






Universitat Autònoma de Barcelona

ADVERTIMENT. L'accés als continguts d'aquesta tesi queda condicionat a l'acceptació de les condicions d'ús establertes per la següent llicència Creative Commons:  http://cat.creativecommons.org/?page_id=184

ADVERTENCIA. El acceso a los contenidos de esta tesis queda condicionado a la aceptación de las condiciones de uso establecidas por la siguiente licencia Creative Commons:  <http://es.creativecommons.org/blog/licencias/>

WARNING. The access to the contents of this doctoral thesis it is limited to the acceptance of the use conditions set by the following Creative Commons license:  <https://creativecommons.org/licenses/?lang=en>

DOCTORAL THESIS

Development of a high-quality annotated reference genome
and evolutionary genomics analysis of chromosomal
inversions in *Drosophila subobscura*.

Author

Charikleia Karageorgiou

Directors

Dr. Francisco J. Rodríguez-Trelles Astruga and Dr. Rosa M. Tarrío
Fernández

Tutor

Dr. Francisco J. Rodríguez-Trelles Astruga



**Universitat Autònoma
de Barcelona**

Departament de Genètica i de Microbiologia

Facultat de Biociències

Universitat Autònoma de Barcelona

2022

Development of a high-quality annotated reference genome
and evolutionary genomics analysis of chromosomal
inversions in *Drosophila subobscura*.

Memòria presentada per Charikleia Karageorgiou
per a optar al grau de Doctora en Genètica
per la Universitat Autònoma de Barcelona

Autora

Directors

Charikleia Karageorgiou

Dr. Francisco J. Rodríguez-Trelles
Astruga
&
Dr. Rosa M. Tarrío Fernández

Bellaterra, 27 de Julio de 2022

Contents

Abstract	1
1. Introduction	3
1.1 Inversions	3
1.1.1 Classification of chromosomal inversions	4
1.1.2 Origins of polymorphic inversions	5
1.1.3 Identification and characterisation of inversions and their breakpoints	7
1.1.4 Inversions' recombination suppression effect	7
1.1.5 Mechanisms that lead to the establishment of polymorphic inversions	8
1.1.6 Maintenance of inversion polymorphisms	9
1.1.7 The double-edged sword of recombination suppression	10
1.1.8 Role of inversions in the evolution of supergenes	10
1.1.9 Outstanding questions regarding polymorphic inversions	11
1.2 <i>Drosophila subobscura</i>	12
1.2.1 <i>Drosophila subobscura</i> geographic distribution and species ecology	12
1.2.2 Inversion polymorphism in <i>D. subobscura</i>	13
1.2.3 Adaptive inversion polymorphism in <i>D. subobscura</i>	14
1.2.4 <i>D. subobscura</i> as a model species, challenges and limitations	15
2. Objectives	17
3. Results	19
3.1 Chapter 1: Long-read based assembly and synteny analysis of a reference <i>Drosophila subobscura</i> genome reveals signatures of structural evolution driven by inversions recombination-suppression effects	19
3.2 Chapter 2: The cyclically seasonal <i>Drosophila subobscura</i> inversion O ₇ originated from fragile genomic sites and relocated immunity and metabolic genes	41
4. Discussion	65
4.1 <i>De novo</i> genome assembly and short-read limitations	65
4.2 Genome browser a resource to handle data efficiently	66
4.3 Insights and future perspectives into inversion polymorphisms in <i>Drosophila subobscura</i>	67
5. Conclusions	69
6. Publications from this Thesis	70
Appendices	71
A. Supplementary material of “Long-read based assembly and synteny analysis of a reference <i>Drosophila subobscura</i> genome reveals signatures of structural evolution driven by inversions recombination-suppression effects”	71
i. Supplementary Figures	71
ii. Supplementary Tables	82

B. Supplementary material of “ The cyclically seasonal <i>Drosophila subobscura</i> inversion O ₇ originated from fragile genomic sites and relocated immunity and metabolic genes”	89
i. Supplementary Figures	89
ii. Supplementary Tables	96
References	97

Abstract

Drosophila subobscura belongs to the obscura species group, which is one of the known ten species groups of the subgenus *Sophophora* of the genus *Drosophila*. Originally endemic from the Palearctic region, it has recently colonized North and South America. The chromosomal inversion system of *D. subobscura* represents one of the most interesting models to investigate the evolution of this type of genome rearrangement, because i) it shows extremely high levels of polymorphism, exhibiting inversions of all kinds regarding length and chromosomal location; and ii) it has been identified to be involved in the species' adaptation to contemporary global climate warming. The lack of a reference genome for the species has, however, stood as a major obstacle to its study. To overcome this limitation, here we have tackled a *de novo* assembly of the genome of *D. subobscura* using PacBio long-read technology. Raw PacBio reads were assembled using the Canu assembler. A semi-automated pipeline for assessing the quality of the contigs and scaffolding the genome has been developed that combined both synteny information from previously assembled genomes of *D. melanogaster* and *D. pseudoobscura*, and published data from 560 genetic markers derived from *in situ* hybridization experiments and genetic linkage analyses. Canu assembled contigs were then scaffolded using SSPACE-LongRead, which resulted in 186 scaffolds with a N50 of approximately 6Mb. Chromosomal assignment, ordering and orientation of the scaffolds resulted in six pseudochromosomes, one for each of the six *D. subobscura* chromosomes. Annotation of the newly assembled genome was conducted with the MAKER annotation pipeline, using *ab initio* gene model predictions and available proteomes from 12 sequenced and annotated *Drosophila* species. Repetitive elements were identified by RepeatMasker, and two previously described species-specific satellites (sat290 and SGM-sat) were identified. A total of 13,939 protein-coding genes were predicted, and 13% of the species genome was found to consist of repetitive sequences. Finally, the amounts of genome rearrangement between *D. subobscura*, *D. guanche*, *D. melanogaster* and *D. pseudoobscura* was assessed. The breakpoints of the fixed inversions between *D. subobscura* and *D. guanche* were determined and characterized. Here we illustrate that genome structure evolution in *D. subobscura* is driven indirectly, through the inversions' recombination-suppression effects in maintaining sets of adaptive alleles together in the face of gene flow.

Presently, *D. subobscura* is experiencing a rapid replacement of high-latitude by low-latitude inversions associated with global warming. However, not all low-latitude inversions are correlated with the secular warming trend. The mixed behavior of O₇ inversion across components of the ambient temperature suggests that it is driven by selective factors other than temperature alone. Research into this question has been hindered by lacking knowledge of the genomic breakpoint sequences of the inversion, which would inform both its mechanism of origin and what genes it altered. To tackle this limitation, we generated a PacBio long read-based chromosome-scale genome assembly, from an O₃₊₄₊₇ isogenic line following the previously described pipeline. The complete continuous sequence of O₇ was isolated using synteny analysis with the reference genome. Inversion O₇ was shown to stretch 9.936 Mb, containing over 1,000 annotated genes. We illustrate that inversion O₇ had a complex origin, involving multiple breaks associated with non-B DNA motifs, formation of a

microinversion, and ectopic repair in trans with the two homologous chromosomes. Our findings support that inversion O₇ breakpoints carry a pre-inversion record of fragility, including a sequence insertion, and transposition with later inverted duplication of an *Attacin* immunity gene. The O₇ inversion was found to have relocated the major insulin signaling forkhead box subgroup O (*foxo*) gene bringing it in tight linkage with its antagonistic regulatory partner serine/threonine-protein kinase B (*Akt1*). Further, its distal breakpoint disrupted concerted evolution of the two inverted *Attacin* duplicates, reattaching them to dFOXO metabolic enhancers. We suggest that O₇ exerts antagonistic pleiotropic effects on reproduction and immunity, setting a framework to understand its relationship with climate change. Our findings have general implications for current theories on the molecular mechanisms of formation of inversions and the contribution of breakage versus repair in shaping inversion-breakpoint junctions.

1. Introduction

In the last two decades extensive efforts have been devoted to characterizing genetic variation in different species. This effort was initially concentrated on a few model species, but ongoing methodological advances have widened the focus to virtually any organism. *Drosophila* species have been predominantly used in genetics research in an effort to describe the effects of drift and selection in shaping patterns of genetic polymorphism and divergence across the genome. Recent advances in next generation sequencing (NGS) have greatly facilitated research into the patterns of diversity. Yet, the majority of the studies conducted focus on single nucleotide polymorphisms (SNPs) commonly overlooking structural variants (SVs).

SVs can span in size from a few to thousands of basepairs, and can include insertions, deletions, duplications, inversions, and translocations. SVs account for a large proportion of the genetic variation within and between species, nevertheless they represent one of the least studied classes of genetic variation. SVs can be classified as unbalanced, if they increase or decrease the total amount of DNA (insertions, deletions and duplications), and balanced, if they revert the orientation (inversions) or alter the location (translocations) of DNA sequences without altering the total amount of genetic sequence. Particularly inversions have attracted much interest currently, being thought to be involved in central evolutionary processes, such as local adaptation (Dobzhansky, 1949; Kirkpatrick & Barton, 2006), speciation (Noor *et al.*, 2001; Hoffmann & Rieseberg, 2008) and sex chromosome evolution (Kirkpatrick, 2010). Evidence about the mechanisms that lead to the formation of inversions has continued to accumulate, yet their mechanisms of establishment and maintenance in populations remain comparatively poorly known.

1.1 Inversions

Chromosomal inversions are ubiquitous across the tree of life. They were the first type of SV to be discovered. Their existence was deduced from their recombination suppression effect by Alfred Sturtevant (1921), before they could be observed. The first direct evidence of their occurrence came from the cytological loops they form in giant polytene chromosomes of *Drosophila*, when in heterozygous combination with the standard orientation (Dobzhansky & Sturtevant, 1938; Cooper, 1938). Shortly after their cytological discovery, Theodosius Dobzhansky and his collaborators began to investigate the population genetics of inversions, gathering extensive information on the spatiotemporal patterns of inversion polymorphisms, particularly in natural populations of *Drosophila pseudoobscura* in an attempt to understand their evolutionary significance.

