-bp TSD.
6. The *D. buzzatii P* element encodes a transposase with 822 residues in seven exons.
7. The *D. buzzatii P* element transposase has a THAP DNA binding domain in the N-terminus and a putative catalitic domain in the C-terminus with a DDE triad and a D(2)H motif.
8. The *D. mojavensis* consensus sequences of *P* element and *BuT5* have 99% identity over 98 bp at the 5′ end, and 90.9% identity over 263 bp at the 3′ end.
9. The *D. buzzatii P* element and the *D. buzzatii BuT5* consensus sequence have 99% identity over the first 96 bp at the 5′ end, and a 96.5% identity over the last 260 bp at the 3′ end.
10. In *D. buzzatii P* element and *BuT5* have two *P* element transposase binding sites (THAP domain binding sites) (8 bp), one at each end.
11. The *P* element present in *D. buzzatii* and *D. mojavensis* is likely *BuT5* master element.

12. In *D. buzzatii* head tissue all introns of the *P* element transposas are spliced, while in ovarian tissue only the last intron is spliced.

13. TEs account for the 8.43% of *D. buzzatii* st-1 genome assembly and for 4.15% of *D. buzzatii* j-19 genome assembly.

14. In *D. buzzatii* st-1 and in *D. mojavensis* chromosomes, considering only the mapped scaffolds, TE density follows similar pattern and increases at the chromosomes proximal end.

15. In *D. buzzatii* st-1 and in *D. mojavensis*, considering only the mapped scaffolds, chromosome 6 (dot) has a 41.22% and 46.30% of TEs respectively and a TE density window distribution significantly different from the rest of the chromosomes.

16. In *D. buzzatii* st-1 and in *D. mojavensis*, considering only the mapped scaffolds, chromosome X has a 8.32% and 11.81% of TEs respectively and a TE density window distribution significantly different from the rest of the chromosomes, except for the dot. The difference is maintained after discarding the three most proximal Mb.

17. *Drosophila* genus genomes sequenced with Sanger have a mean TE fraction of 19.31%. *Drosophila* genus genomes sequenced chiefly with NGS have a mean TE fraction of 10.98%.

18. According to our coverage-based correction method, the TE fraction of *D. buzzatii* st-1 genome is between 10.85% and 11.16%.

19. The LTR-retrotransposons are the most underestimated order in *D. buzzatii* st-1 annotation and have probably been active recently.

20. The abundance of TE orders follow the same order in *D. buzzatii* st-1 and in *D. mojavensis*. Helitrons are the most abundant order and had probably undergone several activity periods in *D. buzzatii*.

BIBLIOGRAPHY

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, and others (2000). The genome sequence of Drosophila melanogaster. *Science (New York, N.Y.)*, 287(5461):2185–2195.

Akagi, K., Li, J., and Symer, D. E. (2013). How do mammalian transposons induce genetic variation? A conceptual framework: the age, structure, allele frequency, and genome context of transposable elements may define their wide-ranging biological impacts. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, 35(4):397–407.

Alkan, C., Sajjadian, S., and Eichler, E. E. (2011). Limitations of next-generation genome sequence assembly. *Nature Methods*, 8(1):61–65.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402.

Anxolabéhère, D., Kidwell, M. G., and Periquet, G. (1988). Molecular characteristics of diverse populations are consistent with the hypothesis of a recent invasion of Drosophila melanogaster by mobile P elements. *Molecular Biology and Evolution*, 5(3):252–269.

Ashburner, M. and Bergman, C. M. (2005). Drosophila melanogaster: a case study of a model genomic sequence and its consequences. *Genome Research*, 15(12):1661–1667.

Bao, W., Jurka, M. G., Kapitonov, V. V., and Jurka, J. (2009). New superfamilies of eukaryotic DNA transposons and their internal divisions. *Molecular Biology and Evolution*, 26(5):983–993.

Bao, Z. and Eddy, S. R. (2002). Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Research*, 12(8):1269–1276.

Barrón, M. G., Fiston-Lavier, A.-S., Petrov, D. A., and González, J. (2014). Population genomics of transposable elements in Drosophila. *Annual Review of Genetics*, 48:561–581.

Bartolomé, C., Bello, X., and Maside, X. (2009). Widespread evidence for horizontal transfer of transposable elements across Drosophila genomes. *Genome Biology*, 10(2):R22.

Bartolomé, C., Maside, X., and Charlesworth, B. (2002). On the abundance and distribution of transposable elements in the genome of Drosophila melanogaster. *Molecular Biology and Evolution*, 19(6):926–937.

Bergman, C. M. and Quesneville, H. (2007). Discovering and detecting transposable elements in genome sequences. *Briefings in Bioinformatics*, 8(6):382–392.

Bergman, C. M., Quesneville, H., Anxolabéhère, D., and Ashburner, M. (2006). Recurrent insertion and duplication generate networks of transposable element sequences in the Drosophila melanogaster genome. *Genome Biology*, 7(11):R112.

Bingham, P. M., Levis, R., and Rubin, G. M. (1981). Cloning of DNA sequences from the white locus of D. melanogaster by a novel and general method. *Cell*, 25(3):693–704.

Boulesteix, M. and Biémont, C. (2005). Transposable elements in mosquitoes. *Cytogenetic and Genome Research*, 110(1-4):500–509.

Boulesteix, M., Weiss, M., and Biémont, C. (2006). Differences in genome size between closely related species: the Drosophila melanogaster species subgroup. *Molecular Biology and Evolution*, 23(1):162–167.

Britten, R. J. and 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 (New York, N.Y.)*, 161(3841):529–540.

Buisine, N., Quesneville, H., and Colot, V. (2008). Improved detection and annotation of transposable elements in sequenced genomes using multiple reference sequence sets. *Genomics*, 91(5):467–475.

Bureau, T. E. and Wessler, S. R. (1992). Tourist: a large family of small inverted repeat elements frequently associated with maize genes. *The Plant Cell*, 4(10):1283–1294.

Bureau, T. E. and Wessler, S. R. (1994). Stowaway: a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants. *The Plant Cell*, 6(6):907–916.

C. elegans Sequencing Consortium (1998). Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science (New York, N.Y.)*, 282(5396):2012–2018.

Casacuberta, E. and González, J. (2013). The impact of transposable elements in environmental adaptation. *Molecular Ecology*, 22(6):1503–1517.

Casacuberta, J. M. and Santiago, N. (2003). Plant LTR-retrotransposons and MITEs: control of transposition and impact on the evolution of plant genes and genomes. *Gene*, 311:1–11.

Casals, F., Cáceres, M., Manfrin, M. H., González, J., and Ruiz, A. (2005). Molecular characterization and chromosomal distribution of Galileo, Kepler and Newton, three foldback transposable elements of the Drosophila buzzatii species complex. *Genetics*, 169(4):2047–2059.

Casals, F., Cáceres, M., and Ruiz, A. (2003). The foldback-like transposon Galileo is involved in the generation of two different natural chromosomal inversions of Drosophila buzzatii. *Molecular Biology and Evolution*, 20(5):674–685.

Casals, F., González, J., and Ruiz, A. (2006). Abundance and chromosomal distribution of six Drosophila buzzatii transposons: BuT1, BuT2, BuT3, BuT4, BuT5, and BuT6. *Chromosoma*, 115(5):403–412.

Casola, C., Lawing, A. M., Betrán, E., and Feschotte, C. (2007). PIF-like transposons are common in drosophila and have been repeatedly domesticated to generate new host genes. *Molecular Biology and Evolution*, 24(8):1872–1888.

Celniker, S. E., Wheeler, D. A., Kronmiller, B., Carlson, J. W., Halpern, A., et al. (2002). Finishing a whole-genome shotgun: release 3 of the Drosophila melanogaster euchromatic genome sequence. *Genome Biology*, 3(12):RESEARCH0079.

Charlesworth, B. and Barton, N. (2004). Genome size: does bigger mean worse? *Current biology: CB*, 14(6):R233–235.

Chen, Z.-X. et al. (2014). Comparative validation of the D. melanogaster modENCODE transcriptome annotation. *Genome Research*, 24(7):1209–1223.

Chiu, J. C., Jiang, X., Zhao, L., Hamm, C. A., Cridland, J. M., et al. (2013). Genome of Drosophila suzukii, the spotted wing drosophila. *G3 (Bethesda, Md.)*, 3(12):2257–2271.

Clark, J. B., Altheide, T. K., Schlosser, M. J., and Kidwell, M. G. (1995). Molecular evolution of P transposable elements in the genus Drosophila. I. The saltans and willistoni species groups. *Molecular Biology and Evolution*, 12(5):902–913.

Clark, J. B. and Kidwell, M. G. (1997). A phylogenetic perspective on P transposable element evolution in Drosophila. *Proceedings of the National Academy of Sciences of the United States of America*, 94(21):11428–11433.

Clark, J. B., Maddison, W. P., and Kidwell, M. G. (1994). Phylogenetic analysis supports horizontal transfer of P transposable elements. *Molecular Biology and Evolution*, 11(1):40–50.

Comeron, J. M., Ratnappan, R., and Bailin, S. (2012). The many landscapes of recombination in Drosophila melanogaster. *PLoS genetics*, 8(10):e1002905.

Cridland, J. M., Macdonald, S. J., Long, A. D., and Thornton, K. R. (2013). Abundance and distribution of transposable elements in two Drosophila QTL mapping resources. *Molecular Biology and Evolution*, 30(10):2311–2327.

Cáceres, M., Puig, M., and Ruiz, A. (2001). Molecular characterization of two natural hotspots in the Drosophila buzzatii genome induced by transposon insertions. *Genome Research*, 11(8):1353–1364.

Cáceres, M., Ranz, J. M., Barbadilla, A., Long, M., and Ruiz, A. (1999). Generation of a widespread Drosophila inversion by a transposable element. *Science (New York, N.Y.)*, 285(5426):415–418.

Daniels, S. B., Peterson, K. R., Strausbaugh, L. D., Kidwell, M. G., and Chovnick, A. (1990). Evidence for horizontal transmission of the P transposable element between Drosophila species. *Genetics*, 124(2):339–355.

David, J. R. and Tsacas, L. (1981). Cosmopolitan, subcosmopolitan and widespread species: different strategies within the Drosophilid family (Diptera). *C. R. Soc Biogéo*, (57):11–26.

de Freitas Ortiz, M., Lorenzatto, K. R., Corrêa, B. R. S., and Loreto, E. L. S. (2010). hAT transposable elements and their derivatives: an analysis in the 12 Drosophila genomes. *Genetica*, 138(6):649–655.

de Freitas Ortiz, M. and Loreto, E. L. S. (2009). Characterization of new hAT transposable elements in 12 Drosophila genomes. *Genetica*, 135(1):67–75.

Delprat, A., Negre, B., Puig, M., and Ruiz, A. (2009). The transposon Galileo generates natural chromosomal inversions in Drosophila by ectopic recombination. *PloS One*, 4(11):e7883.

DeMarco, R., Venancio, T. M., and Verjovski-Almeida, S. (2006). SmTRC1, a novel Schistosoma mansoni DNA transposon, discloses new families of animal and fungi transposons belonging to the CACTA superfamily. *BMC evolutionary biology*, 6:89.

Deprá, M., Ludwig, A., Valente, V. L., and Loreto, E. L. (2012). Mar, a MITE family of hAT transposons in Drosophila. *Mobile DNA*, 3(1):13.

Doolittle, W. F. and Sapienza, C. (1980). Selfish genes, the phenotype paradigm and genome evolution. *Nature*, 284(5757):601–603.

Drosophila 12 Genomes Consortium (2007). Evolution of genes and genomes on the Drosophila phylogeny. *Nature*, 450(7167):203–218.

Drysdale, R. A., Crosby, M. A., and Consortium, T. F. (2005). FlyBase: genes and gene models. *Nucleic Acids Research*, 33(suppl 1):D390–D395.

Edgar, R. C. and Myers, E. W. (2005). PILER: identification and classification of genomic repeats. *Bioinformatics (Oxford, England)*, 21 Suppl 1:i152–158.

English, A. C., Richards, S., Han, Y., Wang, M., Vee, V., et al. (2012). Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology. *PLoS ONE*, 7(11):e47768.

Etges, W., Johnson, W., Duncan, G., Huckins, G., and Heed, W. (1999). Ecological genetics of cactophilic Drosophila. In *Ecology of Sonoran Desert plants and plant communities*, pages 164–214. University of Arizona Press, Tucson (AZ).

Fattash, I., Rooke, R., Wong, A., Hui, C., Luu, T., Bhardwaj, P., and Yang, G. (2013). Miniature inverted-repeat transposable elements: discovery, distribution, and activity. *Genome / National Research Council Canada = Génome / Conseil National De Recherches Canada*, 56(9):475–486.

Fedoroff, N. V. (2012). McClintock's challenge in the 21st century. *Proceedings of the National Academy of Sciences*, 109(50):20200–20203.

Fellows, D. and Heed, W. (1972). Factors Affecting Host Plant Selection in Desert-Adapted Cactiphilic Drosophila. *Ecology*, 53(5):850–&. WOS:A1972N884000008.

Feschotte, C. (2008). Transposable elements and the evolution of regulatory networks. *Nature Reviews. Genetics*, 9(5):397–405.

Feschotte, C., Keswani, U., Ranganathan, N., Guibotsy, M. L., and Levine, D. (2009). Exploring repetitive DNA landscapes using REPCLASS, a tool that automates the classification of transposable elements in eukaryotic genomes. *Genome Biology and Evolution*, 1:205–220.

Feschotte, C. and Pritham, E. J. (2007). DNA transposons and the evolution of eukaryotic genomes. *Annual Review of Genetics*, 41:331–368.

Feschotte, C., Zhang, X., and Wessler, S. R. (2002). Miniature Inverted-repeat Transpo sable Elements (MITEs) and their Relationship with Estab lished DNA Transposons. In *Mobile DNA II*. ASM Press, Washington D. C.

Finnegan, D. J. (1989). Eukaryotic transposable elements and genome evolution. *Trends in genetics: TIG*, 5(4):103–107.

Fiston-Lavier, A.-S., Carrigan, M., Petrov, D. A., and González, J. (2011). T-lex: a program for fast and accurate assessment of transposable element presence using next-generation sequencing data. *Nucleic Acids Research*, 39(6):e36.

Flutre, T., Duprat, E., Feuillet, C., and Quesneville, H. (2011). Considering transposable element diversification in de novo annotation approaches. *PloS One*, 6(1):e16526.

Fonseca, N. A., Morales-Hojas, R., Reis, M., Rocha, H., Vieira, C. P., Nolte, V., Schlötterer, C., and Vieira, J. (2013). Drosophila americana as a model species for comparative studies on the molecular basis of phenotypic variation. *Genome Biology and Evolution*, 5(4):661–679.

Fonseca, N. A., Vieira, C. P., Schlötterer, C., and Vieira, J. (2012). The DAIBAM MITE element is involved in the origin of one fixed and two polymorphic Drosophila virilis phylad inversions. *Fly*, 6(2):71–74.

García Guerreiro, M. P. and Fontdevila, A. (2007). Molecular characterization and genomic distribution of Isis: a new retrotransposon of Drosophila buzzatii. *Molecular genetics and genomics: MGG*, 277(1):83–95.

Gardner, M. J., Hall, N., Fung, E., White, O., Berriman, M., et al. (2002). Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature*, 419(6906):498–511.

Goubert, C., Modolo, L., Vieira, C., ValienteMoro, C., Mavingui, P., and Boulesteix, M. (2015). De Novo Assembly and Annotation of the Asian Tiger Mosquito (Aedes albopictus) Repeatome with dnaPipeTE from Raw Genomic Reads and Comparative Analysis with the Yellow Fever Mosquito (Aedes aegypti). *Genome Biology and Evolution*, 7(4):1192–1205.

Green, M. M. (2010). 2010: A century of Drosophila genetics through the prism of the white gene. *Genetics*, 184(1):3–7.

Gregory, T. R. (2004). Insertion-deletion biases and the evolution of genome size. *Gene*, 324:15–34.

Guillén, Y., Rius, N., Delprat, A., Williford, A., Muyas, F., et al. (2015). Genomics of ecological adaptation in cactophilic Drosophila. *Genome Biology and Evolution*, 7(1):349–366.

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

Hardman, N. (1986). Structure and function of repetitive DNA in eukaryotes. *Biochemical Journal*, 234(1):1–11.

Haring, n., Hagemann, n., and Pinsker, n. (1998). Transcription and splicing patterns of M- and O-type P elements in drosophila bifasciata, D. helvetica, and scaptomyza pallida. *Journal of Molecular Evolution*, 46(5):542–551.

Hasson, E., Naveira, Horacio, and Fontdevila, A. (1992). The breeding sites of Argentinian cactophilic species of the Drosophila mulleri complex. *Revista Chilena de Historia natural*, 65(3):319–326.

Heed, W. and Mangan, R. (1986). Community ecology of the Sonoran Desert Drosophila. In *The genetics and biology of Drosophila*, volume 3, pages 311–345. Academic Press, London.

Hirakata, S. and Siomi, M. C. (2015). piRNA biogenesis in the germline: From transcription of piRNA genomic sources to piRNA maturation. *Biochimica Et Biophysica Acta*.

Holyoake, A. J. and Kidwell, M. G. (2003). Vege and Mar: two novel hAT MITE families from Drosophila willistoni. *Molecular Biology and Evolution*, 20(2):163–167.

Hoskins, R. A., Carlson, J. W., Kennedy, C., Acevedo, D., Evans-Holm, et al. (2007). Sequence finishing and mapping of Drosophila melanogaster heterochromatin. *Science (New York, N.Y.)*, 316(5831):1625–1628.

Hoskins, R. A., Carlson, J. W., Wan, K. H., Park, S., Mendez, I., et al. (2015). The Release 6 reference sequence of the Drosophila melanogaster genome. *Genome Research*, 25(3):445–458.

Hoskins, R. A., Smith, C. D., Carlson, J. W., Carvalho, A. B., Halpern, A., et al. (2002). Heterochromatic sequences in a Drosophila whole-genome shotgun assembly. *Genome Biology*, 3(12):RESEARCH0085.

Hu, T. T., Eisen, M. B., Thornton, K. R., and Andolfatto, P. (2013). A second-generation assembly of the Drosophila simulans genome provides new insights into patterns of lineage-specific divergence. *Genome Research*, 23(1):89–98.

Hua-Van, A., Le Rouzic, A., Boutin, T. S., Filée, J., and Capy, P. (2011). The struggle for life of the genome's selfish architects. *Biology Direct*, 6:19.

Huddleston, J., Ranade, S., Malig, M., Antonacci, F., Chaisson, M., et al. (2014). Reconstructing complex regions of genomes using long-read sequencing technology. *Genome Research*, page gr.168450.113.

Jakubczak, J. L., Zenni, M. K., Woodruff, R. C., and Eickbush, T. H. (1992). Turnover of R1 (type I) and R2 (type II) retrotransposable elements in the ribosomal DNA of Drosophila melanogaster. *Genetics*, 131(1):129–142.

Jiang, C., Chen, C., Huang, Z., Liu, R., and Verdier, J. (2015). ITIS, a bioinformatics tool for accurate identification of transposon insertion sites using next-generation sequencing data. *BMC bioinformatics*, 16:72.

Jiang, N., Feschotte, C., Zhang, X., and Wessler, S. R. (2004). Using rice to understand the origin and amplification of miniature inverted repeat transposable elements (MITEs). *Current Opinion in Plant Biology*, 7(2):115–119.

Jurka, J., Kapitonov, V. V., Kohany, O., and Jurka, M. V. (2007). Repetitive sequences in complex genomes: structure and evolution. *Annual Review of Genomics and Human Genetics*, 8:241–259.

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research*, 110(1-4):462–467.

Kaminker, J. S., Bergman, C. M., Kronmiller, B., Carlson, J., Svirskas, R., Patel, S., Frise, E., Wheeler, D. A., Lewis, S. E., Rubin, G. M., Ashburner, M., and Celniker, S. E. (2002). The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective. *Genome Biology*, 3(12):RESEARCH0084.

Kapitonov, V. V. and 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(11):6569–6574.

Kapitonov, V. V. and Jurka, J. (2008). A universal classification of eukaryotic transposable elements implemented in Repbase. *Nature Reviews Genetics*, 9(5):411–412.

Kidwell, M. G. (2002). Transposable elements and the evolution of genome size in eukaryotes. *Genetica*, 115(1):49–63.

Kidwell, M. G. and Lisch, D. R. (2001). Perspective: transposable elements, parasitic DNA, and genome evolution. *Evolution; International Journal of Organic Evolution*, 55(1):1–24.

Kofler, R., Betancourt, A. J., and Schlötterer, C. (2012). Sequencing of pooled DNA samples (Pool-Seq) uncovers complex dynamics of transposable element insertions in Drosophila melanogaster. *PLoS genetics*, 8(1):e1002487.

Labrador, M. and Fontdevila, A. (1994). High transposition rates of Osvaldo, a new Drosophila buzzatii retrotransposon. *Molecular & general genetics: MGG*, 245(6):661–674.

Le Rouzic, A., Boutin, T. S., and Capy, P. (2007). Long-term evolution of transposable elements. *Proceedings of the National Academy of Sciences of the United States of America*, 104(49):19375–19380.

Le Rouzic, A. and Capy, P. (2006). Population genetics models of competition between transposable element subfamilies. *Genetics*, 174(2):785–793.

Le Rouzic, A. and Deceliere, G. (2005). Models of the population genetics of transposable elements. *Genetical Research*, 85(3):171–181.

Lerat, E. (2010). Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. *Heredity*, 104(6):520–533.

Leung, W., Shaffer, C. D., Reed, L. K., Smith, S. T., Barshop, W., et al. (2015). Drosophila Muller F Elements Maintain a Distinct Set of Genomic Properties Over 40 Million Years of Evolution. *G3: Genes|Genomes|Genetics*, 5(5):719–740.

Li, R., Ye, J., Li, S., Wang, J., Han, Y., Ye, C., Wang, J., Yang, H., Yu, J., Wong, G. K.-S., and Wang, J. (2005). ReAS: Recovery of ancestral sequences for transposable elements from the unassembled reads of a whole genome shotgun. *PLoS computational biology*, 1(4):e43.

Li, Y., Hu, Y., Bolund, L., and Wang, J. (2010). State of the art de novo assembly of human genomes from massively parallel sequencing data. *Human Genomics*, 4(4):271–277.

Locke, J., Howard, L. T., Aippersbach, N., Podemski, L., and Hodgetts, R. B. (1999). The characterization of DINE-1, a short, interspersed repetitive element present on chromosome and in the centric heterochromatin of Drosophila melanogaster. *Chromosoma*, 108(6):356–366.

Loreto, E. L. S., Zambra, F. M. B., Ortiz, M. F., and Robe, L. J. (2012). New Drosophila P-like elements and reclassification of Drosophila P-elements subfamilies. *Molecular genetics and genomics: MGG*, 287(7):531–540.

Lynch, M. (2007). The frailty of adaptive hypotheses for the origins of organismal complexity. *Proceedings of the National Academy of Sciences*, 104(suppl 1):8597–8604.

Mackay, T. F. C., Richards, S., Stone, E. A., Barbadilla, A., Ayroles, J. F., et al. (2012). The Drosophila melanogaster Genetic Reference Panel. *Nature*, 482(7384):173–178.

Marzo, M., Puig, M., and Ruiz, A. (2008). The Foldback-like element Galileo belongs to the P superfamily of DNA transposons and is widespread within the Drosophila genus. *Proceedings of the National Academy of Sciences of the United States of America*, 105(8):2957–2962.

Matthews, B. B., Dos Santos, G., Crosby, M. A., Emmert, D. B., St Pierre, S. E., et al. (2015). Gene Model Annotations for Drosophila melanogaster: Impact of High-Throughput Data. *G3 (Bethesda, Md.)*, 5(8):1721–1736.

Matthews, K. A., Kaufman, T. C., and Gelbart, W. M. (2005). Research resources for Drosophila: the expanding universe. *Nature Reviews. Genetics*, 6(3):179–193.

McClintock, B. (1950). The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences of the United States of America*, 36(6):344–355.

McClintock, B. (1983). THE SIGNIFICANCE OF RESPONSES OF THE GENOME TO CHALLENGE. In *Nobel Lecture*.

McCoy, R. C., Taylor, R. W., Blauwkamp, T. A., Kelley, J. L., Kertesz, M., Pushkarev, D., Petrov, D. A., and Fiston-Lavier, A.-S. (2014). Illumina TruSeq synthetic long-reads empower de novo assembly and resolve complex, highly-repetitive transposable elements. *PloS One*, 9(9):e106689.

Milligan, B. G. (1998). Total DNA Isolation. In Hoelzel, A. R., editor, *Molecular Genetic Analysis of Populations: A Practical Approach*. IRL Press | Practical Approach Series 187, second edition edition.

Myers, E. W., Sutton, G. G., Delcher, A. L., Dew, I. M., Fasulo, D. P., et al. (2000). A whole-genome assembly of Drosophila. *Science (New York, N.Y.)*, 287(5461):2196–2204.

Narzisi, G. and Mishra, B. (2011). Comparing De Novo Genome Assembly: The Long and Short of It. *PLoS ONE*, 6(4):e19175.

Natali, L., Cossu, R. M., Barghini, E., Giordani, T., Buti, M., et al. (2013). The repetitive component of the sunflower genome as shown by different procedures for assembling next generation sequencing reads. *BMC genomics*, 14:686.

Oliveira, D. C. S. G., Almeida, F. C., O'Grady, P. M., Armella, M. A., DeSalle, R., and Etges, W. J. (2012). Monophyly, divergence times, and evolution of host plant use inferred from a revised phylogeny of the Drosophila repleta species group. *Molecular Phylogenetics and Evolution*, 64(3):533–544.

Ometto, L., Cestaro, A., Ramasamy, S., Grassi, A., Revadi, S., et al. (2013). Linking genomics and ecology to investigate the complex evolution of an invasive Drosophila pest. *Genome Biology and Evolution*, 5(4):745–757.

Orgel, L. E. and Crick, F. H. (1980). Selfish DNA: the ultimate parasite. *Nature*, 284(5757):604–607.

Pagel, M. and Johnstone, R. A. (1992). Variation across species in the size of the nuclear genome supports the junk-DNA explanation for the C-value paradox. *Proceedings. Biological Sciences / The Royal Society*, 249(1325):119–124.

Petrov, D. A., Fiston-Lavier, A.-S., Lipatov, M., Lenkov, K., and González, J. (2011). Population genomics of transposable elements in Drosophila melanogaster. *Molecular Biology and Evolution*, 28(5):1633–1644.

Petrov, D. A., Sangster, T. A., Johnston, J. S., Hartl, D. L., and Shaw, K. L. (2000). Evidence for DNA loss as a determinant of genome size. *Science (New York, N.Y.)*, 287(5455):1060–1062.

Piccinali, R. V., Mascord, L. J., Barker, J. S. F., Oakeshott, J. G., and Hasson, E. (2007). Molecular population genetics of the alpha-esterase5 gene locus in original and colonized populations of Drosophila buzzatii and its sibling Drosophila koepferae. *Journal of Molecular Evolution*, 64(2):158–170.

Pinsker, W., Haring, E., Hagemann, S., and Miller, W. J. (2001). The evolutionary life history of P transposons: from horizontal invaders to domesticated neogenes. *Chromosoma*, 110(3):148–158.

Piñol, J., Francino, O., Fontdevila, A., and Cabré, O. (1988). Rapid isolation of Drosophila high molecular weight DNA to obtain genomic libraries. *Nucleic Acids Research*, 16(6):2736.

Prada, C. (2010). *Evolución cromosómica del cluster Drosophila martensis: orígen de las inversiones y reutilización de los puntos de rotura*. PhD thesis, Universitat Autònoma de Barcelona, Barcelona (Spain).

Price, A. L., Jones, N. C., and Pevzner, P. A. (2005). De novo identification of repeat families in large genomes. *Bioinformatics (Oxford, England)*, 21 Suppl 1:i351–358.

Quesneville, H., Nouaud, D., and Anxolabéhère, D. (2006). P elements and MITE relatives in the whole genome sequence of Anopheles gambiae. *BMC genomics*, 7:214.

Ray, D. A., Feschotte, C., Pagan, H. J., Smith, J. D., Pritham, E. J., Arensburger, P., Atkinson, P. W., and Craig, N. L. (2008). Multiple waves of recent DNA transposon activity in the bat, Myotis lucifugus. *Genome Research*, 18(5):717–728.

Rebollo, R., Romanish, M. T., and Mager, D. L. (2012). Transposable elements: an abundant and natural source of regulatory sequences for host genes. *Annual Review of Genetics*, 46:21–42.

Richards, S., Liu, Y., Bettencourt, B. R., Hradecky, P., Letovsky, S., et al. (2005). Comparative genome sequencing of Drosophila pseudoobscura: chromosomal, gene, and cis-element evolution. *Genome Research*, 15(1):1–18.

Ricker, N., Qian, H., and Fulthorpe, R. R. (2012). The limitations of draft assemblies for understanding prokaryotic adaptation and evolution. *Genomics*, 100(3):167–175.

Rio, D. C. (2002). P TransposableElements in Drosophila melanogaster. In *Mobile DNA II*. ASM Press, Washington D. C.

Rizzon, C., Marais, G., Gouy, M., and Biémont, C. (2002). Recombination rate and the distribution of transposable elements in the Drosophila melanogaster genome. *Genome Research*, 12(3):400–407.

Rossato, D. O., Ludwig, A., Deprá, M., Loreto, E. L. S., Ruiz, A., and Valente, V. L. S. (2014). BuT2 is a member of the third major group of hAT transposons and is involved in horizontal transfer events in the genus Drosophila. *Genome Biology and Evolution*, 6(2):352–365.