1.1.1 Classification of chromosomal inversions

Inversions can be classified as pericentric, if they include the centromere, or paracentric if they affect only one chromosome arm. In *Drosophila*, the study of inversion polymorphisms has typically focused on paracentric inversions, which are by far the most common (Krimbas and Powell, 1992). During meiosis, pairing of the inverted and standard chromosome configurations result in the formation of a loop. Single crossovers in heterozygotes for paracentric inversions result in unbalanced recombinant gametes carrying acentric or dicentric chromatids (Sturtevant & Beadle, 1936; Anton *et al.*, 2005). In the particular case of *Drosophila*, males do not recombine and females exhibit specific cytological mechanisms (*e.g.*, formation of the polar body nuclei) that eliminate recombinant chromatids, thereby segregating inversions are assumed to incur no fitness cost (Carson, 1946).

Paracentric inversions were initially postulated to arise via ectopic or non-allelic homologous recombination (NAHR) of segmental duplications, transposons and repetitive elements. This idea was supported by early findings of transposable elements nearby or at breakpoints of polymorphic inversions in different *Drosophila* species (Cáceres *et al.*, 1999; Andolfatto *et al.*, 1999). Ever since, a number of inversion breakpoints have been identified and described at nucleotide resolution, allowing to better pinpoint the originating mechanisms. Currently, inversions are known to form through additional mechanisms to NAHR, including non-homology or microhomology based mechanisms, such as non homologous end-joining (NHEJ) and microhomology-mediated end-joining (MMEJ), microhomology-mediated break-induced repair (MMBIR) (Narayanan *et al.*, 2006) and fork stalling and template switching (FoSTeS) (Lee *et al.*, 2007).

Further, inversions can occur as fixed differences between species or as segregating polymorphisms within species. Fixed inversions were either ancestrally segregating polymorphisms, predating their divergence (Fuller *et al.*, 2018) or originated and spread in one of the two lineages after the speciation event (Kirkpatrick, 2010). Fixed inversions can accumulate at different rates in different lineages. Chromosomal inversions are shown to be more easily fixed in populations when they have weakly underdominant fitness effects or when they are neutral (Kirkpatrick & Barton, 2006). Fixed inversions in *Drosophila* species were found to be involved in speciation via Bateson-Dobzhansky-Muller (BDM) incompatibilities accumulated inside them during population divergence (Noor *et al.*, 2001). In the case of *Drosophila pseudoobscura* and *Drosophila persimilis*, Fuller *et al.* (2018) have demonstrated that nearly all genes involved in reproductive isolation are located within fixed inversions. Their findings suggest that inversions might serve as “fertile grounds” for the formation of hybrid incompatibilities rather than as protectors of existing hybrid incompatibilities, as previously proposed.

Polymorphic inversions can be classified as rare or common/widespread based on their frequencies, and as cosmopolitan or endemic considering their geographic ranges. Polymorphic inversions that have been observed in many populations of a species'

geographical range are considered cosmopolitan. On the contrary, inversions that are geographically restricted and can be observed only in certain populations are thought to be endemic inversions (reviewed in Krimbas and Powell, 1992).

Inversions can be relatively small, spanning less than 1Kb, such as the inversions that are commonly observed in humans, or large, greater than 1Mb, such as many of the inversions detected in dipterans like *Drosophila* and *Anopheles* (Kirkpatrick, 2010). Powell (1997) found that inversion heterozygotes can experience a reduction in fertility. Nonetheless, this decline was not apparent in the *Drosophila* genus, likely due to the formation of the polar body nuclei in *Drosophila*, which can efficiently eliminate abnormal chromatids. This observation may help explain the different distribution of inversion lengths between the aforementioned two lineages. Yet emerging evidence (Cáceres *et al.*, 1999; Porubsky *et al.*, 2020; McBroome *et al.*, 2020; Wright & Schaeffer, 2022) is pointing to the presence of alternative mechanisms that might explain the different length patterns across species (*e.g.*, negative correlation between recombination maps and inversion length, breakpoints to occur at boundaries between TADs, or disruption of TADs).

The length of an inversion might influence its evolutionary fate. The direction and strength of selection on inversions can vary considerably with their lengths, thus information regarding the inversion length can provide insights about its establishment. Large inversions can have enhanced probabilities of establishment since they are more likely to capture beneficial alleles and can affect multiple traits (Cheng & Kirkpatrick, 2019). On the other hand, it is more likely for large inversions to capture deleterious mutations (Nei *et al.*, 1967; Kimura & Ohta, 1970). Connallon and Olito (2021) proposed that larger inversions should evolve under local adaptation scenarios, while smaller inversions are more likely to spread when they are either underdominant or directly beneficial. Empirical studies (Cáceres *et al.*, 1997; Messer, 2009; Corbett-Detig, 2016; Cheng & Kirkpatrick, 2019) as well as theoretical studies employing modeling (Connallon & Olito, 2021) have attempted to address the relation between establishment probability and inversion length.

1.1.2 Origins of polymorphic inversions

Two major molecular mechanisms have been advanced to explain the formation of inversions. The first mechanism is referred to as the intrachromatid ectopic or non-allelic homologous recombination (NAHR) model. The second mechanism involves chromosomal breakage and ectopic repair via non-homologous end joining (NHEJ). The two mechanisms produce distinct signatures. NAHR generates inversions with duplications at their ends in both the inverted and uninverted states (Cáceres *et al.*, 1999) while NHEJ either does not generate duplications at all (Wesley & Eanes, 1994) or when it does generate duplications, they can only be detected at the breakpoints of the inverted state (Kehrer-Sawatzki *et al.*, 2005; Matzkin *et al.*, 2005; Ranz *et al.*, 2007). The two distinct footprints of NHEJ are thought to result from differences in the mode of breakage. So far two models of breakage have been

proposed: clean double-strand breaks (DSBs) that generate blunt ends, also known as “cut-and-paste”, and staggered DSBs which give rise to duplications at the inversions’ ends. The prevalence and distribution of the two mechanisms of inversion formation, namely NAHR and NEHJ, within and across lineages are currently under investigation (Ranz *et al.*, 2007; Delprat *et al.*, 2019; Karageorgiou *et al.*, 2020).

Most of the inversions investigated so far are considered to be monophyletic, *i.e.*, originated from unique mutational events (Krimbas & Powell, 1992) with examples of recurrent cytologically identical inversions being relatively scarce (Goidts *et al.*, 2005; Aguado *et al.*, 2014). With regards to monophyletic inversions, Corbett-Detig (2016) has shown that certain genomic regions are more prone to breakage, and thus inversion breakpoints are more likely to occur within those regions. Regions with the aforementioned properties have been defined as “*sensitive sites*”. The fragile nature of such sites can be attributed to an excess of repetitive sequences, transposable elements (TEs) and low complexity sequences, for instance simple repeats. Another long-standing question with regards to inversions is if they recur, *i.e.*, originate repeatedly over time. To answer this question it is important to disentangle the signatures between multiple independent origins of an inversion and adaptive introgression of an inversion from a common ancestor.

Recurrent inversions are thought to be mediated by NAHR mechanisms, because they are commonly associated with NAHR hotspots (Aguado *et al.*, 2014; Coe *et al.*, 2014; Porubsky *et al.*, 2022). NAHR hotspots are commonly rich in inverted repeats and segmental duplications, thus they are prone to recurrent events. Studies on human chromosomal inversions have revealed a number of cytologically and molecularly identical inversions with multiple origins, also known as polyphyletic origin. The relative high frequency of recurrent inversions observed in humans has been attributed to the presence of large inverted repeats at the breakpoints (Aguado *et al.*, 2014). Several chromosomal regions that have suffered independent breaks reiteratively have been identified in humans and mammals (Murphy *et al.*, 2005). On similar grounds, Cáceres *et al.* (2007) have shown long-term breakpoint reuse during the evolution of mammalian species using a human X-chromosome polymorphic inversion. Similar observations regarding breakpoint reuse have been made using *Drosophila* species (Bhutkar *et al.*, 2008). Breakpoint reuse analysis in 12 species of the *Drosophila* genus suggests that inversion breakpoints tend to be reused at a higher rate in the *Sophophora* lineages than in the *Drosophila* lineages (Bhutkar *et al.*, 2008).

Further, inversions might be introduced into populations through hybridization (dellaTorre *et al.*, 1997). Inversions implicated in adaptation to certain environments have been shown to be acquired via introgression in *Anopheles* (Besansky *et al.*, 2003; White *et al.*, 2007) and *Rhagoletis* (Feder *et al.*, 2003) amongst other species. Kirkpatrick and Barrett (2015) proposed that inversions could serve as adaptive cassettes that can accelerate adaptation by crossing species boundaries.

1.1.3 Identification and characterisation of inversions and their breakpoints

Chromosomal inversions were classically identified using cytological methods, taking advantage of the large polytene chromosomes found in the salivary glands of *Diptera*. (Dobzhansky & Sturtevant, 1938; Kunze-Mühl & Müller, 1957). The identification of polymorphic inversions using the cytological approach can only be applied to dipterans, and can be rather laborious, requiring a large investment of time to familiarize with the experimental methods of chromosomal preparation and inversion identification. Up-to-date, chromosomal preparations and karyotyping are commonly used to identify and verify the presence of inversions in *Drosophila* populations. However, this approach does not permit molecular characterization of inversion breakpoints, instead the breakpoints can only be determined at cytological resolution. The first characterized breakpoints were obtained using probes and *in situ* hybridization on polytene chromosomes (Andolfatto *et al.*, 1999). Until recently, breakpoint identification using probes and chromosome walking remained the approach of choice. Although a widely used approximation, chromosome walking relies on the availability of the genome of a close relative to the target species to be used as a guide for the design of the probes. Recent methodological advances have allowed the detection and characterization of inversion breakpoints with nucleotide resolution. Such breakthroughs have facilitated the detection of polymorphic inversions and their breakpoints in a number of species via the development of polymerase chain reaction (PCR) markers and tag SNPs (Wesley & Eanes, 1994; Andolfatto *et al.*, 1999).

The last decade whole-genome sequencing has further facilitated the detection and identification of inversion breakpoints. Mate pair libraries are widely used to map the locations of inversion breakpoints. Mate pairs that span inversion breakpoints will map at distances remarkably larger than the expected insert size of the library. Additionally, the reverse/forward orientation of the mate pairs is expected to be distorted resulting in pairs where both reads exhibit the same orientation. Long-read sequencing has revolutionized structural variant detection offering the most comprehensive approach so far while enabling population genomics analyses that can answer long-standing questions about the genetic variation within and between arrangements, the demographic history of polymorphic inversions and their origin.

1.1.4 Inversions' recombination suppression effect

The fact that inversions have an associated recombination effect does not imply that they suppress recombination completely. In fact, they can exchange genetic information with alternative arrangements via double crossovers, at their centers, and gene conversion, uniformly across their lengths (Korunes & Noor, 2017). However, the rates at which these phenomena occur do not invalidate the assumption that inversions are inherited as single Mendelian units, particularly at the regions of their breakpoints (Powell, 1997). Inversions' recombination suppression effects can alter the recombination landscape among loci and have

been studied thoroughly. The prevalent view holds that inversions can be favored by natural selection because they link together coadapted alleles at multiple loci in the face of gene flow (Sturtevant & Beadle, 1936; Corbett-Detig & Hartl, 2012; Fuller *et al.*, 2019). Dobzhansky proposed that new inversions spread in populations because they hold together epistatically interacting coadapted alleles against the dissociative effects of recombination (Dobzhansky and Epling, 1948). Recent theoretical studies (Kirkpatrick and Barton, 2006) have shown that epistasis is not a requirement for inversions to confer local adaptation. On the other hand, the suppression of recombination can come at a cost, since reduction of the efficiency of natural selection can result in accumulation of deleterious mutations and/or rapid loss of beneficial ones (Hill & Robertson, 1968; Felsenstein, 1974). Lastly, suppression of recombination can distort a population's demography generating structure and reducing the effective population size.

1.1.5 Mechanisms that lead to the establishment of polymorphic inversions

Most newly arising inversions are expected to be deleterious, and thus to be quickly removed by negative selection, or (less so) to be neutral, in which case they can remain drifting in populations for long periods of time. Early field and laboratory evidence of adaptive inversions came from Dobzhansky and his collaborators working with *D. pseudoobscura*. This early evidence was corroborated by subsequent *Drosophila* and *Anopheles* studies that found the frequencies of many cytological inversions showing systematic, thus adaptive patterns of variation both at temporal and/or geographical scales. Recent genomics studies have extended these conclusions to inversions from a number of other organisms such as *Heliconius* butterflies (Jay *et al.*, 2022), frogs (Dufresnes & Crochet, 2022), ants (Kay *et al.*, 2022), Atlantic cod (Matschiner *et al.*, 2022). It is one thing to demonstrate adaptation, and quite another to understand its underlying mechanisms, including targets of selection and the specific selective regimes.

New inversions can spread in a population due to direct effects of the mutational event *per se* on the structure and/or expression of the genes and functional sequences at the breakpoints (McBroome *et al.*, 2020). On the other hand, indirect effects can emerge from the recombination-suppression effect, when inversions “lock” together advantageous combinations of alleles. Dobzhansky (1947) was the first to propose that polymorphic inversions are maintained in populations due to indirect effects, a view that was later expanded by Wasserman (1968) and Kirkpatrick and Barton (2006).

Leaving aside the aforementioned genetic drift, there are currently five major hypotheses, hereon hypothesis 1-5, for the establishment of polymorphic inversions in populations. Two of them place the focus on the effects of the suppression of recombination in maintaining linkage disequilibrium between loci found within the inverted region. Hypothesis 1 followed Dobzhansky's view that inversions spread in a population due to the reduced recombination and epistatic selection between loci. Under this coadaptation model, epistatically interacting

loci would yield higher fitness than predicted by the sum of their independent fitness effects. Hence, coadapted alleles will be favored by selection and inversions are expected to segregate at high frequencies and ultimately reach fixation in absence of migration or any sorts of counteracting selection. Hypothesis 2 is based on theoretical results by Kirkpatrick and Barton (2006), who showed that inversions can be locally adaptive without the need to invoke epistatic interactions between the beneficial alleles. Under this additive model, the inversion would be favored and maintained in a population at migration-selection balance without reaching fixation, due to the introduction of new combinations of alleles by migration.

Hypothesis 3 relies on the direct effects of the chromosomal lesion. Accordingly, inversions would be favored by direct selection due to the positive effects of the breakage and the breakpoints. Here, the inversion itself would be under selection. Be that as it may, the inversion breakpoints can directly disrupt a gene or alternatively they can disrupt gene expression through position effects (Tadin-Strapps *et al.*, 2004). The fate of such inversions depends entirely on the fitness effects caused by the lesion and/or their position effects. Hypothesis 4 is built on overdominance. Under the overdominance scenario, inversion heterozygotes have higher fitness than either the homozygotes for the ancestral non-inverted arrangement or the homozygotes for the inverted arrangement. Overdominance can originate by deleterious alleles in the inverted region. Consequently, if both the inverted and non-inverted regions carry different deleterious mutations there will be a heterozygote advantage, which will lead to the establishment of the inversion as a balanced polymorphism. Lastly, hypothesis 5, proposes that for the spread of an inversion, underdominance can be invoked. In that case selection against the heterozygote would be observed due to the overall reduced fitness of the heterozygotes. Underdominance can arise if single crossover events occur relatively frequently within the inversion resulting in the production of unbalanced gametes. This phenomenon is particularly common in plants (Rieseberg, 2001). The fate of underdominant inversions is to either reach fixation or become loss.

1.1.6 Maintenance of inversion polymorphisms

A controversial issue regarding the establishment and maintenance of inversions is whether the beneficial alleles are captured when the inversion arises or instead they are acquired gradually after its emergence. So far there is supporting evidence for both scenarios. Coughlan and Willis (2019) have shown an inversion that has captured locally adapted alleles when it first arose; while Lamichhaney *et al.* (2016) provide compelling evidence for the acquisition of the locally adapted alleles over time which has led to the establishment and maintenance of the inversion in the population. Be that as it may, the maintenance of adaptive inversions in a population depends on interactions among genetic drift, gene flow, selection and recombination. Adaptive inversions can become globally fixed under positive selection. Hitchhiking can facilitate the spread and maintenance of a new inversion that has captured a beneficial allele in a population. Yet, once established, adaptive inversions can segregate at intermediate frequencies without reaching fixation. Balancing selection can maintain genetic

variation in populations for longer periods than those expected by chance by overcoming its stochastic loss or fixation by genetic drift. Certain inversions were shown to segregate within populations for millions of years (Gutiérrez-Valencia *et al.*, 2021). Such inversions are thought to be maintained via balancing selection mechanisms such as frequency dependent selection (Maynard Smith, 1998; Takahashi & Kawata, 2013), heterozygote advantage /overdominance (Fisher, 1923; Wallace, 1970), via individually overdominant loci or associative overdominance due to recessive deleterious alleles, antagonistic selection (including sexually antagonistic selection), temporally variable selection (Wittmann *et al.*, 2017), spatially fluctuating selection (Levene, 1953; Hedrick, 2006) and disassortative mating (Lewontin *et al.*, 1968). New mutations arise as a structural variant evolves over time resulting in the divergence of the inversion haplotype from ancestral haplotype owing to the suppression of recombination between the two.

1.1.7 The double-edged sword of recombination suppression

Suppressed recombination is one of the prerequisites for the emergence of supergenes, yet long-term suppression of recombination can have detrimental consequences for the organisms that harbor inversions. Reduced rates of recombination can facilitate adaptation to diverse environments but are shown to result in accumulation of repeats, deletions and deleterious mutations, which in turn results in the degeneration of the supergene. The accumulation of deleterious alleles and thus the degeneration of non-recombining regions can be justified by the reduced efficacy of selection on the region.

1.1.8 Role of inversions in the evolution of supergenes

Supergenes are sets of functionally related loci that are so tightly linked as to segregate as a single entity, thus allowing switching between discrete, complex phenotypes maintained in a balanced polymorphism (Thompson & Jiggins, 2014). Some of the best examples of supergenes are the loci controlling polymorphisms for Batesian mimicry in butterflies (Sheppard, 1959). Supergene architecture is generally seen as ensuing from selection for tight linkage driven by benefits of coinheriting alleles from functionally related loci. Therefore, it is not surprising that inversions are increasingly emerging as one main mechanism that might facilitate the evolution of supergenes. Nonetheless, it should be recalled that the relationship between inversions and supergenes has a long history that has been traced back to Dobzhansky's pioneering view of inversions as blocks of coadapted gene complexes (Thompson & Jiggins, 2014). The precise quantitative relevance of inversions for supergene evolution relative to other recombination-suppression mechanisms, such as modifiers controlling the number of crossover events (Charlesworth, 2015) is currently under debate.

1.1.9 Outstanding questions regarding polymorphic inversions

The evolutionary significance of polymorphic inversions has recently been brought to the forefront with the uprise of genomics and next generation sequencing. Several studies have attempted to address questions regarding the origin, establishment, and maintenance of polymorphic inversions utilizing WGS data (Corbett-Detig & Hartl, 2012; Fuller *et al.*, 2016, 2017, 2019; Cheng *et al.*, 2018; Lowry *et al.*, 2019). As it has been previously mentioned, inversions can span thousands of basepairs involving large numbers of genes, hence they exhibit enhanced potential for affecting multiple traits. This can increase inversions' likelihood to be maintained in a population via different mechanisms (Cheng & Kirkpatrick, 2019). In this respect, there are major questions to be answered regarding the mechanisms of selection that are acting on inversions, or questions regarding the ecological factors that lead to the establishment and fixation of inversions.

Growing evidence suggests that inversions can affect gene expression. Such effects have previously been overlooked since the gene spans within inverted regions remain unaltered. Lavington and Kern (2017) have shown that polymorphic inversions in *Drosophila melanogaster* alter the gene expression of hundreds of transcripts in the genome. Similarly, Said *et al.* (2018) observed differentially expressed genes involved in the immune response for the studied *Drosophila melanogaster* inversions. Said *et al.* (2018) proposed that the modified gene expression manifests as a consequence of linked allelic variation that is maintained within inverted regions via suppressed recombination. Nevertheless, the how or why inversions affect gene expression remains unresolved.