Rubin, G. M., Kidwell, M. G., and Bingham, P. M. (1982). The molecular basis of P-M hybrid dysgenesis: the nature of induced mutations. *Cell*, 29(3):987–994.

Ruiz, A., Cansian, A. M., Kuhn, G. C., Alves, M. A., and Sene, F. M. (2000). The Drosophila serido speciation puzzle: putting new pieces together. *Genetica*, 108(3):217–227.

Ruiz, A. and Heed, W. B. (1988). Host-Plant Specificity in the Cactophilic Drosophila mulleri Species Complex. *Journal of Animal Ecology*, 57(1):237–249.

Ruiz, A. and Wasserman, M. (1993). Evolutionary cytogenetics of the Drosophila buzzatii species complex. *Heredity*, 70 (Pt 6):582–596.

Saha, S., Bridges, S., Magbanua, Z. V., and Peterson, D. G. (2008). Empirical comparison of ab initio repeat finding programs. *Nucleic Acids Research*, 36(7):2284–2294.

Salzberg, S. L., Phillippy, A. M., Zimin, A., Puiu, D., Magoc, T., et al. (2012). GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Research*, 22(3):557–567.

Salzberg, S. L. and Yorke, J. A. (2005). Beware of mis-assembled genomes. *Bioinformatics (Oxford, England)*, 21(24):4320–4321.

Schaack, S., Gilbert, C., and Feschotte, C. (2010). Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. *Trends in Ecology & Evolution*, 25(9):537–546.

Schaeffer, S. W., Bhutkar, A., McAllister, B. F., Matsuda, M., Matzkin, L. M., et al. (2008). Polytene chromosomal maps of 11 Drosophila species: the order of genomic scaffolds inferred from genetic and physical maps. *Genetics*, 179(3):1601–1655.

Schafer, D. J., Fredline, D. K., Knibb, W. R., Green, M. M., and Barker, J. S. F. (1993). Genetics and Linkage Mapping of Drosophila buzzatii. *Journal of Heredity*, 84(3):188–194.

Silva, J. C. and Kidwell, M. G. (2000). Horizontal transfer and selection in the evolution of P elements. *Molecular Biology and Evolution*, 17(10):1542–1557.

Siomi, M. C., Sato, K., Pezic, D., and Aravin, A. A. (2011). PIWI-interacting small RNAs: the vanguard of genome defence. *Nature Reviews. Molecular Cell Biology*, 12(4):246–258.

Smit, A. and Hubley, R. (2008). RepeatModeler Open-1.0. <http://www.repeatmasker.org>.

Smit, A., Hubley, R., and Green, P. (1996). RepeatMasker Open-3.0. <http://www.repeatmasker.org>.

Staten, R., Schully, S. D., and Noor, M. A. (2004). A microsatellite linkage map of Drosophila mojavensis. *BMC Genetics*, 5(1):12.

Wallau, G. L., Capy, P., Loreto, E., and Hua-Van, A. (2014). Genomic landscape and evolutionary dynamics of mariner transposable elements within the Drosophila genus. *BMC genomics*, 15:727.

Wallau, G. L., Ortiz, M. F., and Loreto, E. L. S. (2012). Horizontal transposon transfer in eukarya: detection, bias, and perspectives. *Genome Biology and Evolution*, 4(8):689–699.

Wasserman, M. (1992). Cytological evolution of the Drosophila repleta species group. In *Drosophila inversion polymorphism*. CRC Press, Boca Raton FL.

Watson, J. D. and Crick, F. H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738.

Wicker, T., Guyot, R., Yahiaoui, N., and Keller, B. (2003). CACTA transposons in Triticeae. A diverse family of high-copy repetitive elements. *Plant Physiology*, 132(1):52–63.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., et al. (2007). A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics*, 8(12):973–982.

Wyngaard, G. A., Rasch, E. M., Manning, N. M., Gasser, K., and Domangue, R. (2005). The relationship between genome size, development rate, and body size in copepods. *Hydrobiologia*, 532(1-3):123–137.

Xiong, Y., Burke, W. D., Jakubczak, J. L., and Eickbush, T. H. (1988). Ribosomal DNA insertion elements R1bm and R2bm can transpose in a sequence specific manner to locations outside the 28s genes. *Nucleic Acids Research*, 16(22):10561–10573.

Yaakov, B., Ben-David, S., and Kashkush, K. (2013). Genome-wide analysis of Stowaway-like MITEs in wheat reveals high sequence conservation, gene association, and genomic diversification. *Plant Physiology*, 161(1):486–496.

Yang, G., Fattash, I., Lee, C.-N., Liu, K., and Cavinder, B. (2013). Birth of three stowaway-like MITE families via microhomology-mediated miniaturization of a Tc1/Mariner element in the yellow fever mosquito. *Genome Biology and Evolution*, 5(10):1937–1948.

Yang, G., Nagel, D. H., Feschotte, C., Hancock, C. N., and Wessler, S. R. (2009). Tuned for transposition: molecular determinants underlying the hyperactivity of a Stowaway MITE. *Science (New York, N.Y.)*, 325(5946):1391–1394.

Yang, H.-P. and Barbash, D. A. (2008). Abundant and species-specific DINE-1 transposable elements in 12 Drosophila genomes. *Genome Biology*, 9(2):R39.

Young, M. W. (1979). Middle repetitive DNA: a fluid component of the Drosophila genome. *Proceedings of the National Academy of Sciences of the United States of America*, 76(12):6274–6278.

Yuan, Y.-W. and Wessler, S. R. (2011). The catalytic domain of all eukaryotic cut-and-paste transposase superfamilies. *Proceedings of the National Academy of Sciences of the United States of America*, 108(19):7884–7889.

Zhou, Q. and Bachtrog, D. (2012). Sex-specific adaptation drives early sex chromosome evolution in Drosophila. *Science (New York, N.Y.)*, 337(6092):341–345.

Zhou, Q., Zhu, H.-m., Huang, Q.-f., Zhao, L., Zhang, G.-j., et al. (2012). Deciphering neo-sex and B chromosome evolution by the draft genome of Drosophila albomicans. *BMC genomics*, 13:109.

Appendices

A

SUPPLEMENTARY MATERIAL OF *BUT5* AND THE *P ELEMENT* IN *D. REPLETA* GROUP

A.1 *P ELEMENT* TRANSPOSASE ALINGMENT

P element transposase alingment

>Dvir\1360
MAEKKEI----------XISSSKCSLRHCQKSRRDFPG----IKLFKFSERDPXILKQWAEK--C--
NFSEDFVANSSFFFLCQYYFAPE----------KIEC-
RYLKRGTIPDRNVTLVYNCTSKANYL----------
NVLISPSRKRKR----------
SQSPAEICRQCHKNTTTLKYYQKQYFSFKKIAEARKIRINLLK----------
RENTNLKQKIKRLKVNAYE----------
SLISKVDKMDLTGNEPMMIKMLLKEKXGTCLRWNDSE------
KLFAQGIFYRSTSTYKFLR----------
DSLQLNLPSPSSLQKWNSIKKLQPGDNECLYSAFKDAIKGISECDKECILTCDEVAIEKNWTY
NTSVDAIDGLEHL-----
LERSNKMGSHICVFVLRVIFKKWKFILNYFVPETNIKGECLKALILRNINIAENIGFTLRGVV
YDQGGNNRKCTSLFKVTK----------EKPYFYN---
NNKRYYMFYDIPHIFKSIRNNLLKA-NF----ETPDGLVDFDVIRDLYELEQ--GSVTRMT-
KLTKSHVNPTRFELMRVCLATRIFSHTVAAAIRTCNKNKQLQRSSSEVADATATFVEQVN-
DYFDCLNSRVLNDN--NPMKCALQKENVVWKKLKEMQVYL-----RNIRY--QGNLTY----
CIDGLLQITEAIFGLVD-
DLFKDHPDNFFFLTSRINQDPLENIFASVRAKGGNCRNPSVYEFNIIIAKLIS----------
----------LHIFHFTKKSNCESD--DDVMLHVQFDSIIY-----
EPYENNEIRXXEFSVSLSKIIQFNETYFVEHMDNMFTNDL----------PIELTSSRYFVSYIA-----
KGSNCEKCQTYLIKNSEFLTAPSEQFIFEKNYSKDTDFGNLKAPSDLFFNTNKMQLNIKQKR
FSTSRKLPTRVKC------IIEQC-IKCTKE-SVYSSWFDENDSC--
FTHKISLLDKLIKVLLFKHCKWTVMTDSYKKKA---------
KLNILSHK----------
>Dere\1360
MADKKRS----------NISGNKCSLPHCQKSVRDFPT----LKLFSFPEKDPTILMKWSEK--X--
QFTEDFVASTASRFLCQDHFSPD----------DIGV-
KYLRKGTIPDRNQIKYESCKNDDIDF----------
NAVRSPPQKRFR----------
SQSPGEVSRQCHKNTKTLKMLXKKYLAYKKLSDERQIRIRMLR----------
RENSNLKRKVIRLESKSEK----------
NLLSQIEKINLTGNEKILAKTLLKEKKGTNTRXAYSE------
KLFAXSIYYSSTSTYTFLR----------
DSHKLNFPSPSSLQKWNSITKLQPGDNECLYSALNETIKEMNESESECILTCDEVAIKKNLTY
NVSVDLIDGLEHL-----
IDRSNKMGSHICVSVIGSILKKWKFILNYFVPETNIKGDCLKKXIYKKISIVESIGFKVRGLVY
DQGGNNRKCTSLFEVTN----------EKPYFTF---
NNKKYYMFYDNPHLFKPIRNNFLKA-NF----ETPDGLVDFDVIREVYELDQ--KSVTRMT-
KLTRSHVNPTRFELMRVCLATQTLSHTVAAAIN-
CNQNKQFHRNSPEVAASTAAFVQKVYFDYFDRLNSRVLTDK--
NPMKCALQLNNTVWNKLKEMXXYL-----RSVKY--HGNSIY----CLDGLIQTTEAIFGLVK-
DIFRDHTDHFFFLTRGVNQDPLENRFACVRAKCGNCRNPSFNEFNIIIAKLIS----------
----------LHIFKFSQKCNCESD--DDVMLPIEFDSIIY---QPCVEKKE--
IQQDYSVSFSEIVQGNEKYFDQNMDDFXCNDV----------PIELTLSRYFVGYIA-----
KRSCCDKCRSVILKGSEHLTAPSELFIHEKNYSIESDFGKLKAPSDFFFNICKIHIKVIKNIFXN
-NKKQRCIKQF-----IVEQC-IKCTNESSDFPLWFHSNNDC--
YVHKIDILNXYTXKXQA----------KLCILSHK----------
>Dyak\1360
MAEKKRK----------NISGNKCSLPHCQKSRRDFPT----LKLFSFPEKDPTILMKWAEK--C--
QFTEDFVASTSSLFLCHDHFSPD----------DIGV-
RYLRKGTIPDRNLMTYNNNDEINF----------
NAVRSPAQKRYR----------

SQSPVEICRQCHKNTKTLKVFQKKYISYKKMSDERQIRIKMLR----------
KENSNLKRKIIRLESKSEK----------
NLLSQIDKINLTGNEKILVKMLLKEKKGTNTRWNDSE------
KLFAQSIYYRSTSTYTFLR----------
DSLKLNFPSPSSLQKWNSIKKLQPGDNECLYSALRETIKEMNESDKECILTCDEVAIKKNLTY
NVSVDIIDGLEHL-----
IDRSNKMGSHICVFVIRGILKKWKFILNYFVPETNIKGDCLKKLIYKNINIAENIGFKVRGVV
YDQGGNNRKCTSLLEVTN----------EKPYFTL---
NSKKYYMFYDIPHLFKSIRNNFLKA-NF----ETPDGRVDFDVIREVYELDQ--GSVTRMT-
KLTRSHVNPTRFELMRVCLATQTLSHTVAAAIKACNQNKQFNRNSPEVAASTAAFVQKVN-
DYFDCLNSRVLTDK--NPMKCALQVNNTVWHKLKEMQEYL-----RSVKY--HGNNIY----
CLDGLIQTTEAIFGLVK-
DIFKDHTDHFFFLTSRVNQDPLENIFACVRAKGGNCRNPSVNEFNVIIAKLIS----------
----------LHIFKFSQNSNCESD--DDVMLPIEFDSIIY---QPCIEKKEI-
QQQEYSVSFSEIVEGNERYFDQNIDNFLCNDI----------PIELTSSRYFVGYIA-----
KGSSCDKCRSVILKETEHLTAPSELFIHEKNYSTESDFGKLKAPSDLFFNICKIHIKVFENIFK
N-NKKQMCIQKF-----IVDQC-IKCTNESSDFSLWFHVENEC--
YEHKIDLLKKLIKVLLFKHCKWTVIANRQKSQA---------
KLSILSHK----------
>Dmel\1360
MAKKKRR----------YISGKKCSLSHCRKSRRDFPA----LKLFRFPERDPTMLIRWAEK--C--
QFTEDFVASTSSLFLCQDHFSPY----------DIGV-
KYLKKGAIPDRNLINYKSCNNADINL----------
NAVGSPPRKRHR----------
SQSPVEVCRQCHKNAKTLKVFQKKCIAFKKLSDERQIKIKMLR----------
KENSNLKRKLVRLESKTEK----------
NIYSEIDKINLTGNEKILTKMLLKEKRGTNTRWNDCE------
KLFAQSIYYRSTSTYTFLR----------
DSLKLNFPSPSSLQKWNSIKKLQPGDNECLYSALKEAIKEMNASDKECILACDEVAIKKNLT
YNVSVDIIDGIEHL-----
LDRSNKIGSHICVFVLRGILKKWKFILNYFVAETNIKGDCLKSLIYKNIIIAETIGFKVRGVVY
DQGGNNRKCTSLLEVTN----------EKPYFTL----
NNKKYMFYDIPHLFKSVRNNFLRA-NF----ETPDGLVDFDVIREVYELDH--GSVTRMT-
KLTRSHVNPTRFELMRVCLATQTLSHTVAAAIKTCNQNKQLHRNSSEVAASTAAFVQKDN-
DYFDCLNSRVLTDK--NPMKCALQVNNGVWNKLKEMQEYL-----KSVKY--HGNKIY----
CVDGLIQTTEAIFGLVE-
DLFKDHTDHFFFLTSRVNQDPLENIFACVRAKGGNCRNPSVNEFNIIIAKLIS----------
----------LHIFKFSQKSNCESD--DDVMLPIEFDSIIY---QPFVEKKEIQQ-
QEYSVSFSKIVQDNERYFDQNIDNFLCNDV----------PIELTSSRYFVGYIA-----
KGSSCDKCRSVILKETEHLTAPSELFIHEKNYSIESDFGKLRAPSDLFFNIYKIHIKAFENIFKN
-NKKQMCIKKF-----IVEQC-IKCTNESSAFPLWFYENNEC--
YAHRTDLLNKLIKVLLFKHCKWTVIADRQKKQA---------
KLSILSHE----------
>Dsim\1360
MAKKKRR----------FISGNKCSLSHCRKSRRDFPA----LKLFRFPERDPTMLIRWAEK--C--
QFTDNFMASTSSLLLCQDHFSPD----------DIGV-
KYLKKGTIPDRNLIKYKSCNNDYINL----------
NAVGSPPRKRHR----------
SQSPVEVCRQCHKNAKTLKVFQKKCIAFKKLSDERQTLIKNLR----------
KENSNLKRKLVRLESKSKK----------
NIYSEIEKINLTGNEKILAKMLIKEKKGTNTRWNDCE------
KLFAQSIYYRSTSTYTFLR----------

DSLKLNFPSPSSLQKWNSIKKLQPGDNECLYSALKESIKEMNASDKECILACDEVAIKKXLA
YNVSVDXIYXXEHL-----
LDRSNKIGSHICVFVVRGILKKWKFIINYFVAETNIKGDCLKSLIYKNIIIAEKIGFKVRGVVY
NQGGDNRKCTSLLEVTN----------EKPYFTL----
NNKKYMFYDIPHLFKSIRNNFLKA-NF----ETADGLVDFDVIREVYELDH--GSVTRMT-
KLTRSHVNPTRFELMRVCLATQTLSHTVAAXXXTCNQNKQLHRNSSEVAASTAAFVQKVN-
DYFDCLNSRVLTDI--NPMKCALQVNNAVWNKLKEMQEYI-----KSVKY--HGNKIY----
CVAGLIQTTEAIFGLVE-
DLFRDHADHFFFLTSRVNQDPLENIFACVRAKGGNXRNPSVNEFNIIIAKLIS----------
----------LHIFKFSQKSNCESD--DDVILPIEFDSIIY---QPFIEKKEI-
QQQEYSVSFSKILQDNERYFDQNIDNFLCNDV----------PIELTSSRYFVGYVA-----
KGSSCDKCRSVILKETEHLTAPSELFIHEKNYSIDSDFGKLKAPSDLFFYICKIHIKVFENIFKN
-NKKQMCIKKF-----IVEQC-IKCTNESSAFSLWFYENNEC--
YAHRTDLLNKLIKVLPFKHCKXTVMIADKQKKQA--------
KLSILSHK---------------------------------
>Dsec\1360
MAKKKRR----------FISGNKCSLSHCRKSRRDFPA----LKLFRFPERDPTMLIRWAEK--C--
QFTEDFMASTSSLFLCQDHFSPD----------DIGV-
KYLKQGTIPDRNLIKYKSCNDYNNL----------
NAVGSHPRKRHS----------
SQSPVEVCRQCHKNAKTLKVFQKKYFASKKLSDERQTLIKNLR----------
KENSNLKRKLVRLESKSKK----------
NIYSEIEKINLTGNEKILAKMLIKEKKGTNTRWNDCE------
KLFAQSIYYRSTSTYTFLR----------
DSLKLNFPSPSSLQKWNGIKKLQPGDNEWLYSALKESIKEMNASDKECILACNEVAIKKTXA
YNVSVDIIDGIEHL-----
LDRSNKIGSHICVFVVRGILKKWKFIINYFVAETNIKGDCLKSLIYKNIIIAEKIGFKVRGVVY
DQGGNNRKCTSLLEVTN----------EKPYFTL---
NNKKYYMFYDIPHLFKSIRNNFLKA-NF----ETPDGLVDFDVIREVYELDH--GSVTRMT-
KLTRSHVNPTRFELMRVCLATQTLSHTVAAAIKTSNQNKQLHRNSSEVAASTAAFVQKVN-
DYFDCLNSRVLTDK--NPMKCALQVNNAVWNKLKEMQEYL-----KSVKY--HGNKIY----
CVDGLIQTTEAIFGLVE-
DLFKDHADHFFFLTSRVNQDPLENIFACVRAKGGNCRNPSVNEFNIIIAKLIS----------
----------LHIFKFSQKSNCESD--DDVMLPIEFASIIY---QPFIEKKEIQQ-
QEYSVSFSKILQDNERYFDQNIDNFLCNDV----------PIELTSSRYFVGYIA-----
KGSSCDKCRSVILKETEHLTTPSELFIHEKNYSIDSDFGKLKAPSDLFFNICKIHIKVFENIFKN
-NKKQMCIKKF-----IVEQC-IKCTNESSAFSLWFYENNEC--
YAHRTDLLNKLIKVLLFKHCQWTVMIADKQKK-QA-------
KLSILSHK---------------------------------
>Dper\1360
MSGIRKGD----------AKSGRQCIVKTCQKSRRTNPG----IKMFNFPPADSIFGQIWREK--C--
CVALENYEQKL--IICDDHFSSA----------YIGK-
KKLKSEAIPTLNLEIELE----------
NNSVEFIEPEID----------
AEKEVSICEKCKKNSKSNNFYKKKCNRLLAEIKKLKKELKSKK----------
YIPKNKKYVVSQKIK----------DLIDNLE-I--SKESKIFCKLFVGTRPQK---YGEDV------
QFLAQNIYYVSPSTYAFMR----------
NRLNLSLPHVSTLYRWDPIKSLQPGFENTAIDA----------
EMAIRRELRYNEKLDIIDGFEHNG----
FERTSRIAKPVCVFMFKSIFSKTSSLLNYFASENGLTSDHLCEIVKRNISILHSLGVSVCVLVX
DQGSTNRKCFNNLGATI----------ENPFFEY---
ENQKVFCMYDFPHLIKSLKNGLLTC-DL----SSPDSIVSFKVVQELWEMEE--

HAGTKMCPKLSRHHIYPNSFEKMRVKFATQIFSRTVQAAIKTVCETVGFKNSTYQVALSTAE
FINKID-QIFDCMNSGSLYAD--NVYRSAIQLNNVPHKFIQFFLSYI-----KDVNFVDSKKRVY----
FLDGIQITLKSLLLLAD-EL-
LTNSDKIFIMTKSLNQDKLENTFAVVRQKGGNNTNPSVAEINNIFARILN----------
---------IKIVCSSDFGNCEADFEEGAAVQACIEGVFNE--
SNNLKVEHPNDHDELLSKLDFDGSGFENYFEKDSFKTS-KEI----------
NIEVASMRYFVGYVAFKTI-
PRLNFETCSKCMRKEDEVITVPSELFIFYKNYQKFTDFGSLIAPSDCLMEISKKHILIFCKFFEI
-GPQKVGIKNL-----ILKAC-------QDETPLWF--TGEC--
REHKLALLDFVILVLLRKHSLWAIRRGKKNASK---------KLKIMAQ----------
>Dpse\1360
MSGIRKGD----------AKSGRQCIVKTCQKSRRTNPG----IKMFNFPPADSIFGQIWREK--C--
CVALENYEQKL--IICDDHFSSA----------YIGK-
KKLKSEAIPTLNLEIELE----------
NNSVEFIEPEID----------
AEKEVSICEKCKKNSKSNNFYKKKCNRLLAEIKKLKKELKSKK----------
YIPKNKKYVVSQKIK----------DLIDNLE-I--SKESKIFCKLFVGTRPQK---YGEDV------
QFLAQNIYYVSPSTYAFMR----------
NRLNLSLPHVSTLYRWDPIKSLQPGFENTAIDA----------
EMAIRRELRNNEKLDIIDGFEHNG----
FERTSRIAKPVCVFMFKSIFSKTSSLLNYFASENGLTSDHLCEIVKRNISILHSLGVSVKVLVX
DQGSTNRKCFNNLGATI----------ENPFFXY---
ENQKVFCMYDFPHLIKSLKNGLLTC-DL----SSPDSIVSFKVVQELWEMEE--
HAGTKMCPKLSRHHIYPNSFEKMRVRFATQIFSRTVQAAITTVCETVGFKNSTYQVSLSTAE
FINKVD-QIFDCMNSGSLYAD--NVYRSAIQLNNVPHKFIQFFLSYI-----
KDVNFVDSKKRVY----FLDGIQITLKSLLLLAD-EL-
LTNSDKIFIMTKTLNQDKLENTFAVVRQKGGNNTNPSVAELKNIFAKILNKL----------
----------KIVRSSDFGNCEADFEEGIAVQACIEGVFNE--
SNNLKDEHPNHHDELLSKLDFDGSGFENYFEKDSFKTS-KEI----------
NIEVASMRYFXXXXAFKTI-
PRLNFETCSKCMRKEDKVITVPSELFIFYKNYQKFTDFGSLIAPSDCLMEISKKHILIFCKFFEI
-GPQKVGIKNL-----ILKAC-------QDETPLWF--TGEC--
REHKLALLDFVILVLLRKHSLWAIRRGKKNASK---------KLKIMAQ----------
>Dpse\Galileo
MSSANQKDN----------RKTGIKCIISSCLSGYRNSSS-DFPVRLFQFPQNLLMFNK-WVEV--
C--GLDSNLLKKTA--RICDRHFDVK----------YLGV-
RKLKANAVPTLNLSENTLLTAFNADLCNDYEFSE----------
NKQDNEPIGTFLQAKRIKLADVHNIMSNELEQQCVTEAPDKVSLENK----------
-ESIDIKFCENCLKREQNEKYYRNKYYELMADQKKRENGFIKLKDRFLALKRV--
VYRRNRRYRIKTTNLKPNII----------PIIAKMTHV--SAEAKTICIMILKKTLS----YSPAE------
KVIAQNINFYSSRTYEYLR----------
DVLNLKLPSKRTLARWAVLKNMRPVFNPDLLSNLKTIFDGMSSKGKEAVILFDEIKIKRGLH
YNIALDEIQGYENDG----
QNTKSLLGQQVCVFMIRGLFENWKYVLSYTVTANGIKHEALLTKVTANIEQAQVLGLNVR
AAICDQGSNNRAVFRRLGVDI----------KNPSFKV---
QDKEIFAIYNVPHLIKSLRNIVRNI-NL----YTPDGVVSWKIVEELYEIDS--
RNSTRLCPKLTTKHIYPNSFEKMKVKYATQVFSHSVAAALRTMISSGGFLKCK-
ENAEATATFIEKMN-RLFDCLNSHVLYDK--NPFRSALQDKNLVNETLSDMRKYF-----
EEFKY---PQEVY----CIKGMILTINSILSLAQ-
NVWSESAEVFYIATSKLNQDPLENLFYLIRSRGVTNNNPTMYEFNVIISKMLS-
M----------KVLTSTTVSGNCSPD-EDTMLINIIKDNVSS---ESASESKT---
FDDDILISTNEDAELELSIEATDLSIQ-AAF----------NEN--ALRYYAGYLLHKLL-

KKYDCNKCSELLKSSDE-
VRCSSEYLILNKNFGYVSSSLKLKAPSEDFCTLVKIYFDIFNRHFET-KSHKFNLKRS-----
IVQKC-IILTSKIDKYADWFKFSDPC--
FVHRNHLLNQFVLILIRKNFKWLTDRMIGNQNAKSY--EK--
KLKILRQ----------
>Dper\Galileo
MSFANQKDN----------RKTGRKCIISSCLSGYRNSSS-DFPVRLFKFPQNLLMFNK-WVEV--
C--GLDSNLLKKTA--RICDRHFDVK----------YLGV-
CKLKANAVPTLNLSENTLLTAFNADLCNDYEFSE----------
NKQDNEPIGTFLQAKRIKLADVHNIMSNELEQHCVTEEPDKVSLENK----------
-ESINIKFCENCLKREQNEKYYRNKYYELMADQKKRENGFKKLKDRFLALKMV--
VYRRNRRYRIKTTNLKPNII----------PIIAKMTHV--SAEAKMICIMILKKTLS----YSPAE------
KVIAKNINFYSSRTYEYLR----------
DVLNLKLPSKRTLARWAVLKNMRPGFNPDLLSNLKTIFDGMSSKGKEAVILFDEIKIKIAL----
--DEIQGYKNDG----
QNTKSLLGQQVCVFMIRGLFENWKYILSYTVTLNGIKHEALLAKVTANIERAQVLGLNVRA
AICDQGSNNRAVFRRLGVDI----------KNPSFKV---
QDKEIFAIYDVPHLIKSLRNIVRNR-NL----YTPDGVVSWKIVEELYEIDS--
RNSTRLCPKLTTKHIYPNSFEKMKVKYATQVFSHSVAAALRTMISSGGFLKCK-
ENAEATATFIEKIN-RLFDCLNNHVLYDK--NPFRSALQDKNLVNETLSDMRKYF-----
EEFKY---PQEVY----CIKGMILTINSILSLAQ-
NVWSESAEVFYIATSKLNQDPLENLFYLIRSRGVTNNNPTMYEFNVIISKMLS-
M----------KVLTSTTVSGNCSPD-EDTMLISIIKDNVSS---ESASESKT---
FDDDILISTNENAELELSIEATDLSIQ-AAF----------NEN--ALRYYAGYLLHKLL-
NKYDCNKCSELLKSSDE-
VRCSSEYLILNKNFGYVSSSLKLKAPSEDFCTLVKIYFDIFNRHFET-KSHKFNLKRS-----
IVQKC-IILTSKIDKYADWFKFSDPC--
FVHRNHLLNQFVLILIRKNCKWLTDRMIGNQNAKSY--EK--
KLKILRQ----------
>Drho\P-1
MNQHSV----------RKNGRRCIIECCRGSYRDSSD-SFPVKLFSFPSDDRRLKK-WVEL--C--
NLDDNFNRRTS--RICNRHFESK----------YIGI-
NRLKSNAVPTLNLTDNFLLRPDSSVPCEQFEFN----------
DDGAEEPIGIYVNPIRSQNLSDSLDVSPLNFEAPKSDANLVTPSAVQLYVSPSDLVTP----------
-------SVPQLGNCCYCQKREENEMFYRNKNYEMFLAIKTYKQKIEKLRRANVLM-----
KNAMKRKSKIISKKKSINIF----------NIINKLPHV--SEEAKTICKMLIKKSNS----YNSAE------
KVIAQNINFYSSRTYEYMR----------
DVLNLKLPSKRTLSRWALFKNMTPGFQPDLLENLQKIVGEMSEKGKEAVILCDEIKIKKGLQ
YNTALDEIQGFENDG----
EKRTRFLGQQVCVFMARGLFENWKYVISYTVSANGIKHDALMSKVEANIGVSQTLGLNVR
AIICDQGSNNRAAFKKWGVNI----------NKPSFNV---
NGKEIFTIFDAPHLIKSLRNIIMKN-NL----ITPDGEVSWDILKKLYQLES--
RNSTRLCPKLTAKHINPNSFEKMKVKYATQIFSHTVAAAIRTVIDSGGFVECR-
NSAEATANFIEKIN-RLFDCLNSHVLYER--NPYRCGMQKNNNVQRYLVEMRKYF-----
GELKY---PQLVH----CIDGMILSISSVLALAE-SVW--
SSEIFYISTAKLNQDPLENLFYLIRARGATNNNPTMYEFNIIMSKMLS-
M----------KILTSTSVSGNCGPD-EDTMLINVIQDCSLN---KTCNDIKT--
EESTDFDMFSDEETEIEQIFDIATG----NEF----------DSN--ALRYFAGYILFKFL-
QKKECGVCSDLLKTNIE-
TQCSTEHFIINKNYDCADKGLKLKAPSDYYFSLIEIHYNIFKKIFDK-NPYKKRIKQK-----
IIRSC-ILATEKSSIYSNWFSVTHSC--
YEHRMKMLDGFILILLRKNSKWLTEKMYTKGKAVSTAGAR--