Up-to-date, there are several outstanding questions regarding chromosomal inversions. Below I am listing some of them:

- Do adaptive polymorphic inversions capture “coadapted complexes” of epistatically interacting genes, as proposed by Dobzhansky, or do they comprise sets of independently adaptive loci maintained together in strong linkage disequilibrium by recombination-suppression effects, as suggested by Kirkpatrick and Barton (2006)?
- What is the importance of gene conversion and double crossover events in determining the fate and evolution of an inversion in a population?
- How do inversions respond to environmental shifts?
- What is the interaction between selection and demography in shaping inversion polymorphisms?
- How do different inversions interact between them to affect adaptation?
- What are the candidate genes and locally adaptive loci maintained within inversions?
- How can we distinguish amongst the mechanisms that maintain polymorphic inversions within populations?

1.2 *Drosophila subobscura*

D. subobscura (Collin, 1936) belongs to the obscura group of the subgenus *Sophophora*. It was originally encountered in Europe and the palearctic region (Buzzati-Traverso & Scossiroli, 1955) and has only recently colonized North and South America (reviewed in Krimbas, 1993). *D. subobscura* forms the *subobscura* three-species subgroup together with the island endemics *D. madeirensis* and *D. guanche* (Krimbas, 1993; Bächli, 2020). *D. subobscura* is only known to cross with its sister species *D. madeirensis*. Hybridization between these two species can occur between females of *D. subobscura* and males of *D. madeirensis* and in the reciprocal cross, producing sterile males and fertile hybrid females (Krimbas & Loukas, 1984). Both *D. guanche* and *D. madeirensis* can serve as useful outgroups in the study of *D. subobscura*, owing to their island-endemic origin and small effective population size. In addition, chromosome segment homologies for these species have already been established since the early 80s (Krimbas & Loukas, 1984; Moltó *et al.*, 1987; Papacit & Prevosti, 1989). Extensive cytological research on *D. subobscura* salivary gland chromosomes has revealed that it carries the ancestral *Drosophila* karyotype of five large telocentric chromosomes (namely A, J, U, E and O) and a small dot chromosome, while its chromosomes do not show a chromocenter (Emmens, 1937). *D. subobscura* harbors extensive inversion polymorphisms. Although *D. subobscura* has classically served as a model organism in evolutionary genetics, particularly in the study of adaptive character of chromosomal inversion polymorphism, its use was hindered by the lack of a reference genome.

1.2.1 *Drosophila subobscura* geographic distribution and species ecology

D. subobscura can be encountered all over Europe with the exception of Iceland and some parts of Scandinavia, yet these geographic limits are thought to be extended (for more information please consult section 1.2.3). With regards to the distribution of the species to Eastern Europe and Asia the easternmost frontier remains unknown. The species are also present at the southern coast of the Mediterranean sea. *D. subobscura* populations have been found as far south as in Egypt where its southernmost frontier is thought to be (reviewed in Krimbas, 1993). The first populations of *D. subobscura* in the Americas were identified in 1978 at Puerto Montt in Chile expanding its geographic distribution beyond the Palearctic realm (Beckenbach & Prevosti, 1986; Prevosti *et al.*, 1988). The species in less than two decades since the original colonization of North and South America has expanded remarkably its geographic distribution and population density (Mestres *et al.*, 2001).

Understanding factors that influence the species distribution is a fundamental objective in ecology. Up-to-date little is known about the species habitat. Commonly a distinction is made between regions with multiple habitats, which can be defined as ecologically central and regions that exhibit rare habitats, and are also referred to as marginal. Based on the central-marginal hypothesis, populations at geographic range margins exhibit reduced

intra-population genetic diversity and increased inter-population genetic differentiation compared to central populations. The centrality or marginality of *D. subobscura* can be assessed indirectly via population size estimates for the species, relative density information and the frequency of lethal alleles (reviewed in Krimbas, 1993).

D. subobscura is considered a generalist species (Krimbas, 1993), it can be collected in forests of different vegetation and flora as well as in urban areas, associated with human activity (Krimbas, 1993; Kenig *et al.*, 2010). Moreover, the species is polyphagous typically feeding on decaying fruits, fungi and fermenting sap while fruits are shown to be its preferred medium (Begon, 1975). *D. subobscura* is typically encountered in forests and the edges of forests contrary to *D. obscura* which is only encountered in forests and formally considered a forest species (Burla, 1951). Further, it was shown to coexist with other *Drosophilids* across its habitat. Particularly in Central and Northern Europe *D. subobscura* populations coexist with *D. obscura* and both species are found in similar frequencies. This pattern declines towards the South of Europe where *D. subobscura* accounts for the most prevalent species. Lastly the species shows high dispersion capacities. The dispersal of the flies increases with high humidity and at a temperature of 18°C (reviewed in Krimbas, 1993).

1.2.2 Inversion polymorphism in *D. subobscura*

Cytological studies in *D. subobscura* have identified over 65 inversions, including overlapping and non-overlapping inversions that span all its telocentric chromosomes. Owing to the abundance of chromosomal inversions *D. subobscura* could serve as a valuable model organism for structural variation studies and for the investigation of the recombination-suppression effects on the inverted and non-inverted regions of the chromosomes. In the following chapters presented below we focused primarily on the longest of the five telocentric chromosomes, chromosome O, which shows an abundance of polymorphic inversions and corresponds to Müller element E. In total there are 26 chromosomal inversions identified in chromosome O (Krimbas, 1993), which are denoted with the letter O followed by a number as a subscript or “ST” for the standard arrangement. The numbers are arbitrarily assigned based on the time of their discovery. Overlapping inversions are denoted by underlines below number subscripts that correspond to the inversions that overlap on the chromosome. The cytological map of *D. subobscura* is divided into 100 sections and 405 subsections (Kunze-Mühl & Müller, 1958). The O chromosome is conventionally partitioned in two segments and 24 sections (section 75 to 99), segment I that spans from section 91 to its telomere (section 99) and segment II that extends from the centromere (section 75) to section 90. Out of the 26 inversions of the O chromosome that generate 46 gene arrangements, 19 of them are located in segment II while the remaining 6 are found in segment I and there is solely one inversion (namely O₂₅) that partially occupies both regions of segment I and segment II. Currently 12 of the 65 cytologically visible inversions in the species have been studied, having their breakpoints isolated and characterized. None of the characterized breakpoints are shown to have directly disrupted the

structure of protein coding genes (Papaceit *et al.*, 2012; Puerma *et al.*, 2014, 2016a, 2016b, 2017; Orengo *et al.*, 2015; Karageorgiou *et al.*, 2019, 2020).

1.2.3 Adaptive inversion polymorphism in *D. subobscura*

Based on the observations of Dobzhansky (1962), *D. subobscura* inversion polymorphism was initially speculated to not respond to environmental changes, as the population collected near Vienna did not display any seasonal changes. Hence the populations of *D. subobscura* were thought to be genetically “rigid”. Other *Drosophila* species such as *D. pseudoobscura* had already been established to exhibit responses to environmental fluctuations via inversions, thus these species were classified as genetically “flexible”. A few years later this observation was revised by Sperlich and Feuerbach (1966) who characterized *D. subobscura* inversion polymorphism as “semirigid” or “semiflexible”. The aforementioned classification suggests that while some inversions seem to respond to environmental changes others appear quite stable. Currently, there is extensive evidence to support that polymorphic inversions in *D. subobscura* can respond to certain environmental fluctuations. This behavior has been observed for a number of polymorphic inversions over short, mid and long-term shifts (Rodríguez-Trelles & Rodríguez, 1998, 2010; Rodríguez-Trelles *et al.*, 2013; Balanyà *et al.*, 2006).

Polymorphic inversions in *D. subobscura* are shown to correlate with geographic gradients (also referred to as clines). By geographic clines we refer to latitudinal and/or altitudinal clines. Powell (1997) had already described inversions in different *Drosophila* species that exhibit responses to such geographic clines. Up-to-date populations of *D. subobscura* originating from the Palearctic region illustrate responses to a distinct latitudinal component via fluctuations of inversion frequencies (Menozzi & Krimbas, 1992). Similar responses to latitudinal clines were revealed in the American populations of *D. subobscura*. These populations were established by a founder event in the last century, yet the inversion frequency fluctuation to the latitudinal clines follows the same patterns as in the ancestral population (Prevosti *et al.*, 1988; Balanyà *et al.*, 2003; Rego *et al.*, 2010). The recorded inversion clinality cannot be solely explained by genetic drift in such short evolutionary time. This observation is reinforcing the hypothesis regarding the adaptive nature of inversions. Regarding altitudinal clines, currently almost no responses associated with altitude have been uncovered, with the sole exception of J_{ST} which is thought to be increasing in frequency at higher altitudes (Burla *et al.*, 1986).

Besides the aforementioned geographic clines, temporal clines, either as seasonal clines (short-term) or long-term clines have been described for the system. Seasonal shifts in inversion frequencies were initially identified in *D. pseudoobscura* (Dobzhansky, 1948). Less than two decades later similar seasonal clines were observed in *D. subobscura* by Burla and Götz (1965), who pointed out that standard gene arrangements are overall more prevalent in *D. subobscura* for all chromosomes while their frequencies appear to be decreasing in

summer. These observations were conducted using wild populations sampled between 1963 and 1964, in three seasons (spring, summer and autumn) from two sites near Zurich. Further, Fontdevila *et al.* (1983) using natural *D. subobscura* populations sampled in Mount Pedroso over the period of five years identified two O chromosome gene arrangements that show contrasting seasonal frequencies. In particular, the O₃₊₄ inversion complex was established as a warm climate-associated inversion, while the O_{ST} arrangement as a cold climate-associated one, the two inversions were shown to exhibit opposite trends. Rodríguez-Trelles *et al.* (1996) were able to confirm and reproduce this seasonal trend using *D. subobscura* populations that were collected over the period of 15 years in spring, summer and autumn, in Mount Pedroso. Rodríguez-Trelles *et al.* (2013) extended the effort to track seasonal trends by examining inversion frequencies of *D. subobscura* populations sampled over 2011 and 2012 in the Mount Pedroso and in a second locality approximately 600 km eastwards (Berbikiz, Basque Country, Spain). The samplings happened to coincide with a strong heat wave in April of 2011, which allowed them to quantify the intensity of the genetic shift caused by the heat wave. Their findings illustrate increased “warm-climate inversion dose” that resembles typical “warm dose” values of the late summer period. Such findings propose that the rising temperature could be driving adaptive evolutionary shifts in *D. subobscura*.