KLKILKQ----------
>Dana\Galileo
MNRQNV----------RKNGRRCIIDSCQGSYRNSSG--FPIRLFSFPSDDRTLKK-WVDL--C--
QLDDTFNRRTA--RICNRHFESK----------YIGK-
SRLRSNAVPTLNLFDNSLLCPNSPVPCEKFDFDI----------
NDEAEEPIGIYVNPIRTEQFSNSLDSPSNLETPCAP----------
QLEDSPFCSNCQKREENEMFYRNKNYEMFLELKKYKEKMQKMGRALVLM-----
KNASRRRGKRNSRKTTPNLF----------NIINKLPHV--SDEAKTLCKMLLKKSNS----
YNSAE------KVIAQNINFYSSRTYEYMR----------
DVLKLKLPAKSTLSRWALFKNLTPGFHPDFLENLQKIVGEMSEKGKEAVILCDEIKIKKGLQ
YNTALDEIQGFENDG----
EKRTRFLGQQVCVFMARGLFENWKYVISYTVSANGIKHDALMKKVEANIEVSQTLGLNVR
AIICDQGSNNRAAYKKWGVNI----------NKPSFNV---
NDKEIFVIFDAPHLIKSLRNLLLKN-NL----NTPDGEVSWDIIKKLYQIES--
RNSTRLCPKVTAKHINPNSFEKMKVKYATQIFSHTVAAAIRTVVDSGGFVDCR-
NSAEATANFIENVN-KLFDCLNSHVLYEK--NPDRCALQKNNNVHNYLVEMRKYF-----
AEFKY---PQVVH----CIDGMMLTISSVLALSE-RVW--
SSEIFYISTAKLNQDPLENLFYLIRARGATNNNPLMSEFNNIMSKMLS-
M----------KILTSKSVSGNFGPD-DDTMLINVIQDCSTN---
KICNNLKTDEEESTDFDMFSDEETEIEQIFDIATG----NEF----------GSN--
ALRYFAGYILFKFL-QKNDCGACADLLKKNID-
AQSCTETFIINKNYDCADKTLKLKAPSDSFFSLIEIHFNVFKKIFDK-KPYINRIKKT-----IIQSC-
ISSTEKSSIYSDWFSVTHPC--YEHRMKMLNGFILILLRKNSKWLTEKM-SKEKAVSTASSR--
KLKILKE----------
>Dbip\P-1
MNRPNV----------RKNGRKCIIASCQGSYRDSSH-SIPIRLFSFPSDDRNLKK-WVEL--C--
QLDDTFNRRTA--RICNRHFESK----------YIGK-
SRLRSNAVPTLNLCDNSLLRPNSTVPCEQFEFC----------
DDGAEEPIGTYINPIGRKEFLNRVGSPSNLETPNLES----------
SPQLEDFCSNCQKREENERFYRKKNYEMSLELKKYKEKMQKMRRAFVLM-----
KNSSRRKGMRNSRKSTLNLF----------NIINKLPHV--SDEAKTLCKMLLKKSNS----
YNSAE------KVIAQNINFYSSRTYEYMR----------
DVLKLKLPAKSTLCRWALFKNMTPGFQPDFLENLQKIVGEMSEKGKEAVILCDEIKIKKGLQ
YNTALDEIQGFENDG----
EKRTRFLGQQVCVFMARGLFENWKYVLSYTVSANGIKHDCLMKKVEANIEVSQTLGLNV
RAIICDQGSNNRAAFKKWGVSI----------NKPSFNV---
NGKEIFTIFDAPHLIKSLRNLLIKN-NL----NTPDGDVSWDIIRKLYQIES--
RNSTRLCPKVTAKHINPNSFEKMKVKYATQIFSHTVSAAIRTVIDSGGFVECR-
NSAEATANFIEKVN-KLFDCLNSHVLYEK--NPYRCALQNNNNVHNYLLEMRNYF-----
AEFKY---PQVVH----CIDGMMLTISSVLALSE-SVW--
SSEIFYLSTAKLNQDPLENLFYLIRARGVTNNNPSMSEFNTIMSKMLS-
M----------KILTSKSLSGNCGPD-EDTMLINVIQDCSLN---KTCKELKTD-
EDNTDFEMFSDDETEIEQVLDIATG----DEF----------DSN--ALRYFAGYILFKIL-
EKKDCGVCADLLKKKIE-
AQTCTENFILNKNYDCADKTLKLKAPSDYYFSLIEIHYSVFKKIFEK-KPQMNRIKKT-----
IIQSC-ISSTEKSSIYSNWFSVSHPC--
YEHRMKMLDGFVLILLRKNSKWLTEKMYFKGKAASAAGSK--
KLKILKE----------
>Dbuz\Galileo
MAQISVVNEKSVGVKKXLEKCLGXSGAKCVIETCGASYRQSSV-NFPVKLFSFPSKEIEFNK-
WVKT--C--HLPSNFNRKKA--KICDRHFERK----------YIGK-RKLKANAVPTLNLCDPNFFS-
NSADFNDDFRLVD----------NRGAEHQNENVPNECDDEL-

IENLLIFDDNSKKEFWQNL----------
AVDRTPYCLNCIKREQNEVYYRKKYYEIGLDLKKVQERYTKLKRRFISFKRV--
SNYRGVSFRVRRAKKTVNVF----------TTINSLAHV--SEQSKVLCKMLLKNTNF----
YNSAE------RVLSQNINFYSARAYEYLR----------
DVLHLKLPSKKSLNRWAIFKNLTPGSNPELLENLQGIVEKMSDKGKYAVLVFDEVKIKKGL
QYNSYLDEIQGFENDG----
EKRTKFLGQQVCVFLIRGLFENWKYVLSYTVSANGIRHSDLKSKVEANIGLSQALGLNVKA
VVCDQGSNNRAVFDRWGIDI----------NKPSFHV---
NDKEIFAVFDAPHLVKSLRNILLRH-NI----STTQGTVSXNIIRKLYEIES--
KNLTRLCPKLTSKHVSPNCFEKMKVKYATQVFSHSVAAAIRTVIDSGGFSDCK-
DSAVATAIFIEKIN-RLFDCLNSHVLFDS--NPYRCALTRNNNVHEYLQEMRDYF-----
HDLQY---PQKVY----CITGMIITISSVIALAE-
NIWNDNNDLFFVATSKLNQDPLENLFYLIRSRGATNTNPTIFEFNSIISKMLS-
M----------KVLTSASISGNCILD-EDSMLANIIKDSGST---LSVFHSQCE-
IHSSVYEEPSDPDFEIELSLDSTIVNIQ-NAF----------NEN--ALRYFAGYLLHKLL-
QRTDCEVCTNLLKGSDE-
MQCSSEYLILNKNYNYIHQYLKLKAPSDNFYNIIKIHFDIFQKIFDK-KPFIAKLKEK-----
IILHC-MRATAKSTLHSDWFSPSHPC--FDHRKFMLNQFVLILIRKNCKWLTDSIVSKSSNLS---
KS--KLKMIRE----------
>Dmoj\Galileo
MAETSVIKKSTDGVKKCS----RRNGGKCIIETCGASYRKSSI-HFPVRLFAFPSKEFHLNK-
WIKA--C--NLPNNFDRKRA--RICNRHFERK----------YIGK-
RYLRVNAVPTLHLGNSNLLSNNNADVSDDIYSIDIQEEDITPYSYPKKSLDKVVDDFVKPSD
TEQQHQISNIQNQSDEEENLENFLSFDNSLKNQLWQDL----------
SAGRSSFCSNCLKREQNEVYYRKKYYDMGVDLKKVQKEYLNLKKKYSALKNA--
SRHRNIYHRIRKVKKHVNVF----------TTINNLPHV--SEQSKVLCKMLLKNNHV----
YNSAE------RVIAQNINFYSARTYDYLR----------
DVLNLRLPCKKSLNRWAILKNLVPGFNPELLENLQGIVEKMSAKEKYAVLVCDELKVKRGL
QYNSSLDEIQGFENDG----
VKRSKFLGQQVCVFLVRGLFDNWKYVLSYTVSARGINHTDLKKKFEENIGLSQALGLNVK
AVVCDQGSNNRAVFNRWGIDL----------NNHSFEV---
NGEKIFAIFDAPHLVKSLRNILLKN-NI----STPEGTVSWGIIRKLYETET--
KNLTRLCPKLTLKHVSPNCFEKMKVKFATQIFSHSVAAAIRTVVETGGFADCK-
DSAVATAIFIDKIN-NLFDCLNSHVLFDS--NPYRCALREKNNVHEYLQEMRDYF-----
QNLHY---PHKVY----CIDGMMITISSVIALAE-
NIWNDNNDIFFVATSKLNQDPLENLFYLIRSRGATNTNPTVFEFNTIISKMLS-
M----------KMLTSASVSGNCIPD-EDLMLANIIKDSGSQ---
LSVFHEQCNSCHTPTDIEPLDADLEIELSLDATIANIQ-NDF----------NEN--
ALRYFAGYLLHKLL-QKTDCEVCTNLLKSSDE-
MQCSSEYLILNKNFHYINRYLKLKAPSDHFYNLIKLHFETFRKIFEK-KPYIARIKEK-----
IVLYC-MHATAKSSLDNEWFSPTHPC--
FEHRKFILNQFVLILIRKNCKWQTEKIVGKTSISKR------KLKIIHQ----------
>Dvir\Galileo
MKDK----------RRSGLKCIIKSCLRSYREDGD----VRLFKFPKNDCLRRQ-WVSK--C--
ELPGNTDIDKA--LICDEHFEQN----------FLGK-
TRFKKNAVLSLRLFASLNLNFNITRARGRIDAFT----------
FFKQDDNIETFSAFKNNNITIETTTRIDCEHVIELSDSGSLVKISQYNNPEKTKCHLSNHVGFP
ELYDNISLSAEKDSQDIRFCPNCLKKEQNELYYQNKYWKFFLKCTREYKQVQHYKRKLHR
MQILLRSERIKEASYFKKSRKNINIF----------AILDSKPFL--SXNSNTVCKMLLKNKNS----
WEEEE------KVVAQSIHFYFAKAYDFMR----------
NDLHLNLPSKSSLARWAPVKYLVSGLNECLSNTLLKIFSKMNEKSKQAVLLFDEMSIKRGL
QYNSRRDEIEGFTDDG----

VEKTPELCKQISVFMVRGLYENWIFVLSYFATSTGLLTLKLKRQIESFLRTGYSLGLNIKAIV
CDQGSINRGAFTKYGVNK----------EVPYFTI---
DDKKIYGIYDDPHLFKSLRNILMRN-SL----
ETPDVRVSWQILVKLFQIDTDITSTLFLCPKLSRKHIYPNYFKNMKVKYATQILSHAVASATK
TLIQNGNFADCR-DIALSTAKFIERVN-KLLDCLKSNVLKDK--
NLFESALQNNNIKEKYITEMPNYF-----MKCRY---LKTVY----CINGLILTINSVLKLSQ-
DIWREDSNVFFLILSRLNQDALENLFYLLRDRGITYSNPKLFEFNAIISKMLS-
M----------KIFTAKISSGNCQPN-
GEFMLVNVIELANEKCKAFVLRTKNICPITSSALNISTNV---VCDNDDLPSVAI--SAS----------
SDN--ALRYFAGFVLDKSQ-QEFNCDTCKSFLKEENAKCED-
SEYFLCNKNFKSINNRLKLKDPQDDFFCLIKHCYSIFQTIFQK-SQHVRQKKKRRE--LTIYEC-
ISRNKAFEKFANWFSESHSC--SHHRKYILNYVLQVLFRDNSIWLIEKLCGLSENPSS--AR--
KEKIVQS----------
>Dwil\Galileo
MAKHCRSKSFAGEVTHG-----RKSGCKCIIASCMRSYRDGSE-HFPVRMFKFPKNEFVRRQ-
WVSK--C--NLQYDVNIDRA--LICNLHFEKK----------FLGT-
KFLKAGAIPTLLLTDEPNLNLIATDAKIDLYDF----------
CEETHEEISTFQNIRKHGAEPTNILENIDNNTEKVEDISDI----------
STDCMQFCSNCLKKEQNEAYYRKKCFEMSENLKKEIQKVKLCN----------
KKIRHLRIVLRNERARKLKYLKEKKKI----DIQGLIDKKCVSKNSNTVCKMLLKNKQS----
WEDEE------KIIAQSINFYSSKAYNFMR----------
DDLELNLPCNKSLQRWAPVRNMVPGLNENLLKHLKGIFLKMHNKSKNSVLVLNEISIRKGL
QYNSHRGEVEGFVDDG----
YEKTDALCKQICVFMVRGLYANWKFVLSYVATSTGLSSHKLTQLIDSNIRAARTLGLFIRAV
VCDQGPNNRGAFNKLGIVN----------EAPYFSL---
DDQKIYGIYDVPHLTKSIRNILMRD-SI----ETPDGTVSWHVVVRLLEIDTS-
NTSTRMCPQLTRKHIFQNSFEKMKVKYATQVFSQTVSSAIKTLIQHGKFIDCE-
DVAIATSKFIEKVN-RLFDCLNSSNIYDR--NPNKSAIQKDSDNEQYIIEMRDYF-----KKCLY---
RRKVY----CLDGIVLSINAILMLTS-
DIWNEGHGVFFLMLSRLNQDALEHVFYLIRSRGGTNNNPMLFEFNAIISKMLS-
M----------KLITSKTTTGNCEPD-ED-
MLINVIEETKHELAIENVNDQDQVYYEDFNIILDENMKDEVSEADKEQPTEI--SIA----------
TEN--SLKYFVGFVMHKAQ-QKFNCDTCKELLKEEIANYEE-
SEFFIINKNFKTINNNLKLKAPQNHLFNLMKQHYKFFKN-FPH-ARKIKEK--------IINEC-
LSNIEKDPNYLDWYSESHEC--SEHRKFILNYFLLVLLKKNSKWLMESLCGASEKHKS--NR--
KIEILQS----------
>Agam\P-3
MPRS----------CAAAFCKNNAENVKKRGLNITFHSFPSDDSLPK--WIDF--C--
KRDEHWKPTKIS-TVCSLHFKPDDYQMAKSSLPQTLPVLKRLKPYAIPSL-
IQPADFIQNEPSNMTAP----------
LKECNQPNVEF----------
QTDSEYFSDVENVPSQTIIDMKRELDQVKEDNRKLIEVNTNLR----
DKLHSYFNENKRLKAEIDNLQKHISKKDAGIDEAALVTAMKERLKPT--
LSENQIDIILKKKKRVV----WTKEE------IGSALTLKYFGLRCYKYLA----------
KDRKFPLPADATLKRYTKNLVVKEGILDDVLKLISNLTSTFTEKDRLCALSFDEMKVNRIIEL
DKASDEIIGPHNY----------LQVVMARGLCNKWKQPV-YIGFDKKMTKEILLKIIEK----
LSEININVVAIISDNCSTNVSCWKELGAKDY----------ERPYFQHPT-
TLNNVYVIPDAPHLLKLLRNWFLDS-GF----TYNGKHIKADLLFDMIASRN--
ETEITPLYKLSKTHLVMTPQERQNVRRAAQLLSHTTAISLRRYFKNN-------
AEATDLANFIEKVD-LWFSISNSYSPFAK--LDYKKSYTASDDQIKALDEMFEIV-----
SNMTVIGKHSLQI----FQKSLLMQITSLKLLYD-DL-
HKRHNISFISTHKLNQDVLENFFSQLRQIGGVYDHPSPMSCIHRIKMIILGK----------

---------APTFLKNQTDLEPSTFSCTDEYISSQIRTSIEIEN-
EGSEQANDGDIISASIITSALNQPPVKQSIQSDTLSSRSSEMLSSVSSSAIELPEQDSDGLEYIM
GYI------
GRQCFEKFPHLNLGNLSLNLNSDHSYSHPPSFVKHLSVAGLFVPSEAFLKQGYKMEKIFQK
LHPN-GNFNKKRY-------ISKRLVKRLQKEFPELPLIVVQQF----
AKHRINIRIKFLNMKIANEKRVNNKRKAPSHSKTAKKCE---
KLQISCFV-------------------------------
>Dbuz\P-element
MK-------------------------CSVLHCENYFRKNN-----ISFFKFPLDKKLRKK-WLQF--C-
GKWQSEYNCTNG--KICEEHFESE-
CLIGKIQCGLGLSRKVLLKPGAVPTIKSTHVHEQVRHRTQRSE-------------------------------
LRIRRQLVSEMLQNKENISELAQNNIAEGNTMIVE-------------------------------------
EEMELDINNN---
NDSNLINLKAQIKVLESQVRDLQCENLKLQTALRDKGYAYDKELRKKEFDFSELQKKFEKL
ELESK------NIESCLKVF--FTEGQLIKLKNKDRRQN----WNVDD------
IARSITFYSASPKGYRLLR--------------RNKFPFPAIRTLQHWAQRIDICPGILAPVIKILSAT-
THLSQTQKLCVLSFDEMKIRSTYTYDKPSDSTLPAVNY---------------
VQVAMLRGLVADWKQPI-FYDYDCPMTKTKIQEILKS----
TQNMGYTIVAMVCDLGGTNRSLLSSLEVTY----------------------RQPWFLF---
EGRKVYVFADVPHLIKLIRNHFIDS-GF----VINDKFVNAAIIAELLNITA---
GDLSITHKISHKNLTVSRAERQKVKMATKLFSNTVAAAIQRAASLGYLNG---
HNWSECYELFKTTN-DWFDVLNVRVPRADSRCRMHAYGLAFETQNNILDKMSSLI-----
LNMRVGRNTTLLP----FQKGILQTNNALRMLFH-DI---
KGNVAFLLTYRLNQDVLENFFGLIRARGGMHDHPDRQEFKYRLRSYLLG-------------------------
-----------RNVGVLSNGANVAAD-DTPDLESANPSMTGL------------------
ILSHFRCATQPMATDDEENL----PEL------------EYD--ALENLAGYV-------------
CHRLKMSGHNDENNIDSTF----TWVDQVSEGGLCKPTQEIMNSFKSLEIIFRNLN-G-
DSLLITE-------NYLKKH-IEEASS----------------VNINMKAKQLFFRSRMYFTIRKLNR---
LSLRYTSKFSNN--NYNKTVN--------------------------------
>Aaeg\P-1
METS---

---------------------------------------VKKENLKLKKLLNQEK-----------
RKTKIYSSQLKKSK-----------------ALNSRLAEI--FSKGQLRKLTDPLKRRIQ---WSPDD------
ISRAISLHAAGAKAYRLLL--------------
SRGYPLPAVSTLKKWAGKVKLTPGVLKPVLNLLQN--
SKFNEKERACVLSFDEMKIKSCYEYDRTSDKVLKPTKY----------------VQVAMIRGLYKN?
KQVI-YYNFDTAMTATIIKDIIEK----L?
NIKFRVVAIVCDMGPTNRKLWKDFEISC----------------------EKPFFTV---
DDQKVFTFADTPHLMKLVRNHFLDS-GYSWQKEETTHLIT?RPVVDLVNLQ---
RSELKICHKITLTHLDVKDAARQKVKYAVQLLSNSAAQGIKRAFSLGLISS---P?
ALITSSFLKNMN-DWFDIFNSSAASKGLRERLKAFDHSNTVHQSIIHESNEMI-----
SQLRVPNRRTLLP----CQKAILVNNAALIGLSV-FL-KTIYG?
SYLLTRRLQQDDLERFFGTIRSKGGLHDHPTALEFTYRLRNSILGN------
IIDYFILIKFVLLCVKLWLSIPRYVMIKIYFFQIGRSDASQQ-
VNNSNVEQEQEEPYEELQFTGSLLKDISPGQTDFFQAEDQDYEMEENMDESECSMEYSEL--
----------EED--ALEYIAGYIIKKLKLRIPPPDKSAF--------------------
TWVDQLSEGGLTKPTFDFVTQIMLLDKIFKQQHG--ETFNFNVK-------AVESC-
IDASEN----------------IGLSVEIKKLFFRTRLYIRIRNLNKKNDNIQKTKKR------
KMKKIVN---------------------------------
>Dboc\KBOC
MSF-------------------------CEFCCAVVKT-------EGVKFIRVPKEDRKRKL-WEESLGC--

SLAHNA-------RICDTHFKGS-DFYGETKTKEERKR-
RRLMPNALPRQPTPEPESI----------PTVKPGYSNAYTQT----------IDLENFKLKQKISELE----------
KEIHHLRQQLSESD----------ALRQGLTKI--FTQNQIKMLPNCGKRIR----YNSSD------
MSEAICLHAAGPRAYNHLY----------
RKGYPLPSRATLYRWLSEVEIKTGTLDIVMDLMKN--
EDMDEADKVCVLAFDEMKVSAAYEYDSAADAVYKPASY----------
VQLAMVRGLKKSWKQPV-FFNYNTAMDACTLKAITTK----
LYKSGYIVVAIVCDLGPGNQKLWREFGISE----------
ENTWFSHPVDPALKIFAFSDVPHLIKLVRNHYVGS-GL----LISGTKLTKNTVQQAMNCCS--
SSDLSVLFKLTENHINVRSLQKQKVKMATQLFSNTTASAIRRCYELGYEI----
ENACETADFFKMIN-DWFDTFNSKLSTANSLKYSQPYGLQHDLQKDILDKTSLTM-----
SGKIIEKSQRRLP----FQHGIIVSNKSLDGLYI-YL-
KEKYNMEYILTSRLNQDILEQFFGAMRSKGGLYDHPTPLQLKYRLRKYITA----------
----------KNTELLTGKGNVEDG-EEEEWLNLGDI----------
DEM----------TED--AIEYVAGYMIKKLKLRDMSNKDATY----------
TYVDEVSHGGLKKPSSQFVEQLKKLEAIFQ-LYAK-EEFDLQI-------NVKRTL-
LNAAEK----------LNVPLDIKQLFFKCRIYFRIKHLNKKLAIKNQKQRIVANS--
KLLKIKL----------
>Dmad\P-element
MKW----------CSLCRKVAND---------VKLVHVPKGLEKRKL-WEQCLDC--
SLTVNS-------KICDSHFDAS-QWKSSTIKGQIFKK-
RRLXADAVPQGDPAA----------KLVKLGFANSSTQTEDAFINH-----
SARVEHESLTGRIRLMQ----------KKMDSLREQLEVYK----------DLEISLRTI--
FTETQIKILKNKGKRAV----FNATD------MSAAICLHTAGPPAYYHLY----------
RKGFPLPSRATLXRWLXDVNISTGTLDVVIDLMEN--
EEIPEVDKLCVLSFDEMKVAAAFEHDSSADVDYEPSTY----------
VQLAIARGLNKSWEQAV-XFDFSTLMDADTLHSIINK----
LHKRGYPVVAIVSDFAAGNQTLWTELGISE----------
RKNWFTHPADEDLKIFVFSDTPHLIKLVRDQYVES-GL----IINGKRLTKSTXQXTISHCA--
KXDVSMSFNITDNHLNIXPLAKQNIKLATQLFSNTTGSFIRRCNALGYNV----
QNASETADLFKIIN-DWFAVFNSKSSTSNSIEPTQPYGKQIEIQRGILAKMSEIM-----
SSEILGVGAHRLP----FQKGILVNNASLEGLYC-YL-
SEKYEMQYIFTXRLSQDIVENFFIAMRPKGEQFEHPTPLQFKFMLRKYISAL----------
----
TKLNKPIDK----------
>Dsub\P-element
MKW----------CSLCRKVEND---------VKLVHVPKCLEKRKL-WEQSLDC--
SLTVNS-------KICDSHFDAS-QWKSSTIKGQIFKK-
RRLNADAVPQKPQQ----------EIVRLGFANSSTQTEDKVINH-----
ALRVENESLRKQNRRMQ----------TEMHSLRQQLEDYK----------ELETSLRTI--
FTETQIHILKSGGKRAV----FNATD------ISAAICLHTAGPPAYNHLY----------
RKGFPLPSRATLFRWLADVNISTGTLDVVIDLMEN--
EEMPEVDKLCVLSFDEMKVAAAFEHDGSADVDYEPSTY----------
VQLAIARGLNKSWEQPV-FFDFSTLMDADTLHSIINK----
LHKRGYPVVAIVSDLGAGNQTLWTELGISE----------
RKNWFTHPADEDLKIFVFSDTPHLIKLVRDQYVES-GL----IINGKRLTKSTVQQTISHCA--
KPDVSMSFNITDNHLNIGPLAKQNIKLATQLFSNTTGSFIRRCNALGYNV----

QNASETADLFKIIN-DWFAVFNSKSSTSNSIEPTQPYGKQIEIQRGILAKMSEIM-----
SSEILGVGAHSLP----FQKGILVNNASLEGLYC-YL-
SEKYEMQYIFTSRLSQDIVENFFIAMRPKGEQFEHPTPLQFKFMLRKYISGM----------
---
TKLNKPIDK----------
>Dgua\P-element
MTW----------CSVCGKVANH---------VKLVHVPVCLEKRKL-WEQILDC--
SFAVNS-------KICDSHFDAS-QWRSPPKEGQIYKR-
RRLKADAVPHGEPEP----------
----------KFVKLGFANSSTQTEDNVINH-----
AIRVENESLRKQNRRMQ----------KEMHSLRQQLEDFK----------ELEISLKTI--
FTETQINILKSGGKRAV----FNATD------MSAAICLHTAGPPAYNHLY----------
RKGFPLPSRATLYRWLADVNISTGTLDVVIDLMEN--
EEMPEVDKLCVLSFDEMKVAAAFEHDSSADVDYEPSTY----------
VQLAIARGLNKSWEQPV-FFDFSTLMDADTLHSIINK----
LHKRGYPVVAIVSDLGAGNQTLWTELGISE----------
TRNWFTHPADEDLKIFVFSDTPTLIKLVRDQYVDS-GL----IINGKRLTKSTVQQTISHCA--
KPDVSMSFNITDNHLNIGPLAKQNIKLATQLFSNTTGSFIRRCNALGYNV----
QNASETADLFKIIN-DWFGVFNSKSSTSNSIEPTQPYGKQIEIQRGILAKMSEIM-----
SSEILGVGAHSLP----FQKGILVNNASLEGLYC-
YLSSEKYEIEYIFTSRLSQDIVENFFMPMRPKGEQFEHPTPLQFKFMLRKYISGM----------
---
TKLNKPIDK----------
>Dbif\P-element
MNY----------CYFCRKNVPG---------VKIIHAPKCEMKRKL-WEESLGC--
SLSKNS-------QICDTHFNAS-QWRTAL-KGKIYKK-
RRLNNDAVPQREKED----------
----------ESVKEGYANASTETEDTVINH-----STSMEIKTLRQKIRALE----------
DEVQSLRKLVEDAS----------QLEKSLSTI--FTQTQIKILKSGGKRSE----FNSDD------
ISWAMCLHTAGPRAYNHLY----------
KKGFPLPCRATLYKWLSNVEIQTGCLDVVIDLMDN--
MDMDTADKLCVLAFDEMKVAGTFEYDSSADLVYEPSEY----------
VQLAMVRGLKKSWKQPV-FFDYDTRMDVPTLYELIKK----
LHRRGYFVVSIVSDMGAGNQRLWRELGISE----------
EKTWFGHPEDEDLKIFVFSDAPHLIKLVRNHYLAT-GL----HINGQTLTKSTVEQTITHCC--
KTDVTILFKVNESHLNVRSFAKQKVKLATQLFSNTTASAIRRCYSLGYQV----
ENAVETSDLFKLLN-DWFDVFNSKLSTSNCIETTQPYGKQLELQRDILKQMSHIM------
SNRICGQTHRLP----FQKGILINNASLDGLHA-YC-
NEKYGMEYILTSRLNQDIVENFFGAMRAKGGQHDHPSPLQFKYRLRKYIVA----------
----------KNTELLAGNGNVDED-NCDSWLNLNIT-------
PNGNKENEPDEGKWKGWSKEFEEFEIEMDNNIAAEYIM-DEL----------TED--
AMEYLAGYVVRKLR---LSNESTQSGF----------
TYVDEVSHGGLIKPSDQFTATLKHLESIFINNIHNTIEITKDIKKKLL---
IAAKH----------VQIDNNVKQFYFKTRIYFRLKYLNKKLAIKNQKQRLVGNS--
KLLKIKL----------
>Lmik\P-element
MNY----------CYFCRKNVPG---------VKIIHAPKCEMKRKL-WEESLGC--
SLSKNS-------QICDTHFNAS-QWRTAP-KGKIYKK-