Long-term clines have been described in *D. pseudoobscura* where inversion frequencies were noted to be shifting over a few decades (Anderson *et al.*, 1991). However, these changes in inversion frequencies cannot be directly associated with any pronounced environmental change. As a result a gradual loss of haplotypic diversity can be observed as certain arrangements appear to be replaced by others. It was speculated that inversions which increase in frequency over time are likely under positive direct selection due to their genetic content. In *D. subobscura* the O₃₊₄ arrangement which has already been associated with adaptation to warm climates appears to be increasing in frequency over time while the O_{ST} arrangement wanes (Rodríguez-Trelles *et al.*, 1996; Rodríguez-Trelles & Rodríguez, 1998; Solé *et al.*, 2002; Balanyà *et al.*, 2006). Overall, it has been observed that there is a rapid replacement (in the genetic composition of *D. subobscura*) towards “southern” chromosomal arrangements in recent samplings, which is often seen as evidence of rapid adaptation to contemporary global warming (Orengo & Prevosti, 1996; Balanyà *et al.*, 2006; Rezende *et al.*, 2010). Yet, whether this evolutionary response is driven solely by the ongoing rise in global temperature remains unclear (Karageorgiou *et al.*, 2020).

1.2.4 *D. subobscura* as a model species, challenges and limitations

To summarize, *D. subobscura* harbors a rich inversion polymorphism and has received special attention due to the parallel adaptive variation patterns across latitude that its five telocentric chromosomes are showing (Ayala *et al.*, 1989). Similar adaptive patterns have been observed across seasons (Fontdevila *et al.*, 1983; Rodríguez-Trelles *et al.*, 1996, 2013), and even through a heatwave (Rodríguez-Trelles *et al.*, 2013). Certain polymorphic inversions’ frequencies seem to be rapidly shifting in close association with the ongoing rise

in global temperatures (Rodríguez-Trelles & Rodríguez, 1998, 2010; Balanyà *et al.*, 2006). The investigation of polymorphic inversions in *D. subobscura* can provide insights regarding the role of inversions in adaptation in a broader frame, and further shed light particularly on the evolutionary responses to contemporary climate warming in *D. subobscura*. Moreover, these observed fluctuations in the inversions' frequencies enable the study of the interaction between demography and selection in the species and the evolutionary forces that shape inversion polymorphisms.

Nonetheless, *D. subobscura* exhibits over 65 polymorphic inversions in all telocentric chromosomes; this peculiarity of its genome architecture allows us to investigate if and how different inversions interact between them to affect adaptation. Given recent advances in genomics, it is now feasible to reconstruct all chromosomal rearrangements in the species, identify and characterize their breakpoints and perform population genomics analysis in order to obtain a comprehensive view of how different polymorphic inversions in a population interact between them. On the same grounds, *D. subobscura* presents an excellent model organism to investigate the selection forces that lead to the maintenance and spread of polymorphic inversions. However, advances in these issues have been hindered by the lack of a reference genome for the species. Here, we aimed to fill this gap by first developing an annotated high-quality reference genome for *D. subobscura*, which we then apply to isolate and characterize the molecular breakpoints of the known-to-be adaptive O₇ chromosomal inversion.

2. Objectives

Recent advances in sequencing technologies and genomics have made it possible to explore genome sequences and to assess the DNA changes and genetic responses directly involved in environmental shifts. The aim of this thesis is to contribute to the assembly of a reference genome for *D. subobscura* and the identification and characterization of fixed and polymorphic inversions. Particularly, we have focused on the *D. subobscura* polymorphic inversion O_7 which is identified as involved in a species' adaptation to contemporary climate.

1. Objective 1. To sequence and assemble a high-quality reference genome for *D. subobscura*.
 - 1.1. To functionally characterize and annotate the genome through an *ab initio* approach.
 - 1.2. To annotate all gene models and compare them to other *Drosophila* species.
 - 1.3. To compare orthologs and gene families between *D. subobscura* and the 13 available *Drosophila* species in order to determine molecular divergence.
 - 1.4. To perform synteny analysis between *D. subobscura* and *D. guanche*.
 - 1.5. To characterize the chromosomal inversions fixed in *Drosophila subobscura*.
 - 1.6. To compare the organization of chromosomes between *Drosophila subobscura* and *D. guanche* during the divergence of the two species.
 - 1.7. To map and characterize the breakpoints of the chromosomal inversions fixed in *D. subobscura*.
 - 1.8. To estimate the time of divergence between *D. subobscura* and *D. guanche*.
 - 1.9. To provide an explanation for the accelerated chromosomal evolution of the *D. subobscura* lineage.
 - 1.10. *D. subobscura* lineage.
 - 1.11. To identify *D. subobscura*-specific genes that could be under positive selection and lineage-specific/orphan genes that might be involved in adaptation.

2. Objective 2. To identify and functionally characterize the breakpoints of the O_7 inversion in *D. subobscura*.
 - 2.1. To assemble a high-quality genome using a *D. subobscura* line isogenic for $O_{\underline{3+4+7}}$.
 - 2.2. To annotate all gene models of the $O_{\underline{3+4+7}}$ genome.
 - 2.3. To perform synteny analysis for the isolation of the O_7 breakpoints.
 - 2.4. To characterize the O_7 breakpoints.
 - 2.5. To functionally annotate the O_7 breakpoints.
 - 2.6. To determine the molecular mechanism of formation of the O_7 inversion.
 - 2.7. To investigate the functional effects of the O_7 breakpoints.
 - 2.8. To unravel the selective factors driving the adaptive evolutionary shifts in the frequency of the O_7 inversion.
 - 2.9. To provide clues regarding $O_{\underline{3+4+7}}$ inversion's role in thermal adaptation.

3. Results

3.1 Chapter 1: Long-read based assembly and synteny analysis of a reference *Drosophila subobscura* genome reveals signatures of structural evolution driven by inversions recombination-suppression effects

RESEARCH ARTICLE

Open Access



Long-read based assembly and synteny analysis of a reference *Drosophila subobscura* genome reveals signatures of structural evolution driven by inversions recombination-suppression effects

Charikleia Karageorgiou*, Víctor Gámez-Visairas, Rosa Tarrío* and Francisco Rodríguez-Trelles*

Abstract

Background: *Drosophila subobscura* has long been a central model in evolutionary genetics. Presently, its use is hindered by the lack of a reference genome. To bridge this gap, here we used PacBio long-read technology, together with the available wealth of genetic marker information, to assemble and annotate a high-quality nuclear and complete mitochondrial genome for the species. With the obtained assembly, we performed the first synteny analysis of genome structure evolution in the *subobscura* subgroup.

Results: We generated a highly-contiguous ~ 129 Mb-long nuclear genome, consisting of six pseudochromosomes corresponding to the six chromosomes of a female haploid set, and a complete 15,764 bp-long mitogenome, and provide an account of their numbers and distributions of codifying and repetitive content. All 12 identified paracentric inversion differences in the *subobscura* subgroup would have originated by chromosomal breakage and repair, with some associated duplications, but no evidence of direct gene disruptions by the breakpoints. Between lineages, inversion fixation rates were 10 times higher in continental *D. subobscura* than in the two small oceanic-island endemics *D. guanche* and *D. madeirensis*. Within *D. subobscura*, we found contrasting ratios of chromosomal divergence to polymorphism between the A sex chromosome and the autosomes.

Conclusions: We present the first high-quality, long-read sequencing of a *D. subobscura* genome. Our findings generally support genome structure evolution in this species being driven indirectly, through the inversions' recombination-suppression effects in maintaining sets of adaptive alleles together in the face of gene flow. The resources developed will serve to further establish the *subobscura* subgroup as model for comparative genomics and evolutionary indicator of global change.

Keywords: Genome structure evolution, Inversion originating mechanisms, Inversion fixation and polymorphism, Spatiotemporally fluctuating selection, Adaptation, Global change

* Correspondence: charikleia.karageorgiou@uab.cat; rosamaria.tarrio@uab.cat; franciscojose.rodrigueztrellles@uab.cat

Grup de Genòmica, Bioinformàtica i Biologia Evolutiva (GGBE), Departament de Genètica i de Microbiologia, Universitat Autònoma de Barcelona, Bellaterra, Barcelona, Spain



© The Author(s). 2019 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated.

Background

Drosophila subobscura Collin [1] is a fruitfly species of the *obscura* group of the subgenus *Sophophora* endemic to, and common in Europe and the western Palearctic, where it spans over thirty latitudinal degrees commonly associated to forest fringes, from sea level to the timber line [2]. The species was found to be unusual among *Drosophila* because it is entirely monandrous [3–5], does not mate in the absence of light [6, 7], and does not produce courtship-song by wing vibration [8].

The rise of *D. subobscura* to its current status as model organism for biological research owes to a long-held effort to understand the genetics and evolutionary biology of the species [9]. Early investigations on its salivary gland nucleus revealed that it has the ancestral *Drosophila* karyotype of a small dot and five large acrocentric rods, does not show a chromocenter [10] and, especially, shows extraordinary levels of chromosomal polymorphism caused by large, cytologically visible paracentric inversions segregating on all five rods. Elaboration of detailed polytene drawings [11, 12] and photomaps [13–15] greatly facilitated the study of the inversions, and paved the way for subsequent development of the over 600 linkage [16] and cytologically mapped genetic markers presently available, which cover most of the euchromatic genome [17, 18].

Besides nuclear genetics studies, obtention of the first restriction, and at present the only map available for the *D. subobscura* mitogenome [19] allowed to identify an intriguing geographical pattern of variation with two major mitotypes, named I and II, that segregate at nearly equal frequencies in most populations, and which have associated measurable differences in fitness-related traits [20, 21].

The discovery of the two Macaronesian island-endemic species *D. guanche* Monclús [22], from the Canarian archipelago, and *D. madeirensis* Monclús [23], from Madeira allowed new possibilities for comparison. Together with *D. subobscura*, they form the *subobscura* subgroup [24]. The three species are isolated reproductively from each other, except for *D. madeirensis* and *D. subobscura* [25, 26], which are capable of limited gene exchange in collinear genomic regions not affected by inversions [27]. Hybrid males show extra sex combs, among other anomalies whose genetic basis and role in species formation has only begun to be elucidated [28–30]. Interestingly, the two island endemics show differences in gene ordering between them, and with respect to *D. subobscura*, but are thought to be monomorphic for inversions.