RRLNNDAVPQREKED----------
----------ESVKEGYANASTETEDTVINH-----STSMEIKTLRQKIRALE----------
DEVQSLRKLVEDAS----------QLEKSLSTI--FTQTQIKILKSGGKRSE----FNSDD------
ISWAMCLHTAGPRAYNHLY----------
KKGFPLPCRATLYKWLSNVEIQTGCLDVVIDLMDN--
MDMDTADKLCVLAFDEMKVAGTFEYDSSADVVYEPSEY----------
VQLAMVRGLKKSWKQPV-FFDYDTRMDVPTLYELIKK----
LHRRGYFVVSIVSDMGAGNQRLWRELGISE----------
EKTWFGHPEDEDLKIFVFSDAPHLIKLVRNHYLAT-GL----HINGQTLTKSTVEQTITHCC--
KTDVTILFKVNESHLNVRSFAKQKVKLATQLFSNTTASAIRRCYSLGYQV----
ENAVETSDLFKLLN-
DWFDV----------

>Dmel\P-element
MKY----------CKFCCKAVTG---------VKLIHVPKCAIKRKL-WEQSLGC--
SLGENS-------QICDTHFNDS-QWKAAPAKGQTFKR-
RRLNADAVPSKVIEPEP----------
----------EKIKEGYTSGSTQTE---------SCSLFNENKSLREKIRTLE----------
YEMRRLEQQLRESQ----------QLEESLRKI--FTDTQIRILKNGGQRAT----FNSDD------
ISTAICLHTAGPRAYNHLY----------KKGFPLPSRTTLYRWLSDVDIKRGCLDVVIDLMDS--
DGVDDADKLCVLAFDEMKVAAAFEYDSSADIVYEPSDY----------
VQLAIVRGLKKSWKQPV-FFDFNTRMDPDTLNNILRK----
LHRKGYLVVAIVSDLGTGNQKLWTELGISE----------
SKTWFSHPADDHLKIFVFSDTPHLIKLVRNHYVDS-GL----TINGKKLTKKTIQEALHLCN--
KSDLSILFKINENHINVRSLAKQKVKLATQLFSNTTASSIRRCYSLGYDI----
ENATETADFFKLMN-DWFDIFNSKLSTSNCIECSQPYGKQLDIQNDILNRMSEIM-----RTGIL-
DKPKRLP----FQKGIIVNNASLDGLYK-YL-
QENFSMQYILTSRLNQDIVEHFFGSMRSRGGQFDHPTPLQFKYRLRKYIIGM----------
----------TNLKECVNK-NVIPD-NSESWLNLDFSSKEN---ENKSKDDEPVDDE----------
PVDEMLSNIDFTEM-DEL----------TED--AMEYIAGYVIKKLR----
ISDKVKENLTF----------TYVDEVSHGGLIKPSEKFQEKLKELECIFLHYTNN-
NNFEITN-------NVKEKL-ILAARN----------
VDVDKQVKSFYFKIRIYFRIKYFNKKIEIKNQKQKLIGNS--
KLLKIKL----------
>Dwil\P-element
MKY----------CKFCCKAVTG---------VKLIHVPKCAIKRKL-WEQSLGC--
SLGENS-------QICDTHFNDS-QWKAAPAKGQTFKR-
RRLNADAVPSKVIEPEP----------
----------EKIKEGYTSGSTQTE---------SCSLFNENKSLREKIRTLE----------
YEMRRLEQQLRESQ----------QLEESLRKI--FTDTQIRILKNGGQRAT----FNSDD------
ISTAICLHTAGPRAYNHLY----------KKGFPLPSRTTLYRWLSDVDIKRGCLDVVIDLMDS--
DGVDDADKLCVLAFDEMKVAAAFEYDSSADIVYEPSDY----------
IQLAIVRGLKKSWKQPV-FFDFNTRMDPDTLNNILRK----
LHRKGYLVVAIVSDLGTGNQKLWTELGISE----------
SKTWSSHPADDHLKIFVFSDTPHLIKLVRNHYVDS-GL----TINGKKLTKKTIQEALHLCN--
KSDLSILFKINENHINVRSLAKQKVKLATQLFSNTTASSIRRCYSLGYDI----
ENATETADFFKLMN-DWFDIFNSKLSTSNCIECSQPYGKQLDIQNDILNRMSEIM-----RTGIL-
DKPKRLP----FQKGIIVNNASLDGLYK-YL-
QENFSMQYILTSRLNQDIVEHFFGSMRSRGGQFDHPTPLQFKYRLRKYIIA----------
----------RNTEMLRNSGNIEED-NSESWLNLDFSSKEN---ENKSKDDEPVDD----------

EPVDEMLSNIDFTEM-DEL----------TED--AMEYIAGYVIKKLR----
ISDKVKENLTF----------TYVDEVSHGGLIKPSEKFQEKLKELECIFLHYTNN-
NNFEITNN-------VKEKL-ILAARN----------
VDVDKQVKSFYFKIRIYFRIKYFNKKIEIKNQKQKLIGNS--
KLLKIKL----------
>Dhel\P-element
MKY----------CKFCCKVVTG---------VSLVHVPKCNMKRKL-WEQSLGC--
HLGENS-------QICATHFNDS-QWKSTPNKGETNKR-
RRLNKDAIPTIEIEPEP----------ENVKEGYASSSTQTE---------CCSLSNENKSLRQMIRAME----------
YDLQRLRNQLEESR----------QLEESLGKF--FTEAQIKILKNGGKRST----FTSDD------
LSAAICLHTAGPRAYNHLY----------KKGFPLPSRTTLYRWLSDVEIKTGCLDVAIDLMEN--
DAMDEADKLCVLAFDEMKVAAAFEYDSSADVIYEPSNY----------
VQLAIVRGLKKSWKQPI-FFDFSTRMDADTLNNIIRK----
LHTKGYPVVAIVSDLGSGNQKLWSELGVSE----------
SKSWFSHPTDEHLKISVFPDTPHLIKLVRNHYVDS-GL----TLYGKKLTKTTVQQTLNYCA--
KSDVSILFKISENHLNVRSLDKQKVNLATQLFSNTTASSIRRCYSLGYDV----
ENACETSDLFKLLN-DWFDLFNSKLSTANCIQSTQPYGKQLPFQRDVLEKMSKIM-----
SSEIL-GKSRKLP----FQKGILVNNASLDGLYI-YL-
KDKYKMEYLLTSRLNQDIVDNFFGAMRSRGGQFDHPTPLQFKYRLKKYLIA----------
----------KNTELLRNTGNVEED-NFDSWLNLDFSSKSL---
RNKPEDDEPEDDEQGIANNIPAVIEIDELTEDGMD----------VAGYVIK----
RRRMSDCCKQSPTF----------
TYVDEVSHGGLIKPSDQFKNKLKELKIIFSHYTKEKFEITNNLKEK-----
LAAQN----------VELDKLVISFYFKMRIYLRVKYLNKKMYIKNQKRRLIGNS--
KLLKIKL----------
>Spal\P-element
MKY----------CKFCCKVVTG---------VSLVHVPKCNIKRKL-WEQSLGC--
TLGENS-------QICATHFNDS-QWKTTPNKGQANKR-
RRLNTDAIPTKEKEPEP----------EHVKEGYTSSSTQTECECVC---
RCCSLSTENKSLTETIQAME----------YDLQRLRNQLEESR----------QLEESLAKI--
FTETQIKILKNGGKRST----FTSDD------ISAAICLHTAGPRAYTHLY----------
KKGFPLPSRTTLYRWLSDVEIKPGCLDVAIDLMEN--
DAIDEADKLCVLAFDEMKVAAAFEYDSSADVVYVPSNY----------
VQLAIVRGLKKSWKQPI-FFDFSTRMDADTLNNIIRK----
LHTKGYPVVAIVSDLGSGNQRLWSELGVSECKFFTSIKIKNNLSLIFCNCFLAKIWFSHPTDE
NSKIFVFSDTPHLIKLVRNHYVDS-GF----TLNGKKLTKTTVQQTLNHCA--
KSDVSILYKISENHLNVRSLEKQKVKLATQLFSNTTASSIRRCYSLGYDV----
ENACETSDLFKLLN-DWFDVFNSKLSTANCIQSTQPYGKQLEFQRDVLEKMTQLM-----
CSDILG-RSQKLP----FQKGIIVNNASLDGLFI-YL-
KDKYNMEYLLTSRLNQDIVENFFGAMRSRGGQYDHPTPLQFKYRLRKYLIGMSNLEELCG
VLCVLSFMCFFQFYYLIIIFILYPAKNTELLRNTGNVAED-
NCDSWLNLDFNSKSLEKKENKPEDVEPEDVEPEDEADEDDDCIANNIPADIEM----
DEL----------TED--AIEYVAGYV---IK-RLRLSDCLKQSSTF----------
SYVDEVSTGGLLNRSDEFKNKLKELEIIFSH-FAK-DNFQVTNNNFKVTNNLKEKL-
VVAAQN----------VELDKLVISFYFKIRIYFRVKYLNKKICIKNQKQRLIGNS--
KLLKIKL----------
>Apis_P-1 DNA
MDFRF----------PLKNPERNQLWINAVGRKGFIPTKNS------
AICSSHFVPS---DFKINPGGNYRL--
HLNDTSVPSVFPGNTSEKKSQIEQ----------

NIEQNSPIREIIDTNNLLLTTPSK----------
VTARRPLFSPKTKSKHTSEPLKPRKIESKKKIKLLQQGIRRRD----------
KKINNLKSLLQNMRFKGLIEEQCEQ------LLIDQFEGT--SQEIFCNELKNKGKRPTGYR-
YSIQL------
KEFAATLHYYSPKALKYCRYYTFCYHKLFCYCFRTFLKLPSGKSILNWTSSINGEPGFFKEV
FDTLQT----
MSPDDRHCNLIFDAMSIKKQISWDERLGKFVGYCDYGNAFELEGSETPATETLVFMLTSING
KWKLPIGY-
VFQNKITASIQAELIKSALTHAHNAGMTVWSVTCDGAYTNVCTLKLLRCKISNSYDD-------
--------IESWFEHPV-TRSKVYYIPDACHMLKLARNILANNYVL----
ESDTGYIRWDHIRNLFKVQK--DLTLKLANKLSMVHVNWHN-
NKMRVQYAAQTLSSSTADSLEYLKNINFPGF---ENVEATVEYCRAID-
RIFDFLNSKSKFSNAFKS-PIYYNTIEKREEIIIPLIKYL-----
YTLKFKGSPLHISSKKTFILGFAIAVKSFFSMSR-
SHFVQHPNFKYLLTYKFSQDHLELLFGRIRQRLGSNNNPNTAQFKTAIKQILM----------
----------KNAIKCRSKHNCNTFDDDPIGSLFDFKWKTK---
KDDIDYEKVNFETIDIEALKKLQLLNSYIPESNNYSSTLQDA----------KNN--
ILYYIVGYLIRKLN---LDCSSCENAVLDKFNEHDYCKSLSFT--
KFVNFKNRSGLVFGSKSVFLIILEAEKMFLFLTDNFKSLQIPNLEI----
KIIKHVIVTFSTDKNIFPNLNCENISILERPHKILLITLLTKKCLKLRLKSFSKMYSSDIMNPVS
KRHKLSKLILFSNQ----------
>Hsap\THAP NM_024672
MTRS----------CSAVGCSTRDTVLSR-ERGLSFHQFPTDTIQRSK-
WIRAVNRVDPRSKKIWIPGPGAILCSKHFQES--------
DFESYGIRRKLKKGAVPSVSLYKIPQGVHLKGKARQ----------
----------KILKQPLPDNSQEVATEDHNYSLKT-
PLTIGAEKLAEVQQMLQVSKK-------RLISVKNYRMIKKRKGLRLI----------DALVEEKLL--
SEETECLLRAQFSDFKWELYNWRETDEYSAEMKQFACTLYLCSSKVYDYVR----------
KILKLPHSSILRTWLSKCQPSPGFNSNIFSFLQRRVENGDQLYQYCSLLIKSMPLKQQLQWDP
SSHSLQGFMDFGLGKLDADETPLASETVLLMAVGIFGHWRTPLGYF-
FVNRASGYLQAQLLRLTIGKLSDIGITVLAVTSDATAHSVQMAKALGIHIDG--
DD----------MKCTFQHPSSSSQQIAYFFDSCHLLRLIRNAFQNFQSI----
QFINGIAHWQHLVELVALEE--QELSN-MERIPSTLANLKN-
HVLKVNSATQLFSESVASALEYLLSLDLPPF---QNCIGTIHFLRLIN-
NLFDIFNSRNCYGKGLKGPLLPETYSKINHVLIEAKTIFVTLSDTSNNQIIKGKQKLG----FL-
GFLLNAESLKWLYQNYVFPKVMPFPYLLTYKFSHDHLELFLKMLRQVLVTSSSPTCMAFQK
AYYNLET----------RYKFQ-
DEVFLSKVSIFDISIARRKDLALWTVQRQYGVSVTKTVFHEEGICQDWSHCSLSEALLDL----
--------SDHRRNLICYAGYVANKLS-ALLTCEDCITALYASDLKASKIGSLLFVKKK-------
NGLHFPSESLCRVINICERVVRTHSRM-
AIFELVSKQRELYLQQKILCELSGHINLFVDVNKHLFDGEVCAINHFVKLLKDIIICFLNIRAK
NVAQNPLKHHSERT-------DMKTLSRKHWSSVQDYKCSSFANTSSKFRHLLSNDGYPFK

A.2 SUPPLEMENTARY TABLES

Table S1. Primer sequences

Primer Name	Sequence 5'->3'
BL	AGTAGGTTRCGCAAAGAGC
BR	CACGATTGAGTRAYASTAGG
P1	CTGAGGAAGAAATGGCTGC
P2	AGTCCGAGTGCCTGATAGG
P3	TACTGACGACATTGCCAGG
P4	CTATTCTTTGAGCCCAATGC
P5	GCTGAGAGGCTTAGTGGCAGAC
P6	ACTGTCTCCAGAGCCGAACG
P7	TCTAATACAGTGGCAGCAGC
P8	GAGTACCGAGAGCAGATAGC
P9	TACGGGCTCGCATTTGAAAC
P10	CTGACTTATCGGCTCAACC
P11	CTTCGGTCGTATTTGTTGG
P12	AGTATAAGTCCTGTCATTGAAG
P13	AGCCCTCCTTCACTAACTTGG
P14	CTACTGGCCTCTTCAATATG
P15	ATTTACACTACTGGCCTCTTC
InvL	GGACCAAGTTAGTGAAGGAG
InvR	CTATCAGGCACTCGGACTC

Table S2. Samples and results from *BuT5* screening

Sample ID	Species	BuT5 PCR (BR+BL)	BuT5 Clones			Dot blot
G001	D aldrichi	+	G1_4			
G032	D. aldrichi	+	G32_7			
G036	D. aldrichi	+	G36_3			
G099	D. aldrichi	+	G99_8			
G135	D. aldrichi	+	G135_2	G135_4	G135_6	
G002	D.arizonae	+	G2_2	G2_4		
G003	D.arizonae	+	G3_4	G3_5		
G024	D. arizonae	+	G24_2	G24_3		
G026	D. arizonae	+	G26_4			
G028	D. arizonae	+	G28_2	G28_4		
G029	D. arizonae	+	G29_10			
G095	D. arizonae	+	G95_1	G95_2		
G096	D. arizonae	+	G96_2	G96_3		
G098	D. arizonae	+	G98_2	G98_3		
G005	D. koepferae	+	G5_2			
G044	D. hydei	-				-
G092	D. hydei	-				-

G094	D. hydei	-				-
G007	D. martensis	+	G7_1	G7_2		
G039	D. mercatorum	-				+
G058	D. mercatorum	-				+
G059	D. mercatorum	-				+
G009	D. mojavensis	+	G9_1	G9_2		
G023	D. mojavensis	+	G23_2			
G031	D. mojavensis	+	G31_1			
G033	D. mojavensis	+	G33_2			
G034	D. mojavensis	+	G34_2			
G035	D. mojavensis	+	G35_2			
G037	D. mojavensis	+	G37_2	G37_4		
G090	D. mojavensis	+	G90_1	G90_2	G90_3	
G091	D. mojavensis	-				+
G097	D. mojavensis	+	G97_7			
G010	D. mulleri	+	G10_10			
G011	D. mayaguana	-				+
G012	D. buzzatii	+	G12_3			
G100	D. buzzatii	+	G100_1			
G101	D. buzzatii	+	G101_1	G101_2		
G102	D. buzzatii	+	G102_2	G102_3		
G103	D. buzzatii	+	G103_1	G103_3		
G104	D. buzzatii	+	G104_2	G104_3		
G105	D. buzzatii	+	G105_3			
G106	D. buzzatii	+	G106_2			
G107	D. buzzatii	+	G107_2	G107_3		
G108	D. buzzatii	+	G108_2	G108_3		
G109	D. buzzatii	+	G109_1			
G110	D. buzzatii	+	G110_3			
G111	D. buzzatii	+	G111_4			
G015	D. stalkeri	+	G15_1			
G016	D. starmeri	+	G16_3			
G018	D. uniseta	+	G18_1			
G025	D. wheeleri	+	G25_10	G25_11		
G027	D. wheeleri	+	G27_10			
G022	D. repleta	+	G22_4			
G030	D. navojoa	+	G30_2			
G086	D. navojoa	+	G86_2			
G038	D. mettleri	-				+
G127	D. mettleri	-				
G040	D. anceps	-				+
G041	D. borborema	+	G41_1			
G042	D. fulvimacula	-				+
G118	D. fulvimacula	-				+
G043	D. longicornis	+	G43_1			
G045	D. richardsoni	+	G45_1	G45_2		

G047	D. ritae	+	G47_1			
G052	D. hamatofila	+	G52_4			
G053	D. hexastigma	+	G53_1			
G125	D. hexastigma	+	G125_1			
G054	D. paranaensis	-				+
G055	D. peninsularis	-				+
G056	D. meridiana	-				
G057	D. neorepleta	-				+
G084	D. nigrospiracula	-				-
G085	D. eremophila	-				
G116	D. eremophila	-				+
G088	D. huckinsi	+	G88_10			
G089	D. spenceri	+	G89_2	G89_3		
G120	D. spenceri	+	G120_4			
G117	D. pegasa	-				-
G119	D. desertorum	+	G119_2			
G121	D. huichole	+	G121_2			
G122	D. bifurca	-				+
G123	D. leonis	+	G123_1			
G124	D. mailandi	+	G124_1	G124_3		
G126	D. nigricruria	+	G126_8			
G131	D. racemova	-				+
G129	D. wassermani					-
G130	D. nannoptera					-
Sequenced	D. mojavensis					+
Sequenced	D. virilis					-
Sequenced	D. grimshawi					-
Sequenced	D. willistoni					-
Sequenced	D. persimilis					-
Sequenced	D. pseudoobscura					-
Sequenced	D. ananassae					-
Sequenced	D. erecta					-
Sequenced	D. yakuba					-
Sequenced	D. melanogaster					-
Sequenced	D. sechellia					-
Sequenced	D. simulans					-

a. Locus refers to the coordinates in *D. mojavensis* genome, version CAF1.

Table S3. Combination of primers used to test exon-intron boundaries.

Primers of RT-PCR	Primers of 1st PCR	Primers of 2nd PCR	Clones sequenced[d]

P4	P1 + P4	P2 +P4	H_24_1
			H_24_2
			H_24_3
			H_24_4
			H_24_5
			H_24_6
			H_24_7
			H_24_8
			H_24_9
			H_24_10
			H_24_11
			H_24_12
			H_24_13
			H_24_14
			H_24_15
			H_24_16
			H_24_17
			H_24_18
			H_24_19
			H_24_20
			H_24_21
			H_24_22
			H_24_23
P15	P3 + P15	P5 + P12	H_512_1
			H_512_2
P15	P6 + P15	P7 + P12	H_712_1
			H_712_2
			H_712_3
			H_712_4
P15	P8 + P15	P9 + P12	O_912_1
			O_912_2
			O_912_3
			O_912_4
P15	P10 + P15	P11 + P14	H_1114_1
			H_1114_2
			H_1114_3
			H_1114_4
			O_1114_7
			O_1114_8

a. In clone names, H indicates they were obtained from head cDNA and O, from cDNA from ovaries.

Table S4. *BuT5* copies previously deposited in NCBI

Copy	Gen Bank acc. number	Length (bp)	TIR 5'	TIR 3'	TSD
BuT5-1	AH010797	1039	+	+	ATGAGAGGC
BuT5-2	AY187768	669	-	+	-
BuT5-2	AY187786	339	-	+	-
BuT5-2	AY187787	343	-	+	-
BuT5-2	AY187788	343	-	+	-
BuT5-2	AY187789	339	-	+	-
BuT5-2	AY187790	424	-	+	-
BuT5-3	AY187769	8	-	+	-
BuT5-3	AY187796	8	-	+	-
BuT5-3	AY187797	8	-	+	-
BuT5-3	AY187798	8	-	+	-
BuT5-3	AY187799	8	-	+	-
BuT5-3	AY187800	8	-	+	-
BuT5-4	AY187769	33	-	+	-
BuT5-4	AY187796	33	-	+	-
BuT5-4	AY187797	33	-	+	-
BuT5-4	AY187798	33	-	+	-
BuT5-4	AY187799	33	-	+	-
BuT5-4	AY187800	34	-	+	-
BuT5-5	AY187769	390	+	-	-
BuT5-5	AY187796	391	+	-	-
BuT5-5	AY187797	391	+	-	-
BuT5-5	AY187798	391	+	-	-
BuT5-5	AY187799	390	+	-	-
BuT5-6	AY187769	1041	+	+	CATACAACA
BuT5-7	AY187800	319	-	+	-
BuT5-8	AY900632	201	-	+	-
BuT5-8	GU132446	199	-	+	-
BuT5-8	GU132447	185	-	+	-
BuT5-8	GU132448	202	-	+	-
BuT5-8	GU132449	202	-	+	-
BuT5-8	GU132450	202	-	+	-
BuT5-8	GU132451	202	-	+	-
BuT5-8	GU132452	199	-	+	-
BuT5-8	GU132453	201	-	+	-
BuT5-9[b]	GU132454	1039	+	+	ACTAGAAC

b. BuT5_9 is listed in NCBI as BuT5_7

Table S5. *BuT5* copies from the *Drosophila mojavensis* genome

Copy	Locus[a]	Length (bp)	TIR 5'	TIR 3'	TSD
BuT5-10	scaffold_2697:99..357	259	+	-	-
BuT5-11	scaffold_3716:15..217	203	-	+	-
BuT5-12	scaffold_3911:4327..5358	1032	+	+	GATGGGAGC
BuT5-13	scaffold_3957:2549..3270	722	-	-	-
BuT5-14	scaffold_4398:8859..9480	622	-	-	-
BuT5-15	scaffold_4640:7342..8347	1006	+	+	GGCATCGGC
BuT5-16	scaffold_4852:2580..3222	643	-	-	-
BuT5-17	scaffold_5783:1035..1195	161	-	-	-
BuT5-18	scaffold_5952:755..1428	674	+	-	-
BuT5-19	scaffold_6328:2795805..2796836	1032	+	+	CACTGCTGC
BuT5-20	scaffold_6328:4426476..4426745	270	-	+	-
BuT5-21	scaffold_6350:3789..4455	667	-	-	-
BuT5-22	scaffold_6473:13077948..13078979	1032	+	+	CCGTGGAA
BuT5-23	scaffold_6473:13851404..13851585	182	-	+	-
BuT5-24	scaffold_6473:13851614..13851695	82	-	+	-
BuT5-25	scaffold_6473:16104741..16105452	712	-	-	-
BuT5-26	scaffold_6482:1468156..1468286	131	-	-	-
BuT5-27	scaffold_6482:1790620..1791653	1034	+	+	A(G)TAGG(A/T)TTC
BuT5-28	scaffold_6482:2093876..2094577	702	-	-	-
BuT5-29	scaffold_6482:2104822..2105343	522	-	-	-
BuT5-30	scaffold_6496:6887669..6888680	1012	+	+	-
BuT5-31	scaffold_6498:368474..369511	1038	+	+	TTAACAACC
BuT5-32	scaffold_6498:411006..411978	973	-	-	-
BuT5-33	scaffold_6498:720355..720783	429	+	-	-
BuT5-34	scaffold_6498:1222433..1222578	146	+	-	-
BuT5-35	scaffold_6498:1680527..1681281	755	+	+	GATTTCCTG
BuT5-36	scaffold_6498:2537503..2537583	81	-	-	-
BuT5-37	scaffold_6498:2768890..2769921	1032	+	+	CTCTTCAC
BuT5-38	scaffold_6498:2873056..2873502	447	+	-	-
BuT5-39	scaffold_6498:2937572..2938288	717	+	-	-
BuT5-40	scaffold_6500:5810655..5811686	1032	+	+	TATGTACAT
BuT5-41	scaffold_6500:6701217..6702248	1032	+	+	ATATCTACC
BuT5-42	scaffold_6500:11969560..11970275	716	+	+	AATCGCAGC
BuT5-43	scaffold_6500:18729683..18730714	1032	+	+	TTTTTCAT
BuT5-44	scaffold_6500:27636072..27636504	433	+	+	ATCATTGCT
BuT5-45	scaffold_6500:30489556..30489722	167	-	-	-
BuT5-46	scaffold_6500:30875358..30876365	1008	-	+	-
BuT5-47	scaffold_6500:31171996..31172078	83	-	-	-
BuT5-48	scaffold_6540:4510250..4511281	1032	+	+	GTGAAAAGT
BuT5-49	scaffold_6540:24251254..24252296	1043	+	+	ATTTTATAG
BuT5-50	scaffold_6540:25967333..25968313	981	+	-	-
BuT5-51	scaffold_6541:1506346..1507377	1032	-	+	-
BuT5-52	scaffold_6541:1514826..1515857	1032	+	+	AGTACGCAT
BuT5-53	scaffold_6541:1597411..1599200	996	+	+	ACGTAGTC

BuT5-54	scaffold_6541:1601584..1602548	965	+	-	-
BuT5-55	scaffold_6541:2016973..2017994	1022	-	+	-
BuT5-56	scaffold_6541:2267531..2268515	985	-	+	-
BuT5-57	scaffold_6680:24390307..24390365	59	-	-	-
BuT5-58	scaffold_6680:24391099..24391222	124	-	-	-

Table S6. *P*-element copies copies from the *Drosophila mojavensis* genome

Copy	Locus[c]	Length (bp)	TIR 5'	TIR 3'	BuT5 similarity
P-element_D.moj_1	scaffold_6473:13812021..13809886	2136	-	-	-
P-element_D.moj_2	scaffold_6541:2462460..2465680	3221	+	-	5': 98 bp; 3': 263 bp
P-element_D.moj_3	scaffold_6680:21065852..21068297	2446	+	-	5': 90 bp

c. Locus refers to the coordinates in *D. mojavensis* genome, version CAF1.

B

SUPPLEMENTARY MATERIAL OF TE ANALYSES IN *DROSOPHILA BUZZATII* GENOMES

B.1 TE DENSITY IN *D. BUZZATII* AND *D. MOJAVENSIS* CHROMOSOMES

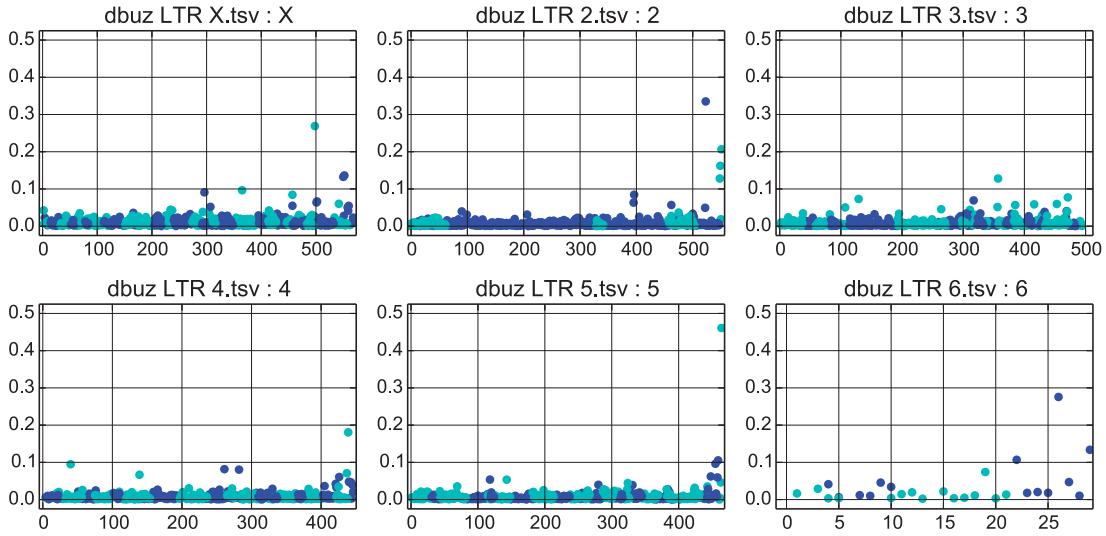

Figure 9.: Chromosomal LTR density in *D. buzzatii* . LTR order density in 50 kb non-overlapping windows. Only mapped and oriented scaffolds are present, N90 for *D. buzzatii* st-1. Changes in dot colors denote scaffold changes.

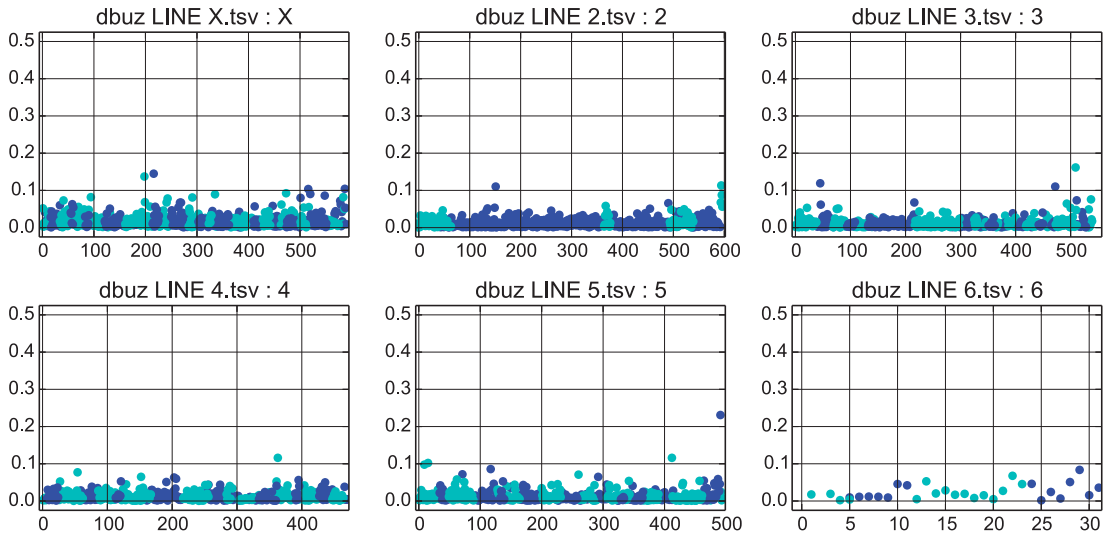

Figure 10.: Chromosomal LINE density in *D. buzzatii* . LINE order density in 50 kb non-overlapping windows. Only mapped and oriented scaffolds are present, N90 for *D. buzzatii* st-1. Changes in dot colors denote scaffold changes.

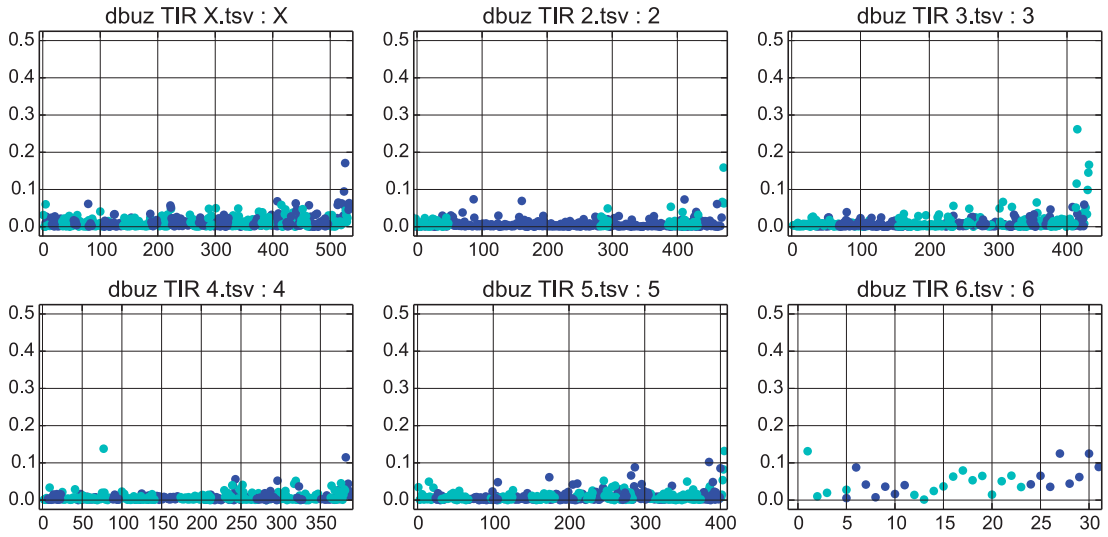

Figure 11.: Chromosomal TIR density in *D. buzzatii* . TIR order density in 50 kb non-overlapping windows. Only mapped and oriented scaffolds are present, N90 for *D. buzzatii* st-1. Changes in dot colors denote scaffold changes.

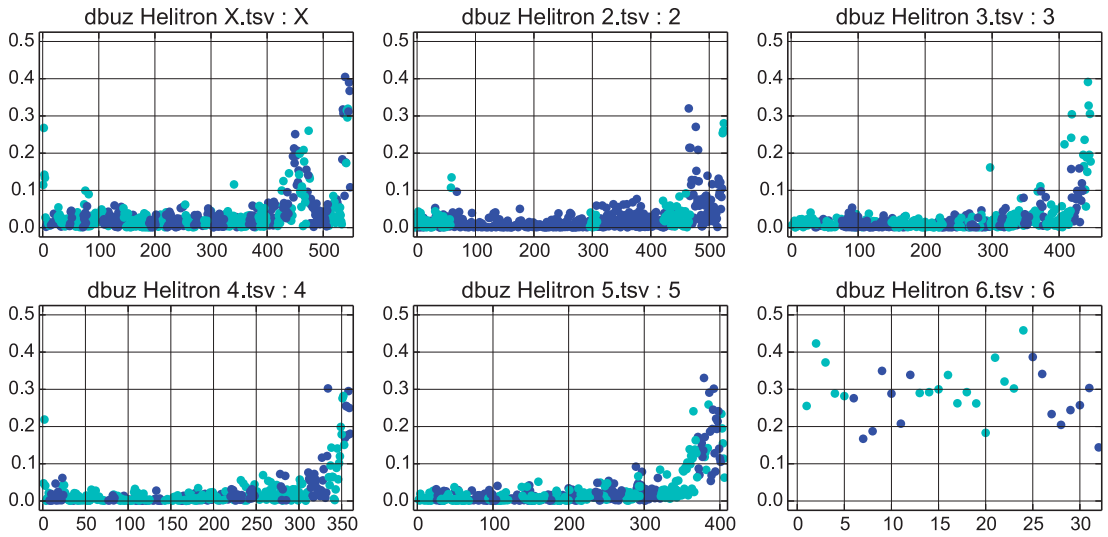

Figure 12.: Chromosomal Helitron density in *D. buzzatii* . Helitron order density in 50 kb non-overlapping windows. Only mapped and oriented scaffolds are present, N90 for *D. buzzatii* st-1. Changes in dot colors denote scaffold changes.

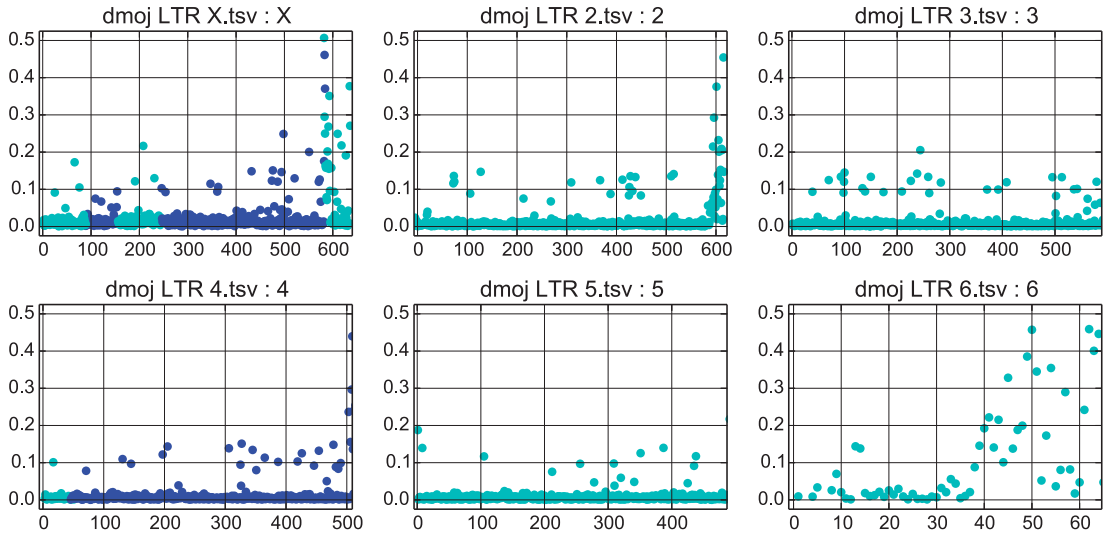

Figure 13.: Chromosomal TE density in *D. mojavensis* . LTR order density in 50 kb non-overlapping windows. Only mapped and oriented scaffolds are present, N80 for *D. mojavensis*. Changes in dot colors denote scaffold changes.

B.2 SUPPLEMENTARY TABLES

4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14,15

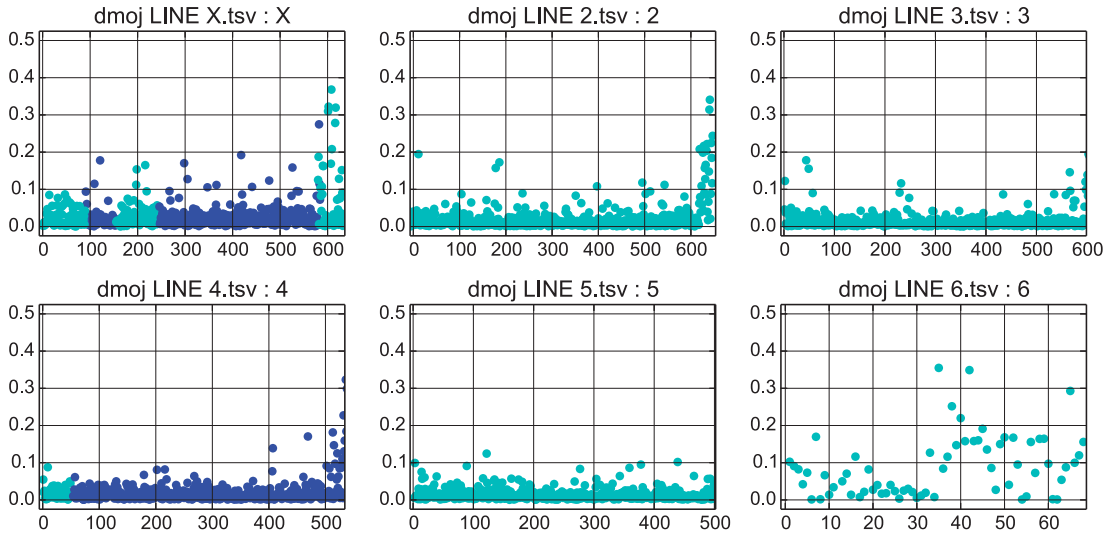

Figure 14.: Chromosomal TE density in *D. mojavensis* . LINE order density in 50 kb non-overlapping windows. Only mapped and oriented scaffolds are present, N80 for *D. mojavensis*. Changes in dot colors denote scaffold changes.

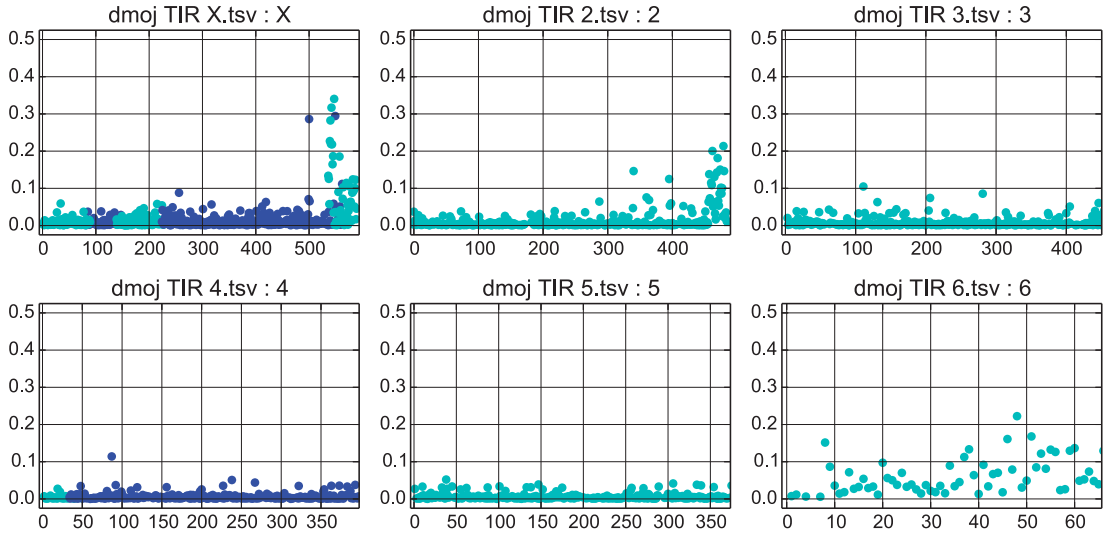

Figure 15.: Chromosomal TE density in *D. mojavensis* . TIR order density in 50 kb non-overlapping windows. Only mapped and oriented scaffolds are present, N80 for *D. mojavensis*. Changes in dot colors denote scaffold changes

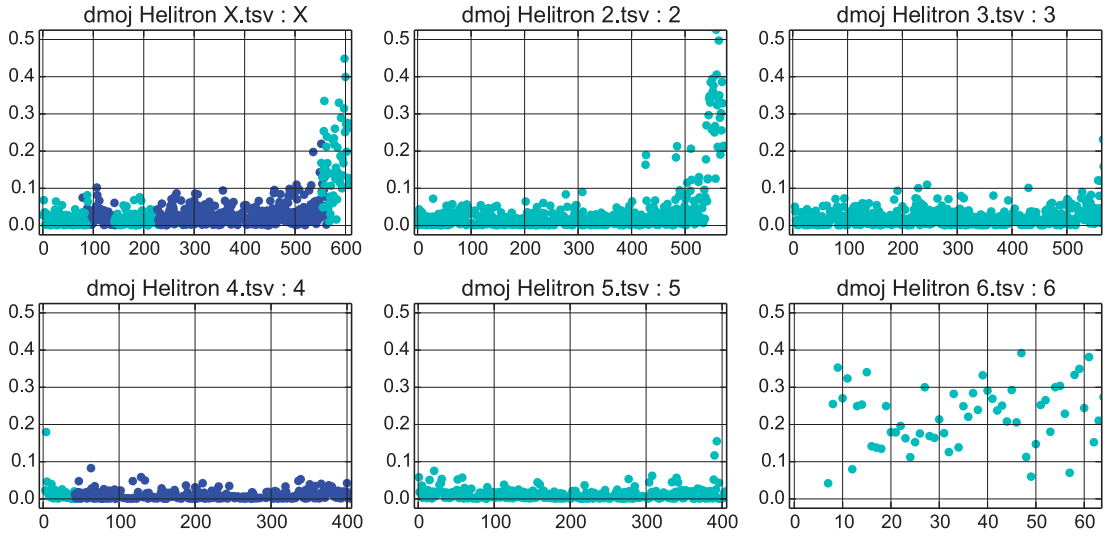

Figure 16.: Chromosomal TE density in *D. mojavensis* . Helitron order density in 50 kb non-overlapping windows. Only mapped and oriented scaffolds are present, N80 for *D. mojavensis*. Changes in dot colors denote scaffold changes

Table 4.: D statistics of two samples Kolmogorov-Smirnov tests comparing the distributions of TE densities of the pairs *D. buzzatii* Proximal

Centromeric	*D. buzzatii*			
D-value	Chr2	Chr3	Chr4	Chr5
ChrX	0.138	0.171	0.158	0.151
Chr2	-	0.179	0.128	0.227
Chr3	-	-	0.121	0.207
Chr4	-	-	-	0.224

Table 5.: D statistics of two samples Kolmogorov-Smirnov tests comparing the distributions of TE densities of the pairs *D. buzzatii* Distal + Central

Distal + Central	*D. buzzatii*			
D-value	Chr2	Chr3	Chr4	Chr5
ChrX	0.354	0.431	0.474	0.404
Chr2	-	0.095	0.144	0.080
Chr3	-	-	0.059	0.053
Chr4	-	-	-	0.080

Table 6.: D statistics of two samples Kolmogorov-Smirnov tests comparing the distributions of TE densities of the pairs *D. buzzatii* total

Total	*D. buzzatii*				
D-value	Chr2	Chr3	Chr4	Chr5	Chr6
ChrX	0.335	0.402	0.423	0.372	0.930
Chr2	-	0.085	0.117	0.063	0.947
Chr3	-	-	0.042	0.056	0.944
Chr4	-	-	-	0.064	0.946
Chr5	-	-	-	-	0.941

Table 7.: D statistics of two samples Kolmogorov-Smirnov tests comparing the distributions of TE densities of the pairs *D. mojavensis* Proximal

Centromeric	*D. mojavensis*			
D-value	Chr2	Chr3	Chr4	Chr5
ChrX	0.297	0.412	0.166	0.480
Chr2	-	0.417	0.183	0.317
Chr3	-	-	0.400	0.683
Chr4	-	-	-	0.383

Table 8.: D statistics of two samples Kolmogorov-Smirnov tests comparing the distributions of TE densities of the pairs *D. mojavensis* Central + Distal

Distal + Central	*D. mojavensis*			
D-value	Chr2	Chr3	Chr4	Chr5
ChrX	0.420	0.362	0.507	0.520
Chr2	-	0.088	0.111	0.117
Chr3	-	-	0.158	0.168
Chr4	-	-	-	0.047

Table 9.: D statistics of two samples Kolmogorov-Smirnov tests comparing the distributions of TE densities of the pairs *D. mojavensis* total

Total	*D. mojavensis*				
D-value	Chr2	Chr3	Chr4	Chr5	Chr6
ChrX	0.394	0.328	0.460	0.485	0.789
Chr2	-	0.087	0.083	0.103	0.837
Chr3	-	-	0.145	0.165	0.807
Chr4	-	-	-	0.051	0.822
Chr5	-	-	-	-	0.863

Table 10.: p-values of two samples Kolmogorov-Smirnov tests comparing the distributions of TE densities of the pairs *D. buzzatii* Proximal

Centromeric	*D. buzzatii*			
p-value	Chr2	Chr3	Chr4	Chr5
ChrX	0.609	0.337	0.437	0.498
Chr2	-	0.280	0.694	0.083
Chr3	-	-	0.765	0.146
Chr4	-	-	-	0.093

Table 11.: p-values of two samples Kolmogorov-Smirnov tests comparing the distributions of TE densities of the pairs *D. buzzatii* Distal + Central

Distal + Central	*D. buzzatii*			
p-value	Chr2	Chr3	Chr4	Chr5
ChrX	0.000	0.000	0.000	0.000
Chr2	-	0.011	0.000	0.063
Chr3	-	-	0.365	0.494
Chr4	-	-	-	0.106

Table 12.: p-values of two samples Kolmogorov-Smirnov tests comparing the distributions of TE densities of the pairs *D. buzzatii* total

Total	*D. buzzatii*				
p-value	Chr2	Chr3	Chr4	Chr5	Chr6
ChrX	0.000	0.000	0.000	0.000	0.000
Chr2	-	0.019	0.001	0.187	0.000
Chr3	-	-	0.734	0.353	0.000
Chr4	-	-	-	0.247	0.000
Chr5	-	-	-	-	0.000

Table 13.: p-values of two samples Kolmogorov-Smirnov tests comparing the distributions of TE densities of the pairs *D. mojavensis* Proximal

Centromeric	*D. mojavensis*			
p-value	Chr2	Chr3	Chr4	Chr5
ChrX	0.008	0.000	0.356	0.000
Chr2	-	0.000	0.239	0.004
Chr3	-	-	0.000	0.000
Chr4	-	-	-	0.000

Table 14.: p-values of two samples Kolmogorov-Smirnov tests comparing the distributions of TE densities of the pairs *D. mojavensis* Distal + Central

Distal + Central	*D. mojavensis*			
p-value	Chr2	Chr3	Chr4	Chr5
ChrX	0.000	0.000	0.000	0.000
Chr2	-	0.018	0.002	0.001
Chr3	-	-	0.000	0.000
Chr4	-	-	-	0.645

Table 15.: p-values of two samples Kolmogorov-Smirnov tests comparing the distributions of TE densities of the pairs *D. mojavensis* total

Total	*D. mojavensis*				
p-value	Chr2	Chr3	Chr4	Chr5	Chr6
ChrX	0.000	0.000	0.000	0.000	0.000
Chr2	-	0.012	0.028	0.003	0.000
Chr3	-	-	0.000	0.000	0.000
Chr4	-	-	-	0.467	0.000
Chr5	-	-	-	-	0.000

C

RESEARCH ARTICLE

In this Appendix the research article describing the *D. buzzatii* st-1 genome is included

C.1 GENOMICS OF ECOLOGICAL ADAPTATION IN CACTOPHILIC DROSOPHILA

GBE

Genomics of Ecological Adaptation in Cactophilic *Drosophila*

Yolanda Guillén[1], Núria Rius[1], Alejandra Delprat[1], Anna Williford[2], Francesc Muyas[1], Marta Puig[1], Sònia Casillas[1,3], Miquel Ràmia[1,3], Raquel Egea[1,3], Barbara Negre[4,5], Gisela Mir[6,7], Jordi Camps[8], Valentí Moncunill[9], Francisco J. Ruiz-Ruano[10], Josefa Cabrero[10], Leonardo G. de Lima[11], Guilherme B. Dias[11], Jeronimo C. Ruiz[12], Aurélie Kapusta[13], Jordi Garcia-Mas[6], Marta Gut[8], Ivo G. Gut[8], David Torrents[9], Juan P. Camacho[10], Gustavo C.S. Kuhn[11], Cédric Feschotte[13], Andrew G. Clark[14], Esther Betrán[2], Antonio Barbadilla[1,3], and Alfredo Ruiz[1,*]

[1]Departament de Genètica i de Microbiologia, Universitat Autònoma de Barcelona, Spain

[2]Department of Biology, University of Texas at Arlington

[3]Institut de Biotecnologia i de Biomedicina, Universitat Autònoma de Barcelona, Spain

[4]EMBL/CRG Research Unit in Systems Biology, Centre for Genomic Regulation (CRG), Barcelona, Spain

[5]Universitat Pompeu Fabra (UPF), Barcelona, Spain

[6]IRTA, Centre for Research in Agricultural Genomics (CRAG) CSIC-IRTA-UAB-UB, Campus UAB, Edifici CRAG, Barcelona, Spain

[7]The Peter MacCallum Cancer Centre, East Melbourne, Victoria, Australia

[8]Centro Nacional de Análisis Genómico (CNAG), Parc Científic de Barcelona, Torre I, Barcelona, Spain

[9]Barcelona Supercomputing Center (BSC), Edifici TG (Torre Girona), Barcelona, Spain and Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

[10]Departamento de Genética, Facultad de Ciencias, Universidad de Granada, Spain

[11]Instituto de Ciências Biológicas, Departamento de Biologia Geral, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil

[12]Informática de Biossistemas, Centro de Pesquisas René Rachou—Fiocruz Minas, Belo Horizonte, MG, Brazil

[13]Department of Human Genetics, University of Utah School of Medicine

[14]Department of Molecular Biology and Genetics, Cornell University

*Corresponding author: E-mail: alfredo.ruiz@uab.cat.

Accepted: December 23, 2014

Abstract

Cactophilic *Drosophila* species provide a valuable model to study gene–environment interactions and ecological adaptation. *Drosophila buzzatii* and *Drosophila mojavensis* are two cactophilic species that belong to the *repleta* group, but have very different geographical distributions and primary host plants. To investigate the genomic basis of ecological adaptation, we sequenced the genome and developmental transcriptome of *D. buzzatii* and compared its gene content with that of *D. mojavensis* and two other noncactophilic *Drosophila* species in the same subgenus. The newly sequenced *D. buzzatii* genome (161.5 Mb) comprises 826 scaffolds (>3 kb) and contains 13,657 annotated protein-coding genes. Using RNA sequencing data of five life-stages we found expression of 15,026 genes, 80% protein-coding genes, and 20% noncoding RNA genes. In total, we detected 1,294 genes putatively under positive selection. Interestingly, among genes under positive selection in the *D. mojavensis* lineage, there is an excess of genes involved in metabolism of heterocyclic compounds that are abundant in *Stenocereus cacti* and toxic to nonresident *Drosophila* species. We found 117 orphan genes in the shared *D. buzzatii–D. mojavensis* lineage. In addition, gene duplication analysis identified lineage-specific expanded families with functional annotations associated with proteolysis, zinc ion binding, chitin binding, sensory perception, ethanol tolerance, immunity, physiology, and reproduction. In summary, we identified genetic signatures of adaptation in the shared *D. buzzatii–D. mojavensis* lineage, and in the two separate *D. buzzatii and D. mojavensis* lineages. Many of the novel lineage-specific genomic features are promising candidates for explaining the adaptation of these species to their distinct ecological niches.

Key words: cactophilic *Drosophila*, genome sequence, ecological adaptation, positive selection, orphan genes, gene duplication.

Introduction

Drosophila species are saprophagous insects that feed and breed on a variety of fermenting plant materials, chiefly fruits, flowers, slime fluxes, decaying bark, leaves and stems, cactus necroses, and fungi (Carson 1971). These substrates include bacteria and yeasts that decompose the plant tissues and contribute to the nutrition of larvae and adults (Starmer 1981; Begon 1982). Only two species groups use cacti as their primary breeding site: *repleta* (Oliveira et al. 2012) and *nannoptera* (Lang et al. 2014). Both species groups originated at the *virilis–repleta* radiation, 20–30 Ma (Throckmorton 1975; Morales-Hojas and Vieira 2012; Oliveira et al. 2012) but adapted independently to the cactus niche. The "cactus-yeast-*Drosophila* system" in arid zones provides a valuable model to investigate gene–environment interactions and ecological adaptation from genetic and evolutionary perspectives (Barker and Starmer 1982; Barker et al. 1990). Rotting cacti provide relatively abundant, predictable, and long-lasting resources that can sustain very large *Drosophila* populations. For instance, a single saguaro rot may weigh up to several tons, last for many months, and sustain millions of *Drosophila* larvae and adults (Breitmeyer and Markow 1998). On the other hand, cacti are usually found in arid climates with middle to high temperatures that may impose desiccation and thermal stresses (Loeschcke et al. 1997; Hoffmann et al. 2003; Rajpurohit et al. 2013). Finally, some cacti may contain allelochemicals that can be toxic for *Drosophila* (see below). Thus, adaptation to use cacti as breeding sites must have entailed a fairly large number of changes in reproductive biology, behavior, physiology, and biochemistry (Markow and O'Grady 2008).

We have sequenced the genome and developmental transcriptome of *Drosophila buzzatii* to carry out a comparative analysis with those of *Drosophila mojavensis*, *Drosophila virilis,* and *Drosophila grimshawi* (*Drosophila* 12 Genomes Consortium et al. 2007). *Drosophila buzzatii* and *D. mojavensis* are both cactophilic species that belong to the *mulleri* subgroup of the *repleta* group (Wasserman 1992; Oliveira et al. 2012), although they have very different geographical distributions and host plants (fig. 1). *Drosophila buzzatii* is a subcosmopolitan species which is found in four of the six major biogeographic regions (David and Tsacas 1980). This species is originally from Argentina and Bolivia but now has a wide geographical distribution that includes other regions of South America, the Old World, and Australia (Carson and Wasserman 1965; Fontdevila et al. 1981; Hasson et al. 1995; Manfrin and Sene 2006). It chiefly feeds and breeds in rotting tissues of several Opuntia cacti but can also occasionally use columnar cacti (Hasson et al. 1992; Ruiz et al. 2000; Oliveira et al. 2012). The geographical dispersal of Opuntia by humans in historical times is considered the main driver of the world-wide expansion of *D. buzzatii* (Fontdevila et al. 1981; Hasson et al. 1995).

On the other hand, *D. mojavensis* is endemic to the deserts of Southwestern United States and Northwestern Mexico. Its primary host plants are *Stenocereus gummosus* (pitaya agria) in Baja California and *Stenocereus thurberi* (organ pipe) in Arizona and Sonora, but uses also *Ferocactus cylindraceous* (California barrel) in Southern California and *Opuntia* sp. in Santa Catalina Island (Fellows and Heed 1972; Heed and Mangan 1986; Ruiz and Heed 1988; Etges et al. 1999). The ecological conditions of the Sonoran Desert are extreme (dry, arid, and hot), as attested by the fact that only four *Drosophila* species are endemic (Heed and Mangan 1986). In addition, *D. mojavensis* chief host plants, pitaya agria and organ pipe, are chemically complex and contain large quantities of triterpene glycosides, unusual medium-chain fatty acids, and sterol diols (Kircher 1982; Fogleman and Danielson 2001). These allelochemicals are toxic to nonresident *Drosophila* species, decreasing significantly larval performance (Fogleman and Kircher 1986; Ruiz and Heed 1988; Fogleman and Armstrong 1989; Frank and Fogleman 1992). In addition, host plant chemistry and fermentation byproducts affect adult epicuticular hydrocarbons and mating behavior (Havens and Etges 2013) as well as expression of hundreds of genes (Matzkin et al. 2006; Etges et al. 2015; Matzkin 2014).

As a first step to understand the genetic bases of ecological adaptation, here we compare the genomes of the two cactophilic species with those of two noncactophilic species of the *Drosophila* subgenus: *D. virilis* that belongs to the *virilis* species group and *D. grimshawi* that belongs to the picture wing group of Hawaiian *Drosophila* (fig. 1). The lineage leading to the common ancestor of *D. buzzatii* and *D. mojavensis* after diverging from *D. virilis* (#3 in fig. 1) represents the lineage that adapted to the cactus niche (likely Opuntia; Oliveira et al. 2012), whereas the lineages leading to *D. buzzatii* (#1) and *D. mojavensis* (#2) adapted to the specific niche of each species. We carried out a genome-wide scan for 1) genes under positive selection, 2) lineage-specific genes, and 3) gene-duplications in the three lineages (fig. 1). Based on the results of our comparative analyses, we provide a list of candidate genes that might play a meaningful role in the ecological adaptation of these fruit flies.