Of the various features of the *D. subobscura* model, its rich inversion polymorphism has received special attention [31]. In total, more than 65 inversions have been identified, which range in length from ~ 1 Mb (e.g. inversion E₂₀) to as long as ~ 11 Mb (O₇). They include both simple and multiple overlapping inversions on the same

chromosome, which appear strongly associated into about 90 different chromosomal rearrangements [9, 32]. All combined, structurally segregating regions represent approximately 83% of the species genome. The inversions are nonrandom as to their lengths and distribution of breakpoints along the chromosomes, with cytological evidence of multiply reused breakpoints in 26 cases (~ 20% [9]). Recently, breakpoint nucleotide sequences were determined for nine polymorphic inversions using in situ hybridization and chromosome walking methods, which found one case of breakpoint reuse, and overall supported a mechanism of inversion formation through chromosomal breakage and repair by non-homologous end joining, rather than through ectopic recombination [33, 34].

Inspired by the work of Dobzhansky et alia on natural populations of its Nearctic sister basal within the *obscura* group *D. pseudoobscura*, research on *D. subobscura* found the inversion frequencies in all major chromosomes to be highly structured according to both spatial and temporal environmental gradients. Specifically, chromosomal polymorphisms vary geographically between cold and warm climates [35], with genomewide warm climate inversion frequencies peaking in summer and dropping in winter repeatedly every year (and the reciprocal for the cold climate arrangements) [36]. The introduction, rapid spread, and successful establishment of *D. subobscura* throughout the southern Neotropical [37] and western Nearctic [38] regions, from few colonizers [39], in contemporary time [40] attested for the high dispersal ability and potential for local adaptation of the species [41]. The establishment of latitudinal patterns of the same sign across three separate territories [42] which, additionally, stood in contrast with the uniformity found for neutral nucleotide markers [39], further corroborated the adaptive significance of the chromosomal polymorphisms. On top of these patterns, southernmost populations of the species were found segregating for a sex-ratio distorting drive arrangement, whose carrier males have offspring consisting of only or mainly females [43, 44]. The realization that the frequencies of cold climate karyotypes are declining with the globally rising temperatures [45–47] expanded the interest on the species as indicator of evolutionary effects of contemporary global-warming [48–50]. In fact, the standing inversion variation, maintained by the spatiotemporally fluctuating thermal environment allowed a rapid genomewide evolutionary response in a time scale as short as “few days” during a sudden heatwave [51].

Although the recombination-suppression effects of inversions may not suffice to suppress gene flow in the inverted regions entirely [52, 53], it is strong enough to cause nucleotide variation in *D. subobscura* to be extensively structured in regions affected by the rearrangements [54], and to allow evolution of genomic islands of

concerted evolution of ecologically-relevant gene families like *Hsp70* [55]. In the wild, inversions covariate with life-history and fitness-related traits [9]. Until now, however, attempts to reproduce observed spatiotemporal patterns of inversions and their phenotypic associations under laboratory conditions have been largely unsuccessful [56, 57].

Many of the above and other findings would not have occurred without the previous development of the *cherry-curved* (*ch-cu*) recessive marker- [16] and the *Varicose/Bare* (*Va/Ba*) balancer-strains [58]. Motivation to use *D. subobscura* as a model to continue research on central issues of evolutionary biology is, however, presently hindered by the lack of a reference genome for the species. Recently, one step to narrow this gap was taken with the publication of a short-read second-generation Illumina-based genome of *D. guanche* [59]. In this paper, we took an additional step using flow-cytometry and long-read third-generation single-molecule real-time (SMRT) PacBio technology, together with the available wealth of genetic marker and synteny data, to assemble and annotate a high-quality nuclear and complete mitochondrial genome for *D. subobscura*, from our laboratory stock of the *ch-cu* strain. Long-read based assemblies are advantageous over short-read based ones because they are better at traversing across common repetitive structures, which results in more contiguous and complete assemblies. Our goals were two-fold. First, to provide a preliminary account of main features of the newly assembled genome and, second, to perform a comparative synteny analysis with *D. guanche* to trace the evolutionary history of fixed chromosomal rearrangement in the *subobscura* subgroup. Until now, this latter issue has been approached using wholly cytological methods [14, 25, 60] which are coarse-grained compared to the single-nucleotide resolution furnished by comparative genomics.

Knowing the sequence identity of synteny breakpoints can help determine both the evolutionary polarity of chromosomal rearrangement states by comparison with an outgroup, and the mechanism of rearrangement formation through assessment of remains of its molecular footprints. *Drosophila* inversions are commonly thought to originate by one of two major mechanisms, namely ectopic recombination, and chromosomal breakage and subsequent repair (reviewed in [61]). The first mechanism predicts occurrence of duplications on the flanks of the inverted segment in both the ancestral and the derived arrangement states, whereas the second predicts absence of duplications or their presence only in the derived state. Knowing how an inversion originated can shed light on why it evolved [62]. Inversions can have direct, indirect, or both types of fitness effects. New inversions can themselves be direct targets of selection because of functional disruption by the breakpoints. The

main importance of inversions, however, might stem indirectly from the fact that they suppress recombination in heterokaryotypes. Through their linkage generation effects, inversions can contribute to keep sets of adapted alleles together in the face of gene flow [63–65].

Results and discussion

Size estimation and de novo long-read assembly of the *D. subobscura* genome

The genome size of the inbred *ch-cu* line was estimated using *k*-mer counting and flow cytometry methods. By the first method, GenomeScope (<http://qb.cshl.edu/genomescope/> [66]) analysis of 21-mer frequencies obtained by Jellyfish (Ver. 2.2.4. [67]) using 20 million Illumina short (300 bp) reads [55] resulted in a genome size of 136.943 Mb. By the second, flow cytometry of PI-stained female brain cell nuclei using a 328.0 Mb genome from *D. virilis* [68] as internal standard resulted in a genome size of 148.069 Mb (0.151 pg \pm 0.001; for the mean plus/minus one standard deviation across five replicates; see Methods). This latest measure conforms to previous flow cytometry-based estimates of the *D. subobscura* genome size (146.7 Mb [69, 70]); rounded to 150 Mb, it was the value set as genome size for the Canu assembler.

The PacBio 7 SMRT cells sequencing of the *ch-cu* genome generated a raw output of 1,252,701 subreads, hereon referred to as reads, with mean and longest read lengths of 8003 bp and 52,567 bp, respectively (Additional file 1: Table S1). These sequences totaled 10,025,366,103 bp, or a \sim 67-fold estimated genome coverage. The average yield per SMRT cell (\sim 1.4 Gb) was on the upper bound of the manufacturer range for the typical SMRT cell (0.75–1.25 Gb [71]), which highlights the suitability of the used high-quality genomic DNA isolation protocol. Canu correction and trimming of the PacBio data retained 1,060,943 reads of 6103 bp average read length, or the equivalent to a 43-fold genome coverage for the assembly, well within Canu's default sensitivity range (30-fold to 60-fold) (Additional file 1: Table S1). Of the 327 Canu-generated contigs, 115 (totaling 6624 Mb) showed evidence of foreign sequences. All the contigs in this subset were solely of bacterial origin, each being exclusively either from *Acetobacter* or from *Providencia*, which are genera known to be part of the *Drosophila* microbiome [72]. After removing these contigs, the primary Canu assembly consisted of 212 contigs spanning 129.183 Mb, with an N50 of 3.129 Mb and a maximum contig length of 15.083 Mb (Additional file 1: Table S1).

A first round of quality control and scaffolding of the Canu contigs carried out combining recursively i) automated BLAT and BLASTN searches against the *D. melanogaster* and *D. pseudoobscura* genomes, and ii) evaluation of consistency with published data on the

chromosomal position of 621 cytological (604) and genetic linkage (17) markers (Additional file 2: Table S2; see Methods) did not detect any misassembling. Scaffolding of the Canu contigs using SSPACE-LongRead (Ver. 1–1. [73]) resulted in 157 scaffolds. Submission of these scaffolds to a second round of quality control and scaffolding as in step one resulted in 186 validated scaffolds with a total length of 129,237 Mb (Additional file 1: Table S1). Half of the assembly was in 7 scaffolds longer than 5.954 Mb, while an additional 45% was in 44 scaffolds longer than 313 Kb. The GC content of the assembly was 45.0%, similar to that found for the close relative *D. pseudoobscura* (45.3%; r3.04 assembly [74]). Based on the available cytogenetic and genetic linkage marker data, it was possible to assign confidently genomic coordinates to 96.6% of the assembled sequence (63 scaffolds spanning 124,862 Mb, with half of it in 6 scaffolds longer than 8.237 Mb). On average, there were 10 markers per scaffold. A total of 38 scaffolds, representing 91.4% of the positioned sequence, were placed using ≥ 2 markers. The remaining 25, relatively shorter scaffolds with only 1 marker (10; average length 0.656 Mb) or 0 markers (15; 0.363 Mb) were placed confidently aided by synteny-based inferences of orthology with the close relative *D. guanche* and/or with *D. melanogaster*. Detailed information about the markers used for the anchoring, ordering and orientation of the scaffolds is provided in Additional file 2: Table S2.