Materials and Methods

We sequenced the genome of a highly inbreed *D. buzzatii* strain, st-1 (Betran et al. 1998). DNA was extracted from male and female adults (Piñol et al. 1988; Milligan 1998). Reads were generated with three different sequencing platforms (supplementary fig. S2 and table S12, Supplementary Material online). The assembly of the genome was performed in three stages (supplementary table S13, Supplementary Material online): Preassembly (Margulies et al. 2005), scaffolding (Boetzer et al. 2011), and gapfilling (Nadalin et al. 2012). In each step, a few chimeric scaffolds were identified and split.

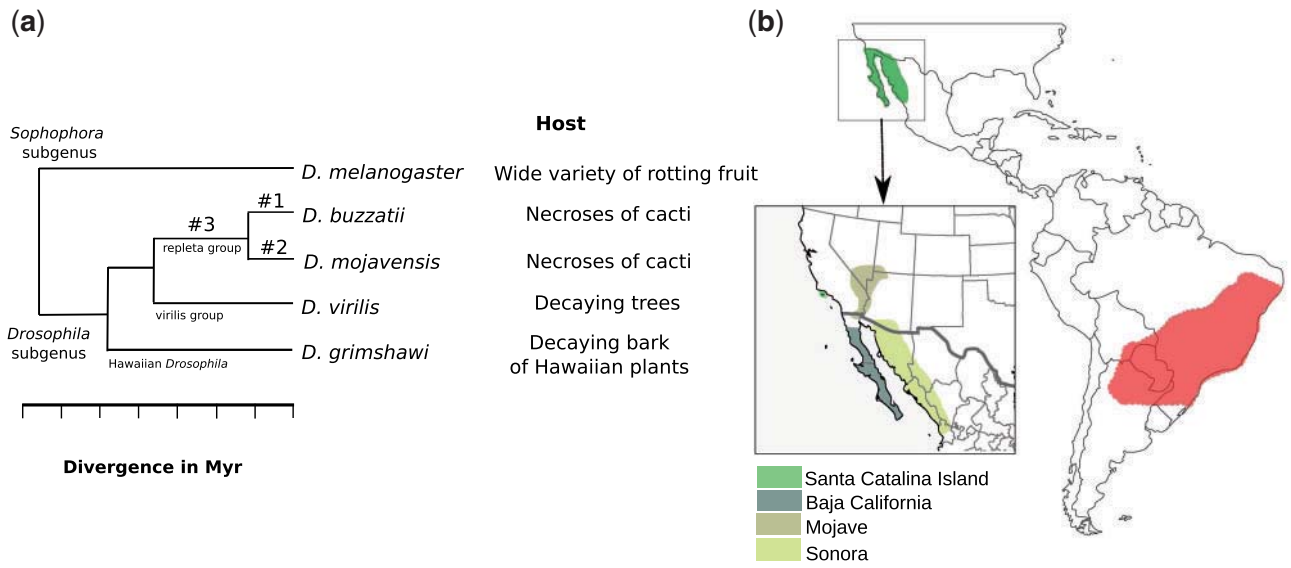

Fig. 1.—(*a*) Phylogenetic relationship of fruit fly species considered in our comparative analysis and their host preference. (*b*) Geographical distribution of cactophilic species *D. buzzatii* (red) and *D. mojavensis* (green) in America.

The final assembly, named Freeze 1, contains 826 scaffolds greater than 3 kb and N50 and N90 index are 30 and 158, respectively. The distribution of read depth in the preassembly showed a Gaussian distribution with a prominent mode centered at approximately 22× (supplementary fig. S3, Supplementary Material online). CG content is approximately 35% overall, approximately 42% in gene regions (including introns) and reaches approximately 52% in exons (supplementary table S14, Supplementary Material online). Unidentified nucleotides (N's) represent approximately 9% overall, approximately 4% in gene regions, and 0.004% in exons. Sequence quality was assessed by comparing Freeze 1 with five Sanger sequenced bacterial artificial chromosomes (BACs) (Negre et al. 2005; Prada 2010; Calvete et al. 2012) and with Illumina genomic and RNA sequencing (RNA-Seq) reads (supplementary fig. S4, Supplementary Material online). Quality assessments gave an overall error rate of approximately 0.0005 and a PHRED quality score of approximately Q33 (supplementary tables S15 and S16, Supplementary Material online). An overall proportion of segregating sites of approximately 0.1% was estimated (supplementary table S17, Supplementary Material online).

The genome size of two *D. buzzatii* strains, st-1 and j-19, was estimated by Feulgen Image Analysis Densitometry. The genome size of *D. mojavensis* 15081-1352.22 strain (193,826,310 bp) was used as reference (*Drosophila* 12 Genomes Consortium et al. 2007). Testicles from anesthetized males were dissected in saline solution and fixed in acetic-alcohol 3:1. Double preparations of *D. mojavensis* and *D. buzzatii* were made by crushing the fixed testicles in 50% acetic acid. Following Ruiz-Ruano et al. (2011), the samples were stained by Feulgen reaction and images obtained by optical microscopy were analyzed with the pyFIA software (supplementary fig. S5 and table S18, Supplementary Material online).

The 826 scaffolds in Freeze 1 were assigned to chromosomes by aligning their sequences with the *D. mojavensis* genome using MUMmer (Delcher et al. 2003). In addition, the 158 scaffolds in the N90 index were mapped, ordered, and oriented (supplementary fig. S1, Supplementary Material online) using conserved linkage (Schaeffer et al. 2008), in situ hybridization, and additional information (González et al. 2005; Guillén and Ruiz 2012). To estimate the number of rearrangements between *D. buzzatii* and *D. mojavensis*, their chromosomes were compared using GRIMM (Tesler 2002; Delprat A, Guillén Y, Ruiz A, in preparation). Genes in the *Hox* gene complex (*HOM-C*) and five other gene complexes were searched in silico in the *D. buzzatii* genome and manually annotated using available information (Negre et al. 2005), the annotated *D. mojavensis* and *Drosophila melanogaster* genomes, and the RNA-seq data generated for *D. buzzatii* (Negre B, Muyas F, Guillén Y, Ruiz A, in preparation). Transposable elements (TEs) were annotated with RepeatMasker using a comprehensive TE library compiled from FlyBase (St Pierre et al. 2014), Repbase (Jurka et al. 2005), and RepeatModeler. Tandem Repeats Finder version 4.04 (Benson 1999) was used to identify satellite DNAs (satDNAs).

For the RNA-Seq experiments, RNA from frozen samples (embryos, larvae, pupae, adult males, and adult females) was

processed using the TruSeq RNA sample preparation kit provided by Illumina. We used a Hi-Seq2000 Illumina Sequencer to generate nonstrand-specific paired-end approximately 100 bp reads from poly(A) + RNA. Between 60 and 89 million reads were generated per sample. A total of approximately 286 million filtered reads were mapped to Freeze 1 with TopHat (Trapnell et al. 2009) representing approximately 180 × coverage of the total genome size (supplementary table S19, Supplementary Material online). Transcripts were assembled with Cufflinks (Trapnell et al. 2010) using Annotation Release 1 as reference (see below).

Protein-coding genes (PCGs) were annotated combining with EVidence Modeler (EVM; Haas et al. 2008) the results of different predictors: Augustus (Stanke and Waack 2003), SNAP (Korf 2004), N-SCAN (Korf et al. 2001), and Exonerate (Slater and Birney 2005). The EVM set contained 12,102 gene models. We noticed that orthologs for a considerable number of *D. mojavensis* PCGs were absent from this data set. Thus, we used the Exonerate predictions to detect another 1,555 PCGs not reported by EVM (Poptsova and Gogarten 2010). Altogether, we predicted a total of 13,657 PCG models in the *D. buzzatii* reference genome (Annotation Release 1). Features of these models are given in supplementary table S20, Supplementary Material online. The RSD (reciprocal smallest distance) algorithm (Wall and Deluca 2007) was used to identify 9,114 1:1 orthologs between *D. mojavensis* and *D. buzzatii*. Orthology relationships among the four species in the *Drosophila* subgenus (fig. 1) were inferred from *D. buzzatii*–*D. mojavensis* list of orthologs and the OrthoDB catalog (version 6; Kriventseva et al. 2008). To test for positive selection, we compared different codon substitution models using the likelihood ratio test (LRT). We run two pairs if site models (SM) on the orthologs set between *D. buzzatii* and *D. mojavensis*: M7 versus M8 and M1a versus M2a (Yang 2007). Then, we used branch-site models (BSM) to test for positive selection in three lineages (fig. 1): *D. mojavensis* lineage, *D. buzzatii* lineage, and the lineage that led to the two cactophilic species (*D. buzzatii* and *D. mojavensis*). We run Venny software (Oliveros 2007) to create a Venn diagram showing shared selected genes among the different models. We identified genes that are only present in the two cactophilic species, *D. mojavensis* and *D. buzzatii*, by blasting the amino acid sequences from the 9,114 1:1 orthologs between *D. mojavensis* and *D. buzzatii* (excluding misannotated genes) against all the proteins from the remaining 11 *Drosophila* species available in FlyBase protein database, excluding *D. mojavensis* (St Pierre et al. 2014).

For gene duplication analysis (DNA- and RNA-mediated duplications), we used annotated PCGs from the four species of the *Drosophila* subgenus (see supplementary methods, Supplementary Material online). Briefly, we ran all-against-all BLASTp and selected hits with alignment length extending over at least 50% of both proteins and with amino acid identity of at least 50%. Markov Cluster Algorithm (Enright et al. 2002) was used to cluster retained proteins into gene families. The data set was further modified to include additional family members based on sequence coverage and to exclude family members with internal stop codons and matches to TEs. Gene counts for each family from the four species were analyzed with an updated version of CAFE (CAFE 3.1 provided by the authors; Han et al. 2013) to identify lineage-specific expansions. The sets of CAFE-identified expanded families in the *D. buzzatii* and *D. mojavensis* genomes were examined for the presence of lineage-specific duplications. Families that included members with d_s <0.4 were examined manually and lineage-specific duplications were inferred when no hits were found in the syntenic region of the genome with a missing copy. *Drosophila buzzatii*-specific RNA-mediated duplications were identified by examining intron-less and intron-containing gene family members. A duplicate was considered a retrocopy if its sequence spanned all introns of the parental gene. The number of families identified by CAFE as expanded along the internal cactophilic branch was reduced by considering only those families that were also found in expanded category after rerunning the analysis with a less stringent cutoff (35% amino acid identity, 50% coverage). The overlapping set of expanded families was manually examined to verify the absence of *D. buzzatii* and *D. mojavensis* new family members in the *D. virilis* genome. Functional annotation (i.e., Gene Ontology [GO] term) for all expanded families was obtained using the DAVID annotation tool (Huang et al. 2009a, 2009b). For genes without functional annotation in DAVID, annotations of *D. melanogaster* orthologs were used. An extended version of these methods is given as supplementary methods, Supplementary Material online.

Results

Features of the *D. buzzatii* Genome

Genome Sequencing and Assembly

We sequenced and de novo assembled the genome of *D. buzzatii* line st-1 using shotgun and paired-end reads from 454/Roche, mate-pair and paired-end reads from Illumina, and Sanger BAC-end sequences (~22× total expected coverage; see Materials and Methods for details). We consider the resulting assembly (Freeze 1) as the reference *D. buzzatii* genome sequence (table 1). This assembly comprises 826 scaffolds greater than 3 kb long with a total size of 161.5 Mb. Scaffold N50 and N90 indexes are 30 and 158, respectively, whereas scaffold N50 and N90 lengths are 1.38 and 0.16 Mb, respectively (table 1). Quality controls (see Materials and Methods) yielded a relatively low error rate of approximately 0.0005 (PHRED quality score $Q = 33$). For comparison, we also assembled the genome of the same line (st-1) using only four lanes of short (100 bp) Illumina paired-end reads (~76× expected coverage) and the SOAPdenovo software (Luo et al. 2012). This resulted in 10,949 scaffolds greater than 3 kb long with a total size of 144.2 Mb (table 1). All scaffolds are

Table 1
Summary of Assembly Statistics for the Genome of *Drosophila buzzatii*

Assembly	Freeze 1	SOAPdenovo
Number of scaffolds (>3 kb)	826	10,949
Coverage	~22×	~76×
Assembly size (bp)	161,490,851	144,184,967
Scaffold N50 index	30	2,035
Scaffold N50 length (bp)	1,380,942	18,900
Scaffold N90 index	158	7,509
Scaffold N90 length (bp)	161,757	5,703
Contig N50 index	1,895	2,820
Contig N50 length (bp)	17,678	3,101

Table 2
Transposable Element Content of *Drosophila buzzatii* Genome

Class	Order	Annotated Base Pair	Genome Coverage (%)
I (retrotransposons)	LTR	2,366,439	1.47
	DIRS	55	0.00
	LINE	2,541,645	1.57
II (DNA transposons)	TIR	2,017,167	1.25
	Helitron	5,531,009	3.42
	Maverick	189,267	0.12
	Unknown	973,759	0.60
Total		13,619,341	8.43

Note.—The classification follows Wicker et al. (2007).

available for download from the *Drosophila buzzatii* Genome Project web page (http://dbuz.uab.cat, last accessed January 7, 2015). This site also displays all the information generated in this project (see below).

Genome Size and Repeat Content

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

To assess the TE content of the *D. buzzatii* genome, we masked the 826 scaffolds of Freeze 1 assembly using a library of TEs compiled from several sources (see Materials and Methods). We detected a total of 56,901 TE copies covering approximately 8.4% of the genome (table 2). The most abundant TEs seem to be Helitrons, LINEs, long terminal repeat (LTR) retrotransposons, and TIR transposons that cover 3.4%, 1.6%, 1.5%, and 1.2% of the genome, respectively (table 2). In addition, we identified tandemly repeated satDNAs with repeat units longer than 50 bp (Melters et al. 2013) (see Materials and Methods). The two most abundant tandem repeat families are the pBuM189 satellite (Kuhn et al. 2008) and the DbuTR198 satellite, a novel family with repeat units 198 bp long (table 3). The remaining tandem repeats had sequence similarity to integral parts of TEs, such as the internal tandem repeats of the transposon Galileo (de Lima LG, Svartman M, Ruiz A, Kuhn GCS, in preparation).

Chromosomal Rearrangements

The basic karyotype of *D. buzzatii* is similar to that of the *Drosophila* genus ancestor and consists of six chromosome pairs: Four pairs of equal-length acrocentric autosomes, one pair of "dot" autosomes, a long acrocentric X, and a small acrocentric Y (Ruiz and Wasserman 1993). Because no interchromosomal reorganizations between *D. buzzatii* and *D. mojavensis* have previously been found (Ruiz et al. 1990; Ruiz and Wasserman 1993) all 826 scaffolds were assigned to chromosomes by BLASTn against the *D. mojavensis* genome. In addition, the 158 scaffolds in the N90 index were mapped to chromosomes, ordered, and oriented (supplementary fig. S1, Supplementary Material online; Delprat A, Guillén Y, Ruiz A, in preparation) using conserved linkage (Schaeffer et al. 2008) and additional information (González et al. 2005; Guillén and Ruiz 2012). A bioinformatic comparison of *D. buzzatii* and *D. mojavensis* chromosomes confirmed that chromosome 2 differs between these species by ten inversions (*2m, 2n, 2z*^{7}, *2c, 2f, 2g, 2h, 2q, 2r*, and *2s*), chromosomes X and 5 differ by one inversion each (*Xe* and *5g*, respectively), and chromosome 4 is homosequential as previously described (Ruiz et al. 1990; Ruiz and Wasserman 1993; Guillén and Ruiz 2012). In contrast, we find that chromosome 3 differs by five inversions instead of the expected two that were previously identified by cytological analyses (Ruiz et al. 1990). These three additional chromosome 3 inversions seem to be specific to the *D. mojavensis* lineage (Delprat A, Guillén Y, Ruiz A, in preparation). One of these inversions, *3f*^{2}, is polymorphic in natural populations of *D. mojavensis*, but, conflicting with previous reports (Ruiz et al. 1990; Schaeffer et al. 2008), appears to be homozygous in the sequenced strain. This has been corroborated by the cytological reanalysis of its polytene chromosomes (Delprat et al. 2014).

Many developmental genes are arranged in gene complexes each comprising a small number of functionally related genes. We checked the organization of six of these gene complexes in the *D. buzzatii* genome: *HOM-C*, *Achaete–scute* complex, *Iroquois* complex, *NK* homeobox gene cluster (*NK-C*), *Enhancer of split* complex, and *Bearded* complex (*Brd-C*) (Negre B, Muyas F, Guillén Y, Ruiz A, in preparation). *Hox* genes were arranged in a single complex in the *Drosophila* genus ancestor (Hughes and Kaufman 2002). However, this *HOM-C* suffered two splits (caused by chromosomal inversions) in the lineage leading to the *repleta* species group (Negre et al. 2005). In order to fully characterize *HOM-C* organization in *D. buzzatii*, we manually annotated all *Hox*

Table 3
Satellite DNAs Identified in the *Drosophila buzzatii* Genome

Tandem repeat Family	Repeat Length	GC Content (%)	Genome Coverage (%)[a]	Consensus Sequence[b]	Distribution
pBuM189	189	29	0.039	GCAAAAGACTCCGTCAATTA GAAAACAAAAAATGTTATAGTTTTGAGGATTAACC GGCAAAAACCGTATTATTTGTTATAT GATTTCTGTATGGAATACCGTTTTAGAA GCGTCTTTTATCGTATTACTCAGATATATCT TAAGATTTAGCATAATCTAAGAACTTTT TGAAATATTCACATTTGTCCA	*D. buzzatii* cluster species *D. mojavensis*
DbuTR198	198	34	0.027	AAGGTAGAAAGGTAGTTGGTGAGATAAACCAGAAAAA GAGCTAAAAACGGCTAAAAACGGCTAGAAAATAGCCA GAAAGGTAGATTGAACATTAATGGGCAAATGG ATGGATAAATAAGACTGGTCATCATCCAA TGAACAGAATCATGATTAAGAGATAGAAATA TGATTAGAAAGTAGGATAGAAAGGTTAGAAAG	*D. buzzatii*

[a]Genome fraction was calculated assuming a genome size of 163,547,398 bp (version 1 freeze of all contigs).
[b]Consensus sequence generated after clustering TRF results (see Materials and Methods).

genes and located them in three scaffolds (2, 5, and 229) of chromosome 2 (Negre B, Muyas F, Guillén Y, Ruiz A, in preparation). The analysis of these scaffolds revealed that only two clusters of *Hox* genes are present. The distal cluster contains *proboscipedia, Deformed, Sex combs reduced, Antennapedia* and *Ultrabithorax*, whereas the proximal cluster contains *labial, abdominal A* and *Abdominal B*. This is precisely the same *HOM-C* organization observed in *D. mojavensis* (Negre and Ruiz 2007). Therefore, there seem to be no additional rearrangements of the *HOM-C* in *D. buzzatii* besides those already described in the genus *Drosophila* (Negre and Ruiz 2007). The other five developmental gene complexes contain 4, 3, 6, 13, and 6 functionally related genes, respectively (Lai et al. 2000; Garcia-Fernàndez 2005; Irimia et al. 2008; Negre and Simpson 2009). All these complexes seem largely conserved in the *D. buzzatii* genome with few exceptions (Negre B, Muyas F, Guillén Y, Ruiz A, in preparation). The gene *slouch* is separated from the rest of the *NK-C* in *D. buzzatii* and also in all other *Drosophila* species outside of the melanogaster species group; in addition, the gene *Bearded*, a member of the *Brd-C*, is seemingly absent from the *D. buzzatii* and *D. mojavensis* genomes, although it is present in *D. virilis* and *D. grimshawi*. On the other hand, genes flanking the complexes are often variable, presumably due to the fixation of chromosomal inversions with breakpoints in the boundaries of the complexes.

PCG Content

We used a combination of ab initio and similarity-based algorithms in order to reduce the high false-positive rate associated with de novo gene prediction (Wang et al. 2003; Misawa and Kikuno 2010) as well as to avoid the propagation of false-positive predicted gene models when closely related species are used as references (Poptsova and Gogarten 2010). A total of 13,657 PCGs were annotated in the *D. buzzatii* genome (Annotation Release 1). These PCG models contain a total of 52,250 exons with an average of 3.8 exons per gene. Gene expression analyses provided transcriptional evidence for 88.4% of these gene models (see below). The number of PCGs annotated in *D. buzzatii* is lower than the number annotated in *D. mojavensis* (14,595, Release 1.3), but quite close to the number annotated in *D. melanogaster* (13,955, Release 5.56), one of the best-known eukaryotic genomes (St Pierre et al. 2014). However PCGs in both *D. buzzatii* and *D. mojavensis* genomes tend to be smaller and contain fewer exons than those in the *D. melanogaster* genome (supplementary table S1, Supplementary Material online), which suggests that the annotation in the two cactophilic species might be incomplete. After applying several quality filters, a total of 12,977 high confidence protein-coding sequences (CDS) were selected for further analysis (see Materials and Methods).

Developmental Transcriptome

To characterize the expression profile throughout *D. buzzatii* development, we performed RNA-Seq experiments using samples from five different stages: Embryos, larvae, pupae, adult females, and adult males. Gene expression levels were calculated based on fragments per kilobase of exon per million fragments mapped (FPKM) values. PCG models that did not show evidence of transcription (FPKM < 1) were classified as nonexpressed PCGs, whereas transcribed regions that did not overlap with any annotated PCG model were tentatively considered noncoding RNA (ncRNA) genes (fig. 2*a*). We detected expression (FPKM > 1) of 26,455 transcripts and 15,026 genes, 12,066 (80%) are PCGs and 2,960 (20%) are

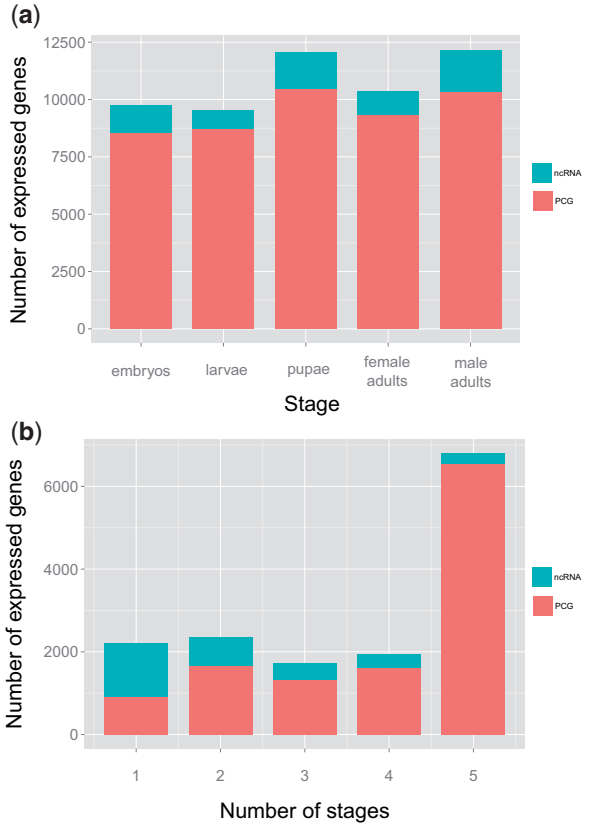

Fig. 2.—Developmental expression profile of *D. buzzatii* genes. (*a*) Number of expressed PCGs (red) and ncRNA genes (blue) along five developmental stages. (*b*) Classification of PCGs and ncRNA genes according to the number of stages where they are expressed.

ncRNA genes. The number of expressed genes (PCGs + ncRNA) increases through the life cycle with a maximum of 12,171 in adult males (fig. 2*a* and supplementary table S2, Supplementary Material online), a pattern similar to that found in *D. melanogaster* (Graveley et al. 2011). In addition, we observed a clear sex-biased expression in adults: Males express 1,824 more genes than females. Previous studies have attributed this sex-biased gene expression mainly to the germ cells, indicating that the differences between ovary and testis are comparable to those between germ and somatic cells (Parisi et al. 2004; Graveley et al. 2011).

We assessed expression breadth for each gene simply as the number of developmental stages with evidence of expression (fig. 2*b* and supplementary table S2, Supplementary Material online). Expression breadth is significantly different ($P < 0.001$) for PCGs and ncRNA genes. A total of 6,546 expressed PCGs (54.2%) are constitutively expressed (i.e., we observed expression in the five stages), but only 260 of ncRNA genes (8.8%) are constitutively expressed (supplementary table S2, Supplementary Material online). In contrast, 925 expressed PCGs (7.7%) and 1,292 ncRNA genes (43.6%) are expressed only in one stage. Mean expression breadth was 3.9 for PCGs and 2.2 for ncRNA genes. Adult males show more stage-specific genes (844 genes) compared with adult females (137 genes).

PCGs with no expression in this study (FPKM < 1) might be expressed at a higher level in other tissues or times, or they might be inducible under specific conditions that we did not test (Weake and Workman 2010; Etges et al. 2015; Matzkin 2014). We also must expect that some remaining fraction of gene models will be false positives (Wang et al. 2003). However, because we used a combination of different annotation methods to reduce the proportion of false-positives, we expect this proportion to be very small. On the other hand, transcribed regions that do not overlap with any annotated PCG models are likely ncRNA genes although we cannot discard that some of them might be false negatives, that is, genes that went undetected by our annotation methods perhaps because they contain small open reading frames (Ladoukakis et al. 2011). One observation supporting that most of them are in fact ncRNA genes is that their expression breadth is quite different from that of PCGs and a high fraction of them are stage-specific genes. In most *Drosophila* species, with limited analyses of the transcriptome (Celniker et al. 2009), few ncRNA genes have been annotated. In contrast, in *D. melanogaster* with a very well-annotated genome, 2,096 ncRNA genes have been found (Release 5.56, FlyBase). Thus, the number of ncRNA found in *D. buzzatii* is comparable to that of *D. melanogaster*.

Website

A website (http://dbuz.uab.cat, last accessed January 7, 2015) has been created to provide free access to all information and resources generated in this work. It includes a customized browser (GBrowse; Stein et al. 2002) for the *D. buzzatii* genome incorporating multiple tracks for gene annotations with different gene predictors, for expression levels and transcript annotations for each developmental stage, and for repeat annotations. It contains also utilities to download contigs, scaffolds, and data files and to carry out Blast searches against all *D. buzzatii* contigs and scaffolds.

Lineage-Specific Analyses

We set up to analyze three lineages for several aspects that could reveal genes involved in adaptation to the cactophilic niche. These lineages are denoted as #1, #2, and #3, respectively, in figure 1: *D. buzzatii* lineage, *D. mojavensis* lineage, and cactophilic lineage (i.e., lineage shared by *D. buzzatii* and *D. mojavensis*). We searched for genes under positive selection, duplicated genes, and orphan genes in those lineages.

Genes under Positive Selection

We first searched for genes evolving under positive selection during the divergence between *D. buzzatii* and *D. mojavensis,*

using codon substitution models implemented in the PAML 4 package (Yang 2007). Two pairs of different SM were compared by the LRT, M1a versus M2a and M7 versus M8 (see Materials and Methods). In each case, a model that allows for sites with $\omega > 1$ (positive selection) is compared with a null model that considers only sites with $\omega < 1$ (purifying selection) and $\omega = 1$ (neutrality). At $P < 0.001$, the first comparison (M1a vs. M2a) detected 915 genes whereas the second comparison (M7 vs. M8) detected 802 genes. Comparison of the two gene sets allowed us to detect 772 genes present in both, and this was taken as the final list of genes putatively under positive selection using SM (supplementary table S3, Supplementary Material online).

Next, we used BSM from PAML 4 package (Yang 2007) to search for genes under positive selection in the phylogeny of the four *Drosophila* subgenus species, *D. buzzatii*, *D. mojavensis*, *D. virilis,* and *D. grimshawi* (fig. 1). Orthologous relationships among the four species were inferred from *D. buzzatii–D. mojavensis* list of orthologs and the OrthoDB catalog (see Materials and Methods). A total of 8,328 unequivocal 1:1:1:1 orthologs were included in the comparison of a BSM allowing sites with $\omega > 1$ (positive selection) and a null model that does not. We selected three branches to test for positive selection (the foreground branches): *D. buzzatii* lineage, *D. mojavensis* lineage, and cactophilic lineage (denoted as #1, #2, and #3 in fig. 1). The number of genes putatively under positive selection detected at $P < 0.001$ in the three branches was 350, 172, and 458, respectively (supplementary table S3, Supplementary Material online). These genes only partially overlap those previously detected in the *D. buzzatii–D. mojavensis* comparison using SM (fig. 3). Although 69.4% and 55.8% of the genes putatively under positive selection in the *D. buzzatii* and *D. mojavensis* lineages were also detected in the *D. buzzatii–D. mojavensis* comparison, only 22.3% of the genes detected in the cactophilic lineage were present in the previous list (fig. 3). Thus, the total number of genes putatively under positive selection is 1,294.

We looked for functional categories overrepresented among the candidate genes reported by both SM and BSM (table 4). We first performed a GO enrichment analysis with the 772 candidate genes uncovered by SM comparing *D. mojavensis* and *D. buzzatii* orthologs using DAVID tools (Huang et al. 2007). Two molecular functions show higher proportion than expected by chance (relative to *D. mojavensis* genome) within the list of candidate genes: Antiporter activity and transcription factor activity. With respect to the biological process, regulation of transcription is the only overrepresented category. A significant enrichment in Src Homology-3 domain was observed. This domain is commonly found within proteins with enzymatic activity and it is associated with protein binding function.