The final assembly resulted in six chromosome-sized pseudomolecules or pseudochromosomes, one for each of the six chromosomes of the haploid *ch-cu* female chromosome set (Table 1 and Additional file 1: Table S1; Fig. 1). The pseudochromosome dot consisted of a single contiguous sequence 1.376 Mb long; the A incorporated 24 scaffolds spanning 22.858 Mb (the largest being 11.265 Mb long; coordinates assigned based on 123 markers); the J, eight scaffolds spanning 23.583 Mb (15.120 Mb; 45); the U, five scaffolds spanning 25.800 Mb (11.275 Mb; 21); the E, seven scaffolds spanning 20.819 Mb (8.237 Mb; 293); and

the O, 18 scaffolds with combined size of 30.426 Mb (8.841 Mb; 140). The number of scaffolds is greater for chromosome A than for the autosomes, probably because we sequenced genomic DNA from a pool of 50:50 males and females, such that the A would be expected to have three-quarters the sequence coverage of the autosomes. The lengths of the pseudochromosomes show nearly perfect correlation with the linear lengths of the corresponding polytene chromosomes measured from the Kunze-Mühl and Müller [12] reference map (Pearson's $r = 0.99$; $P < 10^{-4}$). While the rest of the assembly not included in the pseudochromosomes (3.4%; 4.375 Mb in 123 scaffolds) could not be assigned precise genomic coordinates owing to non-availability of reliable positioning information, for most of it (81.2%; 3.551 Mb in 90 scaffolds) it was possible to at least anchor it to chromosomes (including the rDNA chromosome) using similarity search tools (Additional file 1: Table S1). Only 0.6% (0.824 Mb in 33 scaffolds) of the assembly remained completely unplaced. This included cases where either there was no marker/synteny data available, or the placement of the corresponding BLAT/BLASTN hits in the reference species is unknown.

Ab initio gene prediction and functional annotation

The complete *ch-cu* assembly was predicted to contain 13,939 protein-coding genes, nearly the same number as in the current release of the *D. melanogaster* genome (13,931; r6.18 assembly [75]). Of them, 13,317 (95.5%) were successfully annotated by the MAKER annotation pipeline, which corresponds to a gene density of one gene every 9.70 kb of the genome assembly. The average gene length was 3.502 kb. All genes combined span 46.635 Mb of coding sequence, with a GC content of 55.6%. The average number of exons and introns per gene was 4.6 and 3.6, with average (median) exon and intron lengths of 379 (213) bp and 529 (66) bp, respectively. A total of 87.2% of the genes were multi-exonic.

Of the 13,317 annotated protein-coding genes, 13,181 (99.0%) are placed in the six pseudochromosomes that

Table 1 Overview of *D. subobscura* nuclear pseudochromosome and mitochondrial reference genome assembly (N50: length of the contig for which 50% of the total assembly length is contained in scaffolds of that size or larger; L50: ranking order of the scaffold that defines the N50 length; lengths are in bp)

Component	Length	No. of scaffolds	Largest scaffold	L50	N50	Gene models	% repetitive
Nuclear	124,861,819	63	15,119,984	6	8,236,782	13,181	11.7%
Dot	1,375,632	1	1,375,632	1	1,375,632	91	28.9%
A	22,857,882	24	11,265,230	2	1,077,607	2322	14.6%
J	23,583,473	8	15,119,984	1	15,119,984	2452	11.1%
U	25,800,175	5	11,274,558	3	9,313,524	2496	10.3%
E	20,818,511	7	8,236,782	2	5,954,457	2591	11.3%
O	30,426,146	18	9,011,354	3	4,063,992	3229	10.6%
Mitogenome	15,764	1	15,764	1	15,764	37	

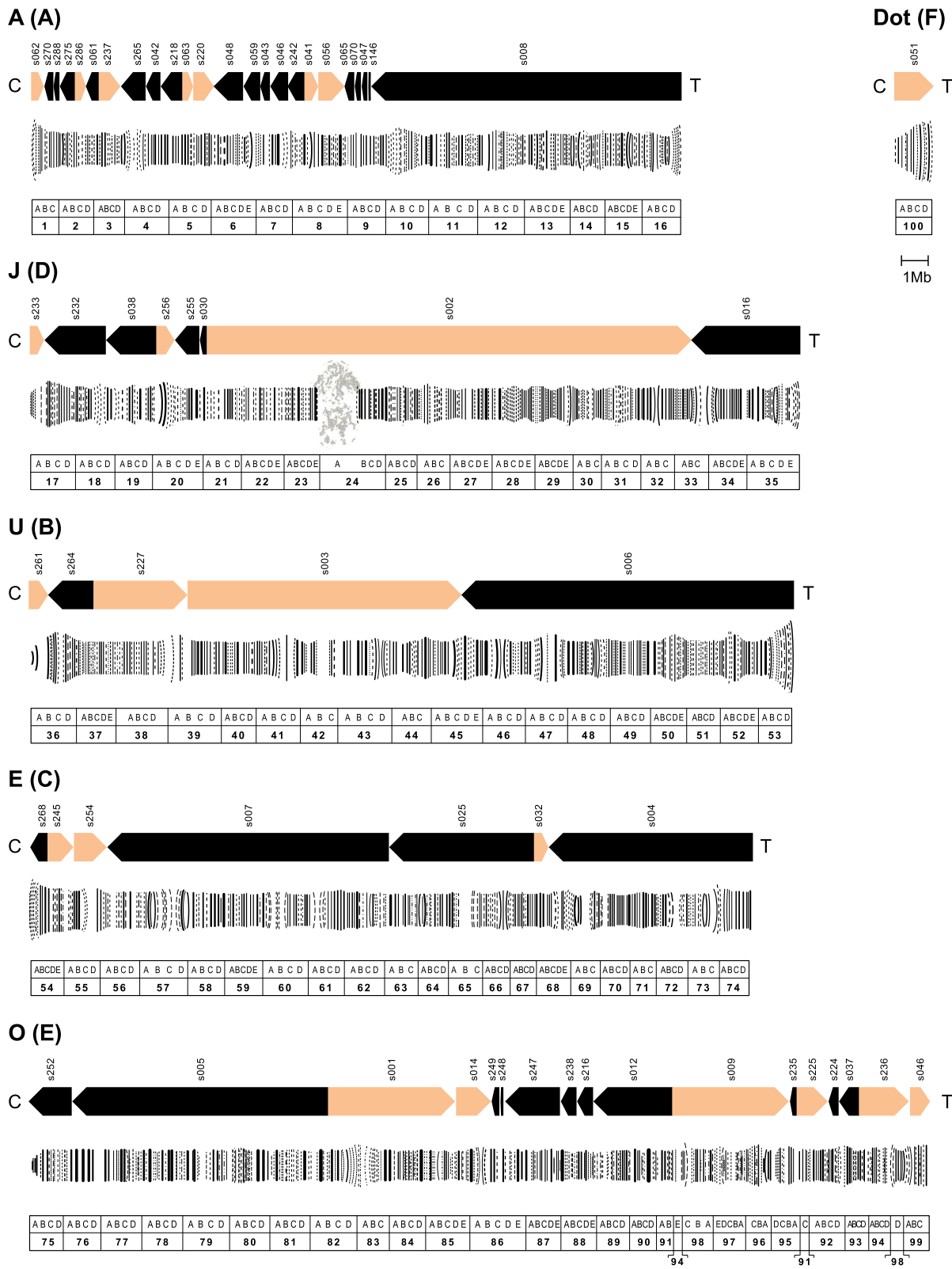


Fig. 1 (See legend on next page.)

(See figure on previous page.)

Fig. 1 De novo assembly of a *D. subobscura* genome from long-read PacBio sequencing data. The six chromosomes are referred to by their corresponding letter (i.e., A, J, U, E, O and dot) and Muller element (i.e., A, D, B, C, E and F, respectively; in parentheses) designations. Chromosomes are shown oriented from centromere (C) to telomere (T). Each chromosome panel includes (top) a scheme of the reconstructed pseudochromosome and their component forward (sepia) and reverse (black) scaffolds with labels (e.g., s062) on them; (center) a drawing of the Kunze-Mühl and Müller [12] reference standard karyotype, modified to take into account that the *ch-cu* strain used for genome sequencing is structurally O_{3+4} (or O_{ms+4} ; see the results and discussion section) and (bottom) a ruler indicating the sections (from 1 to 100) and subsections (each from A to E) of the Kunze-Mühl and Müller [12] map. A 1 Mb-scale bar is shown below the dot

were assigned genomic coordinates (Table 1). The numbers of annotated genes per pseudochromosome (dot: 91; A: 2322; J: 2452; U: 2496; E: 2591; and O: 3229) depart from the expected from pseudochromosome length ($G = 113.61$; d.f. = 5, $P < 10^{-6}$), with E and U showing, respectively, the greatest excess (393 genes) and deficiency (228), in line with previous findings in *D. melanogaster* [76]. With respect to the small subset of genes that could not be assigned precise genomic coordinates (139), most of them (89) were anchored to chromosomes. In addition, 3090 non-coding genes were annotated, including 1191 and 1899 short and long non-coding RNA genes, respectively. Of note, the 5S rDNA gene family was found to consist of > 160 copies of the 5S rDNA repeat unit, tandemly arranged in one cluster located on the distal end of segment II of autosome O, in agreement with early in situ hybridization results [77]. Also, we identified > 80 copies of the 18S–28S rDNA repeat unit distributed over the 19 rDNA annotated scaffolds. With respect to the relatively more rapidly evolving lncRNA genes, BLASTing with Fly-Base lncRNA (Dmel_Release_6) detected 1898 out of the 2965 lncRNA annotated genes, with a strong bias towards the longer ones (10.2 kb vs. 1.2 kb, for the average lengths of detected vs. undetected lncRNAs, respectively).

The high-quality of the genome assembly and annotation is further buttressed on three validation metrics. Firstly, the overall size of the assembly (129.237 Mb) closely matches the estimated size of the genome using the *k*-mer counting (94.4% of 136.943 Mb) and flow-cytometry (87.3% of 148.069 Mb) methods. Secondly, both the low values of the average and median of the MAKER-defined AED scores (0.127 and 0.070, respectively), and the fact that nearly all genes attained AED scores lower than 0.5 (AED₅₀ = 97.9%) are indicative of a good agreement between the annotations and their evidence. And thirdly, BUSCO analysis using the 2799 25-dipterans orthologous gene set resulted in 96.5% (2671) single complete genes, 0.5% (14) duplicated complete genes, and 3.0% (84) fragmented. Only 1.1% (30) of the BUSCO genes were missing, indicating that the assembly is nearly complete.

Phylogenetic placement of the *D. subobscura* genome and age of the *subobscura* subgroup

To further assess the quality of the obtained genome, we subjected it to a phylogenetic analysis together with

closely related species with known relationships. We took advantage of the carefully curated 12 *Drosophila* multiple sequence alignment (MSA) data set used by Obbard et al. [78] (see also [75]). The MSA consists of 67,008 characters from 50 concatenated nuclear protein-coding loci selected for (i) having only 1:1 orthologs, (ii) including an exon longer than 700 bp, and (iii) not showing unusually high codon usage bias. To this MSA, we added the corresponding reciprocal-BLAST-identified orthologs from *D. subobscura* and *D. guanche* using MAFFT (Ver7; <http://mafft.cbrc.jp/alignment/software/>), and then identified the best-fit model of sequence evolution (GTR + G + I; with $\alpha = 0.53$, and $I = 0.27$) for maximum-likelihood (ML) tree estimation using MEGA7 [79]. The resulting tree topology (Additional file 3: Figure S1) is consistent with the known phylogeny of the species. Using this topology, and the RelTime-ML method [80] with the mutation rate-based estimates found by Obbard et al. [78] to perform best as calibration dates, the age of the *subobscura* species subgroup was found to be 1.72 ± 0.51 Mya (Additional file 3: Figure S1). This estimate is at the lower bound of published dates for this divergence, which were all based on one or few available markers (ranging from 1.8 to 8.8Mya, median 2.75Mya [27, 81–84]).

Mitochondrial genome identification and annotation

BLASTN searches against the *ch-cu* assembly found that Canu's tig00002375 contained a complete copy of the *D. subobscura* mitogenome. The mitogenome is 15,764 bp long, and shows the gene number, order and orientation of the typical insect (Table 1 and Additional file 4: Table S3; Fig. 2 [85]), including 13 PCGs (*ND1–6*, *COI-III*, *ND4L*, *Cytb*, *ATP6*, *ATP8*), 2 ribosomal RNAs (*lrRNA* and *srRNA*), 22 tRNAs, and an AT-rich region (control region). The control region is 944 bp long, and is placed between genes rRNAs and tRNAI. The nucleotide composition is biased towards A + T (78.3%), the bias being greatest in the control region (93.0%). The plus strand codes for 23 genes (9 PCGs and 14 tRNAs) and the control region, while the minus strand codes for the remaining 14 genes (4 PCGs, 8 tRNAs and 2 rRNA genes). All PCGs start with the typical ATN codons, except *COI* that starts with a TCG codon, and terminates with the TAA/TAG codons, except *COII* and *ND5* that

less repetitive genome is further supported by available measures of genome size obtained using flow cytometry from brain cell nuclei: its genome is nearly 30 Mb smaller than that of *D. guanche* (167.230 Mb), and the smallest of the ten *obscura* group species values stored in the animal genome size database [70, 89].

Repetitive DNAs were analyzed by classifying them into five categories: long terminal repeat (LTR) and non-LTR retrotransposons, DNA transposons, satellites (including *sat290* and *SGM-sat*), and simple repeats or microsatellites. The extrinsic null of no deviation in repeat number content from the expected from relative chromosomal length was tested using G-tests among all six chromosomes, between A and the four large autosomes, and among the four large autosomes. Overall, there were significant differences in proportion of repetitive sequence among the six chromosomes, whether repeats were considered together or separately by category (all G-tests: $P < 10^{-6}$). The dot showed the largest aggregated excess (2.5-fold; 2.7% of total repeat number content), because it showed 2.6 (3.0%), 4.1 (4.8%) and 4.9-fold (5.7%) more non-LTRs, DNA transposons and satellites than expected from its length, whereas A was the only chromosome that showed a consistent excess of repetitive sequence across all five repeat categories, particularly microsatellites (1.5-fold; 26.9%). Comparatively, the four large autosomes showed a dearth of repetitive DNA. When the dot and the A were excluded from the analysis the magnitude of the deviations in amount of repetitive sequence dropped markedly [with the single exception that the E chromosome shows a 1.3-fold (27.7%) excess of non-LTRs], and no definite pattern emerged.

The distribution of repetitive DNA densities along chromosomes is shown in Fig. 3. For simplicity, non-LTR and LTR retrotransposons, DNA transposons and satellites were aggregated into a single class separately from microsatellites. The two classes differ qualitatively in their patterns of chromosomal distribution. Transposable elements and satellites appeared concentrated in the pericentromeric and (less so) peritelomeric regions. This was so particularly for DNA transposons and satellites, and the pattern became most apparent for the J and U chromosomes. In addition, there were large megabase-scale regions with high density of DNA transposons and satellites in the A and O chromosomes. Interestingly, the distal-most peak of repetitive sequence in chromosome A, in fact consists of telomeric sequence that was repositioned to that location by inversion A_6 (see below). Microsatellites deviate markedly from this pattern, showing nearly monotonic trends to increasing density towards the telomeres, that became statistically significant for the J, U and E chromosomes (simple linear regressions: $r^2 = 0.73$, $P < 10^{-5}$; $r^2 = 0.25$, $P = 0.009$;

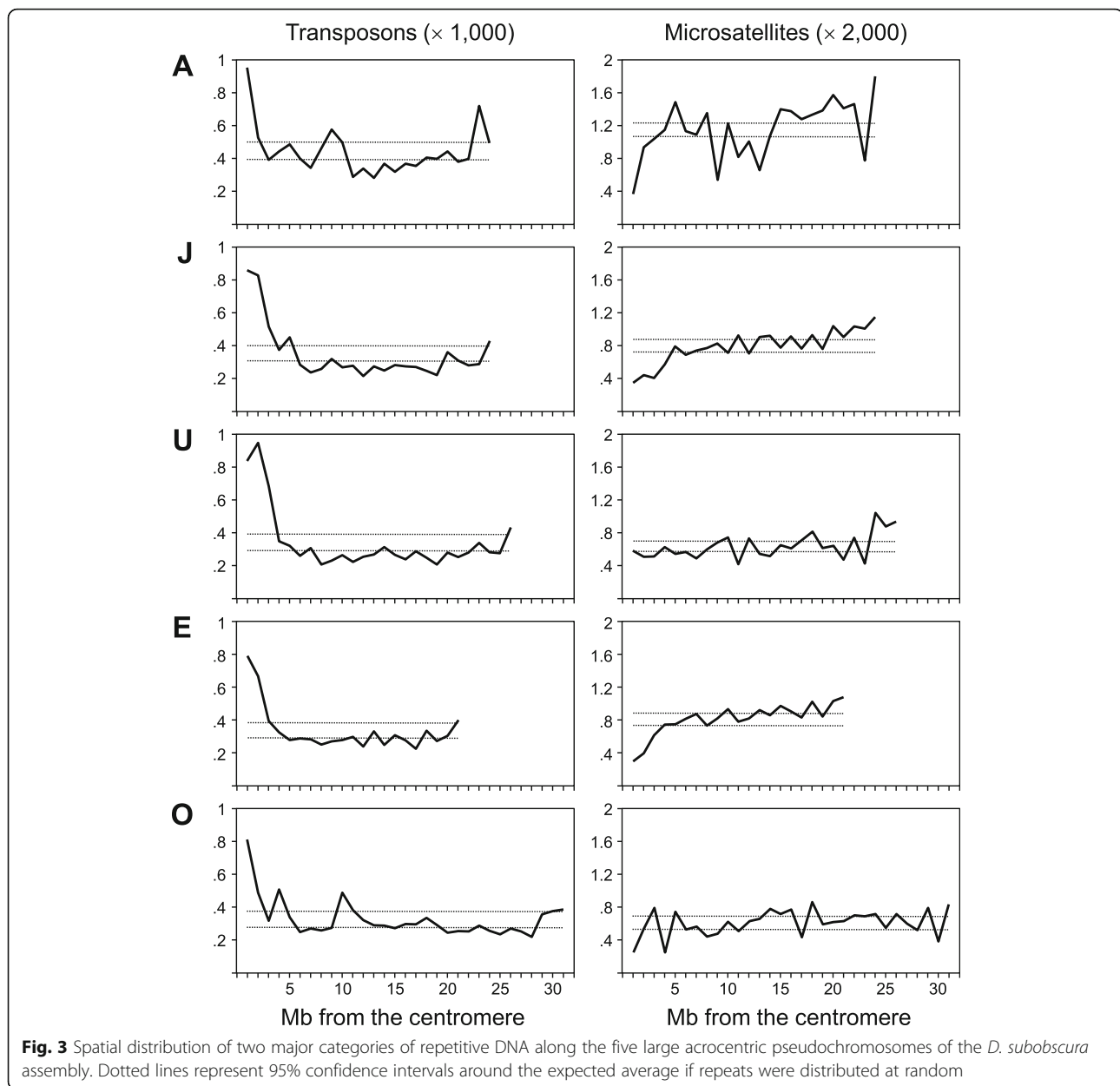
and $r^2 = 0.65$, $P < 10^{-4}$, respectively). Understanding the significance of these differences warrants further in-depth analyses.

Satellites *Sat290* and *SGM-sat* have gathered special interest. *Sat290* is a 290 bp repeat satellite [90]. Early in situ hybridization studies in the three members of the *subobscura* subgroup concluded that *sat290* was absent in *D. madeirensis* and *D. subobscura*, and that in *D. guanche* the repeat comprised a major satDNA class distributed in centromeric regions [90]. *SGM-sat* would be derived from the MITE-like *SGM-IS* transposable element that was already present in the last common ancestor of the *obscura* group [91]. The repeat underwent a species-specific expansion in *D. guanche*, which gave rise to another major satDNA class in this species. Some of these findings were reassessed by a recent study of the *D. guanche* genome combining Illumina short-read whole genome sequencing and dual-color fluorescence in situ hybridization [59]. The study found *sat290* and *SGM-sat* to comprise the first and second most abundant satDNAs of *D. guanche*, respectively, adding up to ~30% of the species' genome. In addition, *SGM-sats* were found to be concentrated in the centromeres, but in more peripheral positions relative to the chromosome ends than *sat290s*. In contrast with this picture, our initial characterization of these repeats in the *ch-cu* assembly showed that *sat290* is present in *D. subobscura*, and in non negligible numbers (637), of which nearly one half are dispersed throughout the euchromatin. *SGM* showed a similar pattern, but conversely to the situation in *D. guanche*, in *D. subobscura* *SGM* sequences are 8-fold more abundant than *sat290s* (Additional file 5: Table S4).

Overall, our preliminary screen of the genomic