A similar GO enrichment analysis was carried out with candidate genes found using BSM in each of the three targeted branches. The 350 candidate genes in *D. buzzatii* lineage

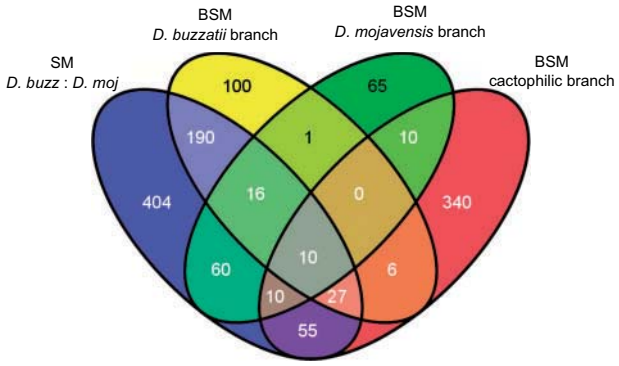

Fig. 3.—Venn diagram showing the number of genes putatively under positive selection detected by two different methods, SM and BSM using three different lineages as foreground branches.

show a significant enrichment in DNA-binding function. DNA-dependent regulation of transcription and phosphate metabolic processes were also overrepresented. We also found a significant enrichment in the Ig-like domain, involved in functions related to cell–cell recognition and immune system. The 172 candidate genes in *D. mojavensis* lineage show a significant excess of genes related to the heterocycle catabolic process ($P = 5.9e\text{-}04$). Interestingly, the main hosts of *D. mojavensis* (columnar cacti) contain large quantities of triterpene glycosides, which are heterocyclic compounds. Among the candidate genes in the branch leading to the two cactophilic species, there are three overrepresented molecular functions related to both metal and DNA binding. The GO terms with the highest significance in the biological process category are cytoskeleton organization and, once again, regulation of transcription.

Using the RNA-Seq data we determined the expression profiles of all 1,294 genes putatively under positive selection. A total of 1,213 (93.7%) of these genes are expressed in at least one developmental stage (supplementary table S2, Supplementary Material online). A comparison of expression level and breadth between candidate and noncandidate genes revealed that genes putatively under positive selection are expressed at a lower level ($X^2 = 84.96$, $P < 2e\text{-}16$) and in fewer developmental stages ($X^2 = 26.99$, $P < 2e\text{-}6$) than the rest.

Orphan Genes in the Cactophilic Lineage

To detect orphan genes in the cactophilic lineage, we blasted the amino acid sequences encoded by 9,114 *D. buzzatii* genes with *D. mojavensis* 1:1 orthologs against all proteins from the 12 *Drosophila* genomes except *D. mojavensis* available in FlyBase (St Pierre et al. 2014). We found 117 proteins with no similarity to any predicted *Drosophila* protein (cutoff value of 1e-05) and were considered to be encoded by putative orphan genes. We focused on the evolutionary dynamics of these orphan genes by studying their properties in comparison

Table 4

GO Analysis of Putative Genes under Positive Selection Detected by Both SM and BSM

Codon substitution Models	Lineage (Branch Number)	Number of Candidates	GO enrichment					
			Molecular Function		Biological Process		Interpro Domain	
			ID	Fold Enrichment	ID	Fold Enrichment	ID	Fold Enrichment
SM	*Drosophila buzzatii* versus *Drosophila mojavensis*	772	Antiporter activity	1.77	Regulation of transcription	4.90	Src homology-3 domain	1.60
			Transcription factor activity	1.56				
BSM	*D. buzzatii* #1	350	DNA binding	1.36	Regulation of transcription DNA dependent	1.36	Immunoglobulin-like	1.33
					Phosphate metabolic process	0.72		
	D. mojavensis #2	172	Dopamine beta-monooxigenase activity	2.35	Heterocycle catabolic process	2.35	DOMON (DOpamine beta-MOnooxygenase N-terminal domain)	2.35
					Cation transport	0.98		
					Histidine family amino acid catabolic process	2.35		
	Cactophilic #3	458	Zinc ion binding	2.01	Cytoeskeleton organization	1.67	Zinc finger, PHD-type	1.93
			Transition metal ion binding	2.01	Regulation of transcription DNA dependent	1.06	Proteinase inhibitor I1 kazal	2.20
			DNA binding	1.66				

NOTE.—Only categories showing an enrichment with a *P* value less than 1.0e-03 are included.

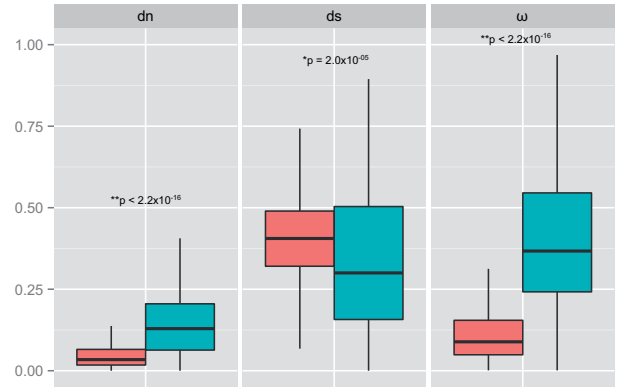

Fig. 4.—Patterns of divergence in orphan and nonorphan genes. Orphan genes (blue) have significantly higher d*n* and ω values compared with that of nonorphan genes (red). Nonorphan genes show significantly higher d*s*.

to the remaining 8,997 1:1 orthologs (fig. 4). We observed that median d*n* of orphan genes was significantly higher than that of nonorphan genes ($dn_{orphan} = 0.1291$; $dn_{nonorphan} = 0.0341$; $W = 846{,}254$, $P < 2.2e\text{-}16$) and the same pattern was observed for ω ($\omega_{orphan} = 0.4253$, $\omega_{nonorphan} = 0.0887$, $W = 951{,}117$, $P < 2.2e\text{-}16$). However, median d*s* of orphan genes is somewhat lower than that for the rest of genes ($ds_{orphan} = 0.3000$, $ds_{nonorphan} = 0.4056$, $W = 406{,}799$, $P = 2.4e\text{-}05$).

We found 19 of the 117 orphan genes in the list of candidate genes detected in the *D. buzzatii*–*D. mojavensis* comparison (see above). This proportion (16.3%) was significantly higher than that found in nonorphan 1:1 orthologs (753/8,997 = 8.4%), which indicates an association between gene lineage-specificity and positive selection (Fisher exact test, two-tailed, $P < 0.0001$). The 19 orphan genes included in the candidate gene group are not associated with any GO category. As a matter of fact, information about protein domains was found for only two of these genes (GYR and YLP motifs in both cases: GI20994 and GI20995). These results should be viewed cautiously as newer genes are functionally undercharacterized and GO databases are biased against them (Zhang et al. 2012). We also compared the protein length between orphan and nonorphan gene products. Our results showed that orphan genes are shorter ($W = 68{,}825.5$, $P < 2.2e\text{-}16$) and have fewer exons than nonlineage-specific genes ($W = 201{,}068$, $P < 2.2e\text{-}16$).

RNA-Seq data allowed us to test for expression of orphan genes. From the 117 gene candidates, 82 (70%) are expressed at least in one of the five analyzed developmental stages. A comparison of the expression profiles between orphan and the rest of 1:1 orthologous genes showed that the expression breadth of orphans is different from that of nonorphans ($X^2 = 101.4$, $P < 0.001$): Most orphan genes are expressed exclusively in one developmental stage with mean expression breadth of 2.56 (vs. 3.94 for nonorphans).

Gene Duplications

The annotated PCGs from four species of the *Drosophila* subgenus were used to study gene family expansions in the *D. buzzatii*, *D. mojavensis*, and cactophilic lineages (fig. 1). Proteins that share 50% identity over 50% of their length were clustered into gene families using Markov Cluster Algorithm. After additional quality filters (see Materials and Methods), the final data set consisted of a total of 56,587 proteins from four species clustered into 19,567 families, including single-gene families (supplementary tables S4–S7, Supplementary Material online).

Considering the *D. buzzatii* genome alone (supplementary table S4, Supplementary Material online), we find 11,251 single-copy genes and 1,851 duplicate genes (14%) clustered in 691 gene families. Among *D. buzzatii* gene families, about 70% of families have two members and the largest family includes 16 members (supplementary table S4, Supplementary Material online). Among single-copy genes, 1,786 genes are only present in the *D. buzzatii* lineage. This number decreases only to 1,624 when proteins are clustered into families with a less stringent cutoff of 35% identity and 50% coverage. Such lineage-specific single-copy genes have been found in all the 12 *Drosophila* genomes that have been analyzed, including *D. mojavensis* (Hahn et al. 2007), and although traditionally they have been viewed as annotation artifacts, many of these genes may be either de novo or fast-evolving genes (Reinhardt et al. 2013; Palmieri et al. 2014).

Lineage-specific expansions were identified by analyzing the gene count for each family from the four species using CAFE3.1 (see Materials and Methods). This analysis detected expansions of 86 families along the *D. buzzatii* lineage. However, 15 families increased in size as the result of extra copies added to the data set after taking into account high sequence coverage. The expansions of these families cannot be confirmed with the current genome assembly. The remaining families were analyzed further in order to confirm *D. buzzatii*-specific duplications. To do that, we first selected gene families with members that have $ds < 0.4$ (median d*s* for *D. mojavensis*–*D. buzzatii* orthologs) and then manually examined syntenic regions in *D. mojavensis* genome. Although this approach might miss some true lineage-specific expansions, it reduces the possibility of including old families into the expansion category that might have been misclassified as a result of incomplete gene annotation in the genomes under study or independent loss of family members in different lineages. Of the 30 gene families whose members had $ds < 0.4$, we confirmed the expansion of 20 families (supplementary table S8, Supplementary Material online). In 12 of the 20 families, new family members are found on the same scaffold in close proximity suggesting unequal crossing over or proximate

segmental duplication as the mechanisms for duplicate formation. The remaining eight families contain dispersed duplicates found in different scaffolds. Six of these families expanded through retroposition, the RNA-mediated duplication mechanism that allows insertion of reverse-transcribed mRNA nearly anywhere in the genome. In most cases, family expansions are due to addition of a new single copy in the *D. buzzatii* lineage (in 25 of total 35 families). Two families that expanded the most, with up to 5 (Family 95) and 9 (Family 126) new members, encode various peptidases involved in protein degradation. Other expanded families are associated with a broad range of functions, including structural proteins of insect cuticle and chorion, enzymes involved in carbohydrate and lipid metabolism, proteins that function in immune response, and olfactory receptors. In addition, Family 128 encodes female reproductive peptidases (Kelleher and Markow 2009) and it appears that new family members have been acquired independently in *D. buzzatii* and *D. mojavensis* lineages (supplementary table S11, Supplementary Material online).

We find six families in *D. buzzatii* that expanded through retroposition in the 11 Myr since the split between *D. buzzatii* and *D. mojavensis* (supplementary table S9, Supplementary Material online). This gives a rate of 0.55 retrogenes/Myr, which is consistent with previous estimates of functional retrogene formation in *Drosophila* of 0.5 retrogenes/Myr (Bai et al. 2007). The expression of all but one retrogene is supported by RNA-Seq data, with no strong biases in expression between the sexes. Four retrogenes are duplicates of ribosomal proteins, and the parental genes from two of these families (*RpL37a* and *RpL30*) have been previously shown to generate retrogenes in other *Drosophila* lineages (Bai et al. 2007; Han and Hahn 2012). Frequent retroposition of ribosomal proteins could be explained by the high levels of transcription of ribosomal genes although other *Drosophila* lineages do not show a bias in favor of retroduplication of ribosomal proteins (Bai et al. 2007; Han and Hahn 2012). The remaining two retrogenes include the duplicate of Caf1, protein that is involved in histone modification, and the duplicate of VhaM9.7-b, a subunit of ATPase complex.

CAFE analysis identified 127 families that expanded along the *D. mojavensis* lineage. Of these families, 86 contain members with $ds < 0.4$. Further examination of syntenic regions confirmed expansion of only 17 families (supplementary table S8, Supplementary Material online). New members in two families (Families 1121 and 1330) are found in different scaffolds and originated through RNA-mediated duplications. These instances have been previously identified as *D. mojavensis*-specific retropositions (Han and Hahn 2012). Members of expanded families encode proteins that function in proteolysis, peptide and ion transport, aldehyde and carbohydrate metabolism, as well as sensory perception (supplementary table S11, Supplementary Material online). At least 4 of the 17 expanded families play a role in reproductive biology: Proteases of Family 128 with three new members have been shown to encode female reproductive peptidases (Kelleher and Markow 2009), and members of three additional families (Families 187, 277, and 1234) encode proteins that are found in *D. mojavensis* accessory gland proteome (Kelleher et al. 2009).

There are 20 gene families that expanded along the cactophilic branch, that is, before the split between *D. buzzatii* and *D. mojavensis* (see Materials and Methods; supplementary table S10, Supplementary Material online). Most families (16 of 20) have expanded through tandem or nearby segmental duplication and are still found within the same scaffold. The remaining families with dispersed duplicates included one retrogene, the duplicate of T-cp1, identified previously in *D. mojavensis* lineage (Han and Hahn 2012). The extent of per-family expansions in the cactophilic lineage is modest, with two new additional members found in four families and a single new copy in the remaining families. Members of the most expanded families encode guanylate cyclases that are involved in intracellular signal transduction, peptidases, and carbon–nitrogen hydrolases. Members of other families include various proteins with metal-binding properties as well as proteins with a role in vesicle and transmembrane transport (supplementary table S11, Supplementary Material online). We also see expansion of three families (Family 775, Family 776, and Family 800) with functions related to regulation of juvenile hormone (JH) levels (see Discussion).

Discussion

The *D. buzzatii* Genome

Drosophila is a leading model for comparative genomics, with 24 genomes of different species already sequenced (Adams et al. 2000; *Drosophila* 12 Genomes Consortium et al. 2007; Zhou et al. 2012; Zhou and Bachtrog 2012; Fonseca et al. 2013; Ometto et al. 2013; Chen et al. 2014). However, only five of these species belong to the species-rich *Drosophila* subgenus, and only one of these species, *D. mojavensis*, is a cactophilic species from the large *repleta* species group. Here we sequenced the genome and transcriptome of *D. buzzatii*, another cactophilic member of the *repleta* group, to investigate the genomic basis of adaptation to this distinct ecological niche. Using different sequencing platforms and a three-stage de novo assembly strategy, we generated a high quality genome sequence that consists of 826 scaffolds greater than 3 kb (Freeze 1). A large portion (>90%) of the genome is represented by 158 scaffolds with a minimum size of 160 kb that have been assigned, ordered, and oriented in the six chromosomes of the *D. buzzatii* karyotype. As expected, the assembly is best for chromosome 2 (because of the use of Sanger generated BAC-end sequences) and worst for chromosome X (because of the three-fourth representation of this chromosome in adults of both sexes). The quality of our Freeze 1 assembly compares favorably with the

assembly generated using only Illumina reads and the SOAPdenovo assembler, and with those of other *Drosophila* genomes generated using second-generation sequencing platforms (Zhou et al. 2012; Zhou and Bachtrog 2012; Fonseca et al. 2013; Ometto et al. 2013; Chen et al. 2014), although our Freeze 1 does not attain the quality of the 12 *Drosophila* genomes generated using Sanger only (*Drosophila* 12 Genomes Consortium et al. 2007).

Drosophila buzzatii is a subcosmopolitan species that has been able to colonize four of the six major biogeographical regions (David and Tsacas 1980). Only two other *repleta* group species (*Drosophila repleta* and *Drosophila hydei*) have reached such widespread distribution. Invasive species are likely to share special genetic traits that enhance their colonizing ability (Parsons 1983; Lee 2002). From an ecological point of view we would expect colonizing species to be r-strategists with a short developmental time (Lewontin 1965). Because there is a correlation between developmental time and genome size (Gregory and Johnston 2008), colonizing species are also expected to have a small genome size (Lavergne et al. 2010). The genome size of *D. buzzatii* was estimated in our assembly as 161 Mb and by cytological techniques as 153 Mb, approximately 20% smaller than the *D. mojavensis* genome. The genome size of a second *D. buzzatii* strain, estimated by cytological techniques, is even smaller, 146 Mb. However, the relationship between genome size and colonizing ability does not hold in the *Drosophila* genus at large. Although colonizing species such as *D. melanogaster* and *Drosophila simulans* have relatively small genomes, specialist species with a narrow distribution such as *Drosophila sechelia* and *Drosophila erecta* also have small genomes. On the other hand, *Drosophila ananassae, Drosophila malerkotliana, Drosophila suzuki, D. virilis*, and *Zaprionus indianus* are also colonizing *Drosophila* species but have relatively large genomes (Nardon et al. 2005; Bosco et al. 2007; *Drosophila* 12 Genomes Consortium et al. 2007; Gregory and Johnston 2008). Further, there seems to be little difference in genome size between original and colonized populations within species (Nardon et al. 2005). Seemingly, other factors such as historical or chance events, niche dispersion, genetic variability, or behavioral shifts are more significant than genome size in determining the current distribution of colonizing species (Markow and O'Grady 2008).

TE content in the *D. buzzatii* genome was estimated as 8.4% (table 2), a relatively low value compared with that of *D. mojavensis*, 10–14% (Ometto et al. 2013; Rius et al., in preparation). These data agree well with the smaller genome size of *D. buzzatii* because genome size is positively correlated with the contribution of TEs (Kidwell 2002; Feschotte and Pritham 2007). However, TE copy number and coverage estimated in *D. buzzatii* (table 2) must be taken cautiously. Coverage is surely underestimated due to the difficulties in assembling repeats, in particular with short sequence reads, whereas the number of copies may be overestimated due to copy fragmentation (Rius N, Guillén Y, Kapusta A, Feschotte C, Ruiz A, in preparation). The contribution of satDNAs (table 3) is also an underestimate and further experiments are required for a correct assessment of this component (de Lima LG, Svartman M, Ruiz A, Kuhn GCS, in preparation). However, we identified the pBuM189 satDNA as the most abundant tandem repeat of *D. buzzatii*. Previous in situ hybridization experiments revealed that pBuM189 copies are located in the centromeric region of all chromosomes, except chromosome X (Kuhn et al. 2008). Thus, pBuM189 satellite is likely the main component of the *D. buzzatii* centromere. Interestingly, a pBuM189 homologous sequence has recently been identified as the most abundant tandem repeat of *D. mojavensis* (Melters et al. 2013). Although the chromosome location in *D. mojavensis* has not been determined, the persistence of pBuM189 as the major satDNA in *D. buzzatii* and *D. mojavensis* may reflect a possible role for these sequences in centromere function (Ugarković 2009).

Chromosome Evolution

The chromosomal evolution of *D. buzzatii* and *D. mojavensis* has been previously studied by comparing the banding pattern of the salivary gland chromosomes (Ruiz et al. 1990; Ruiz and Wasserman 1993). *Drosophila buzzatii* has few fixed inversions (*2m, 2n, 2z*[7], and *5g*) when compared with the ancestor of the *repleta* group. In contrast, *D. mojavensis* showed ten fixed inversions (*Xe, 2c, 2f, 2g, 2h, 2q, 2r, 2s, 3a,* and *3d*), five of them (*Xe, 2q, 2r, 2s,* and *3d*) exclusive to *D. mojavensis* and the rest shared with other cactophilic *Drosophila* (Guillén and Ruiz 2012). Thus, the *D. mojavensis* lineage appears to be a derived lineage with a relatively high rate of rearrangement fixation. Here, we compared the organization of both genomes corroborating all known inversions in chromosomes X, 2, 4, and 5. In *D. mojavensis* chromosome 3, however, we found five inversions instead of the two expected (Delprat A, Guillén Y, Ruiz A, in preparation). One of the three additional inversions is the polymorphic inversion *3f*[2] (Ruiz et al. 1990). This inversion has previously been found segregating in Baja California and Sonora (Mexico) and is homozygous in the strain of Santa Catalina Island (California) that was used to generate the *D. mojavensis* genome sequence (*Drosophila* 12 Genomes Consortium et al. 2007). Previously, the Santa Catalina Island population was thought to have the standard (ancestral) arrangements in all chromosomes, like the populations in Southern California and Arizona (Ruiz et al. 1990; Etges et al. 1999). The presence of inversion *3f*[2] in Santa Catalina Island is remarkable because it indicates that the flies that colonized this island came from Baja California and are derived instead of ancestral with regard to the rest of *D. mojavensis* populations (Delprat et al. 2014). The other two additional chromosome 3 inversions are fixed in the *D. mojavensis* lineage and emphasize its rapid chromosomal evolution. Guillén and Ruiz (2012) analyzed the breakpoint of all chromosome 2

inversions fixed in *D. mojavensis* and concluded that the numerous gene alterations at the breakpoints with putative adaptive consequences point directly to natural selection as the cause of *D. mojavensis* rapid chromosomal evolution. The four fixed chromosome 3 inversions provide an opportunity for further testing this hypothesis (Delprat A, Guillén Y, Ruiz A, in preparation).

Candidate Genes under Positive Selection and Orphan Genes

Several methods have been developed to carry out genome-wide scans for genes evolving under positive selection (Nielsen 2005; Anisimova and Liberles 2007; Vitti et al. 2013). We used here a rather simple approach based on the comparison of the nonsynonymous substitution rate (d*n*) with the synonymous substitution rate (d*s*) at the codon level (Yang et al. 2000; Wong et al. 2004; Zhang et al. 2005; Yang 2007). Genes putatively under positive selection were detected on the basis of statistical evidence for a subset of codons where replacement mutations were fixed faster than mutation at silent sites. Four species of the *Drosophila* subgenus (fig. 1) were employed to search for genes under positive selection using SM and BSM. We restricted the analysis to this subset of the *Drosophila* phylogeny to avoid the saturation of synonymous substitutions expected with phylogenetically very distant species (Bergman et al. 2002; Larracuente et al. 2008), and also because these are the genomes with the highest quality available (Schneider et al. 2009). A total of 1,294 candidate genes were detected with both SM and BSM, which represents approximately 14% of the total set of 1:1 orthologs between *D. mojavensis* and *D. buzzatii*. Positive selection seems pervasive in *Drosophila* (Sawyer et al. 2007; Singh et al. 2009; Sella et al. 2009; Mackay et al. 2012) and, using methods similar to ours, it has been estimated that 33% of single-copy orthologs in the *melanogaster* group have experienced positive selection (*Drosophila* 12 Genomes Consortium et al. 2007). The smaller fraction of genes putatively under positive selection in our analyses may be due to the fewer lineages considered in our study. In addition, both studies may be underestimating the true proportion of positively selected genes because only 1:1 orthologs were included in the analyses and genes that evolve too fast may be missed by the methods used to establish orthology relationships (Bierne and Eyre-Walker 2004). At any rate, the 1,294 candidate genes found here should be evaluated using other genomic methods for detecting positive selection, for example, those comparing levels of divergence and polymorphism (Vitti et al. 2013). Furthermore, functional follow-up tests will be necessary for a full validation of their adaptive significance (Lang et al. 2012).

BSM allowed us to search for positively selected genes in the three-targeted lineages (*D. buzzatii*, *D. mojavensis,* and cactophilic branch). We then performed GO enrichment analyses in order to identify potential candidates for environmental adaptation given the ecological properties of both cactophilic species (table 4). The most interesting result of this analysis is that genes putatively under positive selection in *D. mojavensis* branch are enriched in genes involved in heterocyclic catabolic processes. Four candidate *D. mojavensis* genes, *GI19101, GI20678, GI21543* and *GI22389,* that are orthologous to *D. melanogaster* genes *nahoda*, *CG5235*, *slgA* and *knk*, respectively, participate in these processes and might be involved in adaptation of *D. mojavensis* to the *Stenocereus cacti*, plants with particularly large quantities of heterocyclic compounds (see Introduction). A difficulty with this interpretation is the fact that the *D. mojavensis* genome sequence was generated using a strain from Santa Catalina Island where *D. mojavensis* inhabits Opuntia cactus (*Drosophila* 12 Genomes Consortium et al. 2007). However, the evidence indicates that the ancestral *D. mojavensis* population is the agria-inhabiting Baja California population and that the Mainland Sonora population split from Baja California approximately 0.25 Ma whereas the Mojave Desert and Mainland Sonora populations diverged more recently, approximately 0.125 Ma (Smith et al. 2012). Moreover, the presence of inversion $3f^2$ in the Santa Catalina Island population suggests that the flies that colonized this island came from Baja California populations, where this inversion is currently segregating, and not from the Mojave Desert, where this inversion is not present (Delprat et al. 2014). This is compatible with mitochondrial DNA sequence data (Reed et al. 2007) although in contrast to other data (Machado et al. 2007). Finally, the transcriptional profiles of the four *D. mojavensis* subpopulations reveal only minor gene expression differences between individuals from Santa Catalina Island and Baja California (Matzkin and Markow 2013).

Orphan genes are genes with restricted taxonomic distribution. Such genes have been suggested to play an important role in phenotypic and adaptive evolution in multiple species (Domazet-Lošo and Tautz 2003; Khalturin et al. 2009; Chen et al. 2013). The detection of orphan genes is highly dependent on the availability of sequenced and well-annotated genomes of closely related species, and the total number of lineage-specific genes tend to be overestimated (Khalturin et al. 2009). We were as conservative as possible by considering only high-confidence 1:1 orthologs in two species, *D. buzzatii* and *D. mojavensis*. The result is a set of 117 orphans in the cactophilic lineage.

We observe that orphan genes clearly show a different pattern of molecular evolution compared with that of older conserved genes. Orphans exhibit a higher d*n* that can be attributed to more beneficial mutations fixed by positive selection or to lower constraint, or both (Cai and Petrov 2010; Chen et al. 2010). However, as the number of genes putatively under positive selection within the set of orphan genes is higher than expected by chance, we suggest that the elevated d*n* likely reflects adaptive evolution.

Orphans also have fewer exons and encode shorter proteins than nonorphans. This observation has been reported in

multiple eukaryotic organisms such as yeasts (Carvunis et al. 2012), fruitflies (Domazet-Lošo and Tautz 2003) and primates (Cai and Petrov 2010), and it is further supported by a positive correlation between protein length and sequence conservation (Lipman et al. 2002) (see above). We did not find expression support for all the orphan genes detected. This suggests to us that either orphans are more tissue- or stage-specific than nonorphans (Zhang et al. 2012) or we are actually detecting artifactual CDS that are not expressed. However, given the patterns of sequence evolution of orphan genes, we favor the first explanation for the majority of them. Collectively, all these results support the conclusion that orphan genes evolve faster than older genes, and that they experience lower levels of purifying selection and higher rates of adaptive evolution (Chen et al. 2010).

It has been widely reported that younger genes have lower expression levels than older genes on average (Cai and Petrov 2010; Tautz and Domazet-Lošo 2011; Zhang et al. 2012). Here, we observe that orphan genes that are being transcribed are less expressed than nonorphans (Kruskal test, $X^2 = 9.37$, $P = 0.002$). One of the proposed hypotheses to explain these observations is that genes that are more conserved are indeed involved in more functions (Pál et al. 2006; Tautz and Domazet-Lošo 2011).

Different studies have demonstrated that newer genes are more likely to have stage-specific expression than older genes (Zhang et al. 2012). Here, we show that the number of stage-specific expressed orphans is significantly higher than that of older genes. It has been proposed that newer genes tend to be more developmentally regulated than older genes (Tautz and Domazet-Lošo 2011). This means that they contribute most to the ontogenic differentiation between taxa (Chen et al. 2010). In *D. buzzatii* the vast majority of stage-specific orphan genes are expressed in larvae (15/29), indicating that expression of younger genes is mostly related to stages in which *D. buzzatii* and *D. mojavensis* lineages most diverge from each other.

Gene Duplication

The study of gene duplications in the *D. buzzatii* and *D. mojavensis* lineages aims at understanding the genetic bases of the ecological specialization associated with colonization of novel cactus habitats. Although we only considered expanded families, it is known that specialization sometimes involves gene losses. For example, *D. sechellia* and *D. erecta,* which are specialized to grow on particular substrates, have lost gustatory receptors and detoxification genes (*Drosophila* 12 Genomes Consortium et al. 2007; Dworkin and Jones 2009). Sometimes the losses are driven by positive selection, as has been suggested in the case of the *neverland* gene in *Drosophila pachea* (Lang et al. 2012) where positive selection appears to have favored a novel *neverland* allele that has lost the ability to metabolize cholesterol. In our study of gene families, the incompleteness of the annotation of *D. buzzatii* PCGs precludes us from being able to reliably identify gene families that lost family members.

To minimize the possibility of missing gene copies that were potentially collapsed into single genes during *D. buzzatii* genome assembly, we used sequence coverage to adjust the size of gene families. Two of the families that expanded as a result of this correction encoded chorion genes. However, chorion genes are known to undergo somatic amplifications in ovarian follicle cells (Claycomb and Orr-Weaver 2005), and the use of sequence coverage to correct for "missing" copies can be misleading in these cases. As there is no easy way to verify families that were placed into the expanded category due to high sequence coverage alone, our discussion below is limited to gene duplicates that were annotated in the *D. buzzatii* genome.

A recent survey of the functional roles of new genes across various taxa offers evidence for the rapid recruitment of new genes into gene networks underlying a wide range of phenotypes including reproduction, behavior, and development (Chen et al. 2013). A number of lineage-specific duplicates identified in our study fit this description, but further experimental confirmation of their functions through loss-of-function studies and characterization of molecular interactions are necessary. Among families that expanded in the *D. buzzatii*, the *D. mojavensis*, and the cactophilic lineages, 35% have functional annotations that are similar to those of rapidly evolving families identified in the analysis of the 12 *Drosophila* genomes (Hahn et al. 2007). These families include genes that are involved in proteolysis, zinc ion binding, chitin binding, sensory perception, immunity, and reproduction. A fraction of these expanded families may reflect physiological adaptations to a novel habitat. For example, given the importance of olfactory perception in recognition of the host cactus plants (Date et al. 2013), the duplication of an olfactory receptor in *D. buzzatii* may represent an adaptation to cactophilic substrates. Another *D. buzzatii* family includes *ninjurin*, a gene involved in tissue regeneration that is one of the components of the innate immune response (Boutros et al. 2002). In *D. mojavensis*, we also observe the duplication of an odorant receptor and, coinciding with a previous report (Croset et al. 2010), of an ionotropic glutamate receptor that belongs to a novel family of diversified chemosensory receptors (Benton et al. 2009; Croset et al. 2010). An aldehyde dehydrogenase is also duplicated in the *D. mojavensis* lineage and might reveal a role in detoxification of particular aldehydes and ethanol (Fry and Saweikis 2006). In the *D. buzzatii–D. mojavensis* lineage, one family contains proteins with the MD-2-related lipid recognition domain involved in pathogen recognition and in *D. mojavensis* we find a duplicate of a phagosome-associated peptide transporter that is involved in bacterial response in *D. melanogaster* (Charrière et al. 2010).

Several of the *D. mojavensis*-specific gene duplicates have been described as male and female reproductive proteins.

Unlike the accessory gland proteins of *D. melanogaster*, the proteome of *D. mojavensis* accessory glands is rich in metabolic enzymes and nutrient transport proteins (Kelleher et al. 2009). Three of the expanded families include metabolic proteins previously identified as candidate seminal fluid proteins specific to *D. mojavensis* lineage (Kelleher et al. 2009). We also detect an increase of female reproductive tract proteases as a possible counter adaptation to fast-evolving male ejaculate (Kelleher and Markow 2009).

Three gene families are of particular interest among those that were expanded in the lineages leading to *D. buzzatii* and *D. mojavensis*, as they contain duplicates of genes with functions related to the regulation of JH levels. One family includes a new duplicate of JH esterase duplication gene (*Jhedup* in *D. melanogaster*). JH esterases are involved in JH degradation (Bloch et al. 2013), although *Jhedup* has much lower level of JH esterase activity than *Jhe* (Crone et al. 2007). Another family includes new duplicate that encodes protein with sequence similarity to hemolymph JH-binding protein (*CG5945* in *D. melanogaster*). JH-binding proteins belong to a large gene family regulated by circadian genes and affect circadian behavior, courtship behavior, metabolism, and aging (Vanaphan et al. 2012). This family includes JH-binding proteins that function as carriers of JH through the hemolymph to its target tissues (Bloch et al. 2013). The third family includes a new duplicate of a dopamine synthase gene (*ebony* in *D. melanogaster*). *ebony* is involved in the synthesis of dopamine, and it is known that dopamine levels affect behavior and circadian rhythms through regulation of hormone levels including JH (Rauschenbach et al. 2012). All three duplicates are expressed in *D. buzzatii* adults. At insect adult stage, JHs play a role in physiology and behavior, and their levels oscillate daily (Bloch et al. 2013). Gene duplications of JH-binding proteins (JHBP), JH esterases (JHE), and *ebony* may change the timing and levels of active JHs which, in turn, alter the behavior and physiology regulated by JHs. One interesting effect of mutations in circadian rhythm genes, or of direct perturbations of the circadian rhythm, is a reduced ethanol tolerance in *D. melanogaster* (Pohl et al. 2013). Intriguingly, *Jhedup* and another gene duplicated in the cactophilic lineage, *Sirt2* (a protein deacetylase), have been also shown to affect ethanol tolerance and sensitivity when mutated (Kong et al. 2010). Given that both *D. mojavensis* and *D. buzzatii* breed and feed on rotting fruit, a shift in tolerance to ethanol and other cactus-specific compounds is one of the expected adaptations associated with a switch to a cactus host. Future functional studies of these new duplicates are required to understand their role in physiological and behavioral changes associated with a change to a new habitat.

Supplementary Material

Supplementary methods, figures S1–S6, and tables S1–S20 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

This work was supported by grants BFU2008-04988 and BFU2011-30476 from Ministerio de Ciencia e Innovación (Spain) to A.R., by an FPI fellowship to Y.G. and a PIF-UAB fellowship to N.R, and by the National Institute of General Medical Sciences of the National Institute of Health under award number R01GM071813 to E.B. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

Literature Cited

Adams MD, et al. 2000. The genome sequence of *Drosophila melanogaster*. Science 287:2185–2195.
Anisimova M, Liberles DA. 2007. The quest for natural selection in the age of comparative genomics. Heredity 99:567–579.
Bai Y, Casola C, Feschotte C, Betrán E. 2007. Comparative genomics reveals a constant rate of origination and convergent acquisition of functional retrogenes in *Drosophila*. Genome Biol. 8:R11.
Barker JSF, Starmer WT. 1982. Ecological genetics and evolution: the cactus-yeast-*Drosophila* model system. Sydney (NSW): Academic Press.
Barker JSF, Starmer WT, MacIntyre RJ. 1990. Ecological and evolutionary genetics of *Drosophila*. New York: Plenum Press.
Begon M. 1982. Yeasts and *Drosophila*. In: Ashburner M, Carson HL, Thompson JN Jr, editors. The genetics and biology of *Drosophila*, Vol. 3b. London: Academic Press. p. 3345–3384.
Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27:573–580.
Benton R, Vannice KS, Gomez-Diaz C, Vosshall LB. 2009. Variant ionotropic glutamate receptors as chemosensory receptors in *Drosophila*. Cell 136:149–162.
Bergman CM, et al. 2002. Assessing the impact of comparative genomic sequence data on the functional annotation of the *Drosophila* genome. Genome Biol. 3:research0086.
Betran 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.
Bierne N, Eyre-Walker A. 2004. The genomic rate of adaptive amino acid substitution in *Drosophila*. Mol Biol Evol. 21:1350–1360.
Bloch G, Hazan E, Rafaeli A. 2013. Circadian rhythms and endocrine functions in adult insects. J Insect Physiol. 59:56–69.
Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. 2011. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics 27:578–579.
Bosco G, Campbell P, Leiva-Neto JT, Markow TA. 2007. Analysis of *Drosophila* species genome size and satellite DNA content reveals significant differences among strains as well as between species. Genetics 177:1277–1290.
Boutros M, Agaisse H, Perrimon N. 2002. Sequential activation of signaling pathways during innate immune responses in *Drosophila*. Dev Cell. 3: 711–722.
Breitmeyer CM, Markow TA. 1998. Resource availability and population size in cactophilic *Drosophila*. Funct Ecol. 12:14–21.
Cai JJ, Petrov DA. 2010. Relaxed purifying selection and possibly high rate of adaptation in primate lineage-specific genes. Genome Biol Evol. 2: 393–409.
Calvete O, González J, Betrán E, Ruiz A. 2012. Segmental duplication, microinversion, and gene loss associated with a complex inversion breakpoint region in *Drosophila*. Mol Biol Evol. 29:1875–1889.
Carson HL 1971. The ecology of *Drosophila* breeding sites. No. 2. Honolulu (Hawaii): University of Hawaii Foundation Lyon Arboretum Fund.

Carson HL, Wasserman M. 1965. A widespread chromosomal polymorphism in a widespread species, *Drosophila buzzatii*. Am Nat. 99: 111–115.
Carvunis A-R, et al. 2012. Proto-genes and *de novo* gene birth. Nature 487:370–374.
Celniker SE, et al. 2009. Unlocking the secrets of the genome. Nature 459: 927–930.
Charrière GM, et al. 2010. Identification of *Drosophila* Yin and PEPT2 as evolutionarily conserved phagosome-associated muramyl dipeptide transporters. J Biol Chem. 285:20147–20154.
Chen S, Krinsky BH, Long M. 2013. New genes as drivers of phenotypic evolution. Nat Rev Genet. 14:645–660.
Chen S, Zhang YE, Long M. 2010. New genes in *Drosophila* quickly become essential. Science 330:1682–1685.
Chen ZX, et al. 2014. Comparative validation of the *D. melanogaster* modENCODE transcriptome annotation. Genome Res. 24: 1209–1223.
Claycomb JM, Orr-Weaver TL. 2005. Developmental gene amplification: insights into DNA replication and gene expression. Trends Genet. 21: 149–162.
Crone EJ, et al. 2007. Only one esterase of *Drosophila melanogaster* is likely to degrade juvenile hormone in vivo. Insect Biochem Mol Biol. 37: 540–549.
Croset V, et al. 2010. Ancient protostome origin of chemosensory ionotropic glutamate receptors and the evolution of insect taste and olfaction. PLoS Genet. 6:e1001064.
Date P, et al. 2013. Divergence in olfactory host plant preference in *D. mojavensis* in response to cactus host use. PLoS One 8:e70027.
David J, Tsacas L. 1980. Cosmopolitan, subcosmopolitan and widespread species: different strategies within the Drosophilid family (Diptera). C R Soc Biogéogr. 57:11–26.
Delcher AL, Salzberg SL, Phillippy AM. 2003. Using MUMmer to identify similar regions in large sequence sets. Curr Protoc Bioinformatics. Chapter 10:Unit 10.3.
Delprat A, Etges WJ, Ruiz A. 2014. Reanalysis of polytene chromosomes in *Drosophila mojavensis* populations from Santa Catalina Island, California, USA. Drosoph Inf Serv. 97:53–57.
Domazet-Lošo T, Tautz D. 2003. An evolutionary analysis of orphan genes in *Drosophila*. Genome Res. 13:2213–2219.
Drosophila 12 Genomes Consortium, et al. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. Nature 450:203–218.
Dworkin I, Jones CD. 2009. Genetic changes accompanying the evolution of host specialization in *Drosophila sechellia*. Genetics 181: 721–736.
Enright AJ, Van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 30: 1575–1584.
Etges WJ, et al. 2015. Deciphering life history transcriptomes in different environments. Mol Ecol. 24:151–179.
Etges WJ, Johnson WR, Duncan GA, Huckins G, Heed WB. 1999. Ecological genetics of cactophilic *Drosophila*. In: Robichaux R, editor. Ecology of Sonoran Desert plants and plant communities. Tucson (AZ): University of Arizona Press. p. 164–214.
Fellows DP, Heed WB. 1972. Factors affecting host plant selection in desert-adapted cactiphilic *Drosophila*. Ecology 53:850–858.
Feschotte C, Pritham EJ. 2007. DNA transposons and the evolution of eukaryotic genomes. Annu Rev Genet. 41:331–368.
Fogleman JC, Armstrong L. 1989. Ecological aspects of cactus triterpene glycosides I. Their effect on fitness components of *Drosophila mojavensis*. J Chem Ecol. 15:663–676.
Fogleman JC, Danielson PB. 2001. Chemical interactions in the cactus-microorganism-*Drosophila* model system of the Sonoran Desert. Am Zool. 41:877–889.
Fogleman JC, Kircher HW. 1986. Differential effects of fatty acid chain length on the viability of two species of cactophilic *Drosophila*. Comp Biochem Physiol A Physiol. 83:761–764.
Fonseca NA, et al. 2013. *Drosophila americana* as a model species for comparative studies on the molecular basis of phenotypic variation. Genome Biol Evol. 5:661–679.
Fontdevila A, Ruiz A, Alonso G, Ocaña J. 1981. Evolutionary history of *Drosophila buzzatii*. I. Natural chromosomal polymorphism in colonized populations of the Old World. Evolution 35:148–157.
Frank MR, Fogleman JC. 1992. Involvement of cytochrome P450 in host-plant utilization by Sonoran Desert *Drosophila*. Proc Natl Acad Sci U S A. 89:11998–12002.
Fry JD, Saweikis M. 2006. Aldehyde dehydrogenase is essential for both adult and larval ethanol resistance in *Drosophila melanogaster*. Genet Res. 87:87–92.
Garcia-Fernàndez J. 2005. The genesis and evolution of homeobox gene clusters. Nat Rev Genet. 6:881–892.
González J, et al. 2005. A BAC-based physical map of the *Drosophila buzzatii* genome. Genome Res. 15:885–889.
Graveley BR, et al. 2011. The developmental transcriptome of *Drosophila melanogaster*. Nature 471:473–479.
Gregory TR, Johnston JS. 2008. Genome size diversity in the family Drosophilidae. Heredity 101:228–238.
Guillén Y, Ruiz A. 2012. Gene alterations at *Drosophila* inversion breakpoints provide *prima facie* evidence for natural selection as an explanation for rapid chromosomal evolution. BMC Genomics 13:53.
Haas BJ, et al. 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol. 9:R7.
Hahn MW, Han MV, Han S-G. 2007. Gene family evolution across 12 *Drosophila* genomes. PLoS Genet. 3:e197.
Han MV, Hahn MW. 2012. Inferring the history of interchromosomal gene transposition in *Drosophila* using n-dimensional parsimony. Genetics 190:813–825.
Han MV, Thomas GWC, Lugo-Martinez J, Hahn MW. 2013. Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. Mol Biol Evol. 30:1987–1997.
Hasson E, et al. 1995. The evolutionary history of *Drosophila buzzatii*. XXVI. Macrogeographic patterns of inversion polymorphism in New World populations. J Evol Biol. 8:369–384.
Hasson E, Naveira H, Fontdevila A. 1992. The breeding sites of Argentinian cactophilic species of the *Drosophila mulleri* complex (subgenus *Drosophila-repleta* group). Rev Chil Hist Nat. 65:319–326.
Havens JA, Etges WJ. 2013. Premating isolation is determined by larval rearing substrates in cactophilic *Drosophila mojavensis*. IX. Host plant and population specific epicuticular hydrocarbon expression influences mate choice and sexual selection. J Evol Biol. 26:562–576.
Heed WB, Mangan RL. 1986. Community ecology of the Sonoran Desert *Drosophila*. In: Ashburner M, Carson HL, Thompson JN, editors. The genetics and biology of *Drosophila*. Vol. 3e. London: Academic Press. p. 311–345.
Hoffmann AA, Sørensen JG, Loeschcke V. 2003. Adaptation of *Drosophila* to temperature extremes: bringing together quantitative and molecular approaches. J Therm Biol. 28:175–216.
Huang DW, et al. 2007. DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. Nucleic Acids Res. 35:W169–W175.
Huang DW, Sherman BT, Lempicki RA. 2009a. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 37:1–13.
Huang DW, Sherman BT, Lempicki RA. 2009b. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 4:44–57.

Hughes CL, Kaufman TC. 2002. Hox genes and the evolution of the arthropod body plan. Evol Dev. 4:459–499.
Irimia M, Maeso I, Garcia-Fernàndez J. 2008. Convergent evolution of clustering of Iroquois homeobox genes across metazoans. Mol Biol Evol. 25:1521–1525.
Jurka J, et al. 2005. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res. 110:462–467.
Kelleher ES, Markow TA. 2009. Duplication, selection and gene conversion in a *Drosophila mojavensis* female reproductive protein family. Genetics 181:1451–1465.
Kelleher ES, Watts TD, LaFlamme BA, Haynes PA, Markow TA. 2009. Proteomic analysis of *Drosophila mojavensis* male accessory glands suggests novel classes of seminal fluid proteins. Insect Biochem Mol Biol. 39:366–371.
Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TCG. 2009. More than just orphans: are taxonomically-restricted genes important in evolution? Trends Genet. 25:404–413.
Kidwell MG. 2002. Transposable elements and the evolution of genome size in eukaryotes. Genetica 115:49–63.
Kircher HW. 1982. Chemical composition of cacti and its relationship to Sonoran Desert *Drosophila*. In: Barker JSF, Starmer WT, editors. Ecological genetics and evolution: the cactus-yeast-*Drosophila* model system. Sydney (NSW): Academic Press. p. 143–158.
Kong EC, et al. 2010. Ethanol-regulated genes that contribute to ethanol sensitivity and rapid tolerance in *Drosophila*. Alcohol Clin Exp Res. 34: 302–316.
Korf I. 2004. Gene finding in novel genomes. BMC Bioinformatics 5:59.
Korf I, Flicek P, Duan D, Brent MR. 2001. Integrating genomic homology into gene structure prediction. Bioinformatics 17(Suppl 1): S140–S148.
Kriventseva EV, Rahman N, Espinosa O, Zdobnov EM. 2008. OrthoDB: the hierarchical catalog of eukaryotic orthologs. Nucleic Acids Res. 36: D271–D275.
Kuhn GCS, Sene FM, Moreira-Filho O, Schwarzacher T, Heslop-Harrison JS. 2008. Sequence analysis, chromosomal distribution and long-range organization show that rapid turnover of new and old pBuM satellite DNA repeats leads to different patterns of variation in seven species of the *Drosophila buzzatii* cluster. Chromosome Res. 16: 307–324.
Ladoukakis E, Pereira V, Magny EG, Eyre-Walker A, Couso JP. 2011. Hundreds of putatively functional small open reading frames in *Drosophila*. Genome Biol. 12:R118.
Lai EC, Bodner R, Posakony JW. 2000. The enhancer of split complex of *Drosophila* includes four Notch-regulated members of the bearded gene family. Development 127:3441–3455.
Lang M, et al. 2012. Mutations in the *neverland* gene turned *Drosophila pachea* into an obligate specialist species. Science 337:1658–1661.
Lang M, et al. 2014. Radiation of the *Drosophila nannoptera* species group in Mexico. J Evol Biol. 27:575–584.
Larracuente AM, et al. 2008. Evolution of protein-coding genes in *Drosophila*. Trends Genet. 24:114–123.
Lavergne S, Muenke NJ, Molofsky J. 2010. Genome size reduction can trigger rapid phenotypic evolution in invasive plants. Ann Bot. 105: 109–116.
Lee CE. 2002. Evolutionary genetics of invasive species. Trends Ecol Evol. 17:386–391.
Lewontin RC. 1965. Selection for colonizing ability. In: Baker HG, Stebbins GL, editors. The genetics of colonizing species. New York: Academic Press. p. 77–94.
Lipman DJ, Souvorov A, Koonin EV, Panchenko AR, Tatusova TA. 2002. The relationship of protein conservation and sequence length. BMC Evol Biol. 2:20.
Loeschcke V, Krebs RA, Dahlgaard J, Michalak P. 1997. High-temperature stress and the evolution of thermal resistance in *Drosophila*. EXS 83: 175–190.
Luo R, et al. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. GigaScience 1:18.
Machado CA, Matzkin LM, Reed LK, Markow TA. 2007. Multilocus nuclear sequences reveal intra- and interspecific relationships among chromosomally polymorphic species of cactophilic *Drosophila*. Mol Ecol. 16:3009–3024.
Mackay TFC, et al. 2012. The *Drosophila melanogaster* genetic reference panel. Nature 482:173–178.
Manfrin MH, Sene FM. 2006. Cactophilic *Drosophila* in South America: a model for evolutionary studies. Genetica 126:57–75.
Margulies M, et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–380.
Markow TA, O'Grady P. 2008. Reproductive ecology of *Drosophila*. Funct Ecol. 22:747–759.
Matzkin LM. 2014. Ecological genomics of host shifts in *Drosophila mojavensis*. Adv Exp Med Biol. 781:233–247.
Matzkin LM, Markow TA. 2013. Transcriptional differentiation across the four subspecies of *Drosophila mojavensis*. In: Pawel M, editor. Speciation: natural processes, genetics and biodiversity. New York: Nova Scientific Publishers.
Matzkin LM, Watts TD, Bitler BG, Machado CA, Markow TA. 2006. Functional genomics of cactus host shifts in *Drosophila mojavensis*. Mol Ecol. 15:4635–4643.
Melters DP, et al. 2013. Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. Genome Biol. 14:R10.
Milligan B. 1998. Total DNA isolation. In: Hoelzel AR, editor. Molecular genetic analysis of populations: A practical approach. Oxford (UK): IRL Press. p. 29–64.
Misawa K, Kikuno RF. 2010. GeneWaltz—a new method for reducing the false positives of gene finding. BioData Min. 3:6.
Morales-Hojas R, Vieira J. 2012. Phylogenetic patterns of geographical and ecological diversification in the subgenus *Drosophila*. PLoS One 7: e49552.
Nadalin F, Vezzi F, Policriti A. 2012. GapFiller: a *de novo* assembly approach to fill the gap within paired reads. BMC Bioinformatics 13(Suppl 14): S8.
Nardon C, et al. 2005. Is genome size influenced by colonization of new environments in dipteran species? Mol Ecol. 14:869–878.
Negre B, et al. 2005. Conservation of regulatory sequences and gene expression patterns in the disintegrating *Drosophila* Hox gene complex. Genome Res. 15:692–700.
Negre B, Ruiz A. 2007. HOM-C evolution in *Drosophila*: is there a need for Hox gene clustering? Trends Genet. 23:55–59.
Negre B, Simpson P. 2009. Evolution of the *achaete-scute* complex in insects: convergent duplication of proneural genes. Trends Genet. 25:147–152.
Nielsen R. 2005. Molecular signatures of natural selection. Annu Rev Genet. 39:197–218.
Oliveira DCSG, et al. 2012. Monophyly, divergence times, and evolution of host plant use inferred from a revised phylogeny of the *Drosophila repleta* species group. Mol Phylogenet Evol. 64:533–544.
Oliveros J. 2007. VENNY. An interactive tool for comparing lists with Venn diagrams. Madrid (Spain): BioinfoGP CNB-CSIC.
Ometto L, et al. 2013. Linking genomics and ecology to investigate the complex evolution of an invasive *Drosophila* pest. Genome Biol Evol. 5: 745–757.
Pál C, Papp B, Lercher MJ. 2006. An integrated view of protein evolution. Nat Rev Genet. 7:337–348.
Palmieri N, Kosiol C, Schlötterer C. 2014. The life cycle of *Drosophila* orphan genes. eLife 3:e01311.
Parisi M, et al. 2004. A survey of ovary-, testis-, and soma-biased gene expression in *Drosophila melanogaster* adults. Genome Biol. 5: R40.

Parsons P. 1983. The evolutionary biology of colonizing species. New York: Cambridge University Press.
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.
Pohl JB, et al. 2013. Circadian genes differentially affect tolerance to ethanol in *Drosophila*. Alcohol Clin Exp Res. 37:1862–1871.
Poptsova MS, Gogarten JP. 2010. Using comparative genome analysis to identify problems in annotated microbial genomes. Microbiology 156: 1909–1917.
Prada CF 2010. Evolución cromosómica del cluster *Drosophila martensis*: origen de las inversiones y reutilización de los puntos de rotura [PhD thesis]. Barcelona (Spain): Universitat Autònoma de Barcelona.
Rajpurohit S, Oliveira CC, Etges WJ, Gibbs AG. 2013. Functional genomic and phenotypic responses to desiccation in natural populations of a desert drosophilid. Mol Ecol. 22:2698–2715.
Rauschenbach IY, Bogomolova EV, Karpova EK, Shumnaya LV, Gruntenko NE. 2012. The role of D1 like receptors in the regulation of juvenile hormone synthesis in *Drosophila* females with increased dopamine level. Dokl Biochem Biophys. 446:231–234.
Reed LK, Nyboer M, Markow TA. 2007. Evolutionary relationships of *Drosophila mojavensis* geographic host races and their sister species *Drosophila arizonae*. Mol Ecol. 16:1007–1022.
Reinhardt JA, et al. 2013. *De novo* ORFs in *Drosophila* are important to organismal fitness and evolved rapidly from previously non-coding sequences. PLoS Genet. 9:e1003860.
Ruiz A, Cansian AM, Kuhn GC, Alves MA, Sene FM. 2000. The *Drosophila serido* speciation puzzle: putting new pieces together. Genetica 108: 217–227.
Ruiz A, Heed WB. 1988. Host-plant specificity in the cactophilic *Drosophila* mulleri species complex. J Anim Ecol. 57:237–249.
Ruiz A, Heed WB, Wasserman M. 1990. Evolution of the mojavensis cluster of cactophilic *Drosophila* with descriptions of two new species. J Hered. 81:30–42.
Ruiz A, Wasserman M. 1993. Evolutionary cytogenetics of the *Drosophila buzzatii* species complex. Heredity 70:582–596.
Ruiz-Ruano FJ, et al. 2011. DNA amount of X and B chromosomes in the grasshoppers *Eyprepocnemis plorans* and *Locusta migratoria*. Cytogenet Genome Res. 134:120–126.
Sawyer SA, Parsch J, Zhang Z, Hartl DL. 2007. Prevalence of positive selection among nearly neutral amino acid replacements in *Drosophila*. Proc Natl Acad Sci U S A. 104:6504–6510.
Schaeffer SW, et al. 2008. Polytene chromosomal maps of 11 *Drosophila* species: the order of genomic scaffolds inferred from genetic and physical maps. Genetics 179:1601–1655.
Schneider A, et al. 2009. Estimates of positive Darwinian selection are inflated by errors in sequencing, annotation, and alignment. Genome Biol Evol. 1:114–118.
Sella G, Petrov DA, Przeworski M, Andolfatto P. 2009. Pervasive natural selection in the *Drosophila* genome? PLoS Genet. 5:e1000495.
Singh ND, Larracuente AM, Sackton TB, Clark AG. 2009. Comparative genomics on the *Drosophila* phylogenetic tree. Annu Rev Ecol Evol Syst. 40:459–480.
Slater GS, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6:31.
Smith G, Lohse K, Etges WJ, Ritchie MG. 2012. Model-based comparisons of phylogeographic scenarios resolve the intraspecific divergence of cactophilic *Drosophila mojavensis*. Mol Ecol. 21: 3293–3307.
St Pierre SE, Ponting L, Stefancsik R, McQuilton P, FlyBase Consortium. 2014. FlyBase 102—advanced approaches to interrogating FlyBase. Nucleic Acids Res. 42(Database issue):D780–D788.
Stanke M, Waack S. 2003. Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics 19(Suppl 2): ii215–ii225.
Starmer WT. 1981. A comparison of *Drosophila* habitats according to the physiological attributes of the associated yeast communities. Evolution 35:38–52.
Stein LD, et al. 2002. The generic genome browser: a building block for a model organism system database. Genome Res. 12: 1599–1610.
Tautz D, Domazet-Lošo T. 2011. The evolutionary origin of orphan genes. Nat Rev Genet. 12:692–702.
Tesler G. 2002. GRIMM: genome rearrangements web server. Bioinformatics 18:492–493.
Throckmorton L. 1975. The phylogeny, ecology and geography of *Drosophila*. In: King R, editor. Handbook of genetics. Vol. 3. New York: Plenum Press. p. 421–469.
Trapnell C, et al. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 28:511–515.
Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25:1105–1111.
Ugarković Đ. 2009. Centromere-competent DNA: structure and evolution. In: Ugarković Đ, editor. Centromere. Structure and evolution. Progress in molecular and subcellular biology. Vol. 48. Berlin (Germany): Springer. p. 53–76.
Vanaphan N, Dauwalder B, Zufall RA. 2012. Diversification of takeout, a male-biased gene family in *Drosophila*. Gene 491:142–148.
Vitti JJ, Grossman SR, Sabeti PC. 2013. Detecting natural selection in genomic data. Annu Rev Genet. 47:97–120.
Wall DP, Deluca T. 2007. Ortholog detection using the reciprocal smallest distance algorithm. Methods Mol Biol. 396:95–110.
Wang J, et al. 2003. Vertebrate gene predictions and the problem of large genes. Nat Rev Genet. 4:741–749.
Wasserman M. 1992. Cytological evolution of the *Drosophila repleta* species group. In: Krimbas CB, Powell JR, editors. *Drosophila* inversion polymorphism. Boca Raton (FL): CRC Press. p. 455–552.
Weake VM, Workman JL. 2010. Inducible gene expression: diverse regulatory mechanisms. Nat Rev Genet. 11:426–437.
Wicker T, et al. 2007. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 8:973–982.
Wong WS, Yang Z, Goldman N, Nielsen R. 2004. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. Genetics 168: 1041–1051.
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24:1586–1591.
Yang Z, Nielsen R, Goldman N, Pedersen AM. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155:431–449.
Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol. 22:2472–2479.
Zhang YE, Landback P, Vibranovski M, Long M. 2012. New genes expressed in human brains: implications for annotating evolving genomes. Bioessays 34:982–991.
Zhou Q, Bachtrog D. 2012. Sex-specific adaptation drives early sex chromosome evolution in *Drosophila*. Science 337:341–345.
Zhou Q, et al. 2012. Deciphering neo-sex and B chromosome evolution by the draft genome of *Drosophila albomicans*. BMC Genomics 13:109.

Associate editor: Mar Alba

6

ACKNOWLEDGMENTS

This is the end of a chapter in my life. A chapter that it would not be as it has been without the people who accompanied me.

First, I want to thank Alfredo, because he accepted me as a master student and later as a PhD student giving me projects that allowed me not only to discover the TE world, or the Drosophila field, but to grow as a scientist. For giving me the freedom to learn but being there when I needed guidance. I want to thank Alejandra, for teaching me how to move around the lab and more importantly for teaching me to not give up even when I could not see the light. From my stay in Texas I want to thank Cedric for hosting me and giving me the possibility to see another reality. I enjoyed your classes and I still remember the passion you transmitted. Thanks to Esther and all the people from UTA. From that period I am most grateful for meeting Aurelie and the friends I made there. Thank you for making me feel at home. Quiero agradecer también a mis compañeras Andrea y Yolanda, el brindarme la oportunidad de aprender de ellas en muchos sentidos y a Mar que fue desde el principio un ejemplo muy cercano. A David, Maite y Miquel, que pese a estar en otro edificio han estado siempre cerca. Als meus pares que m'han ajudat sempre en tot el que han pogut, i al meu germà que és la persona més generosa que conec, mai us ho podré agrair prou. A les meves amigues que m'han acompanyat en aquest procés, una abraçada! A tots els Sols que han fet els meus dies més càlids i agradables, moltes gràcies.

Thank you. Muchas gracias. Moltes gràcies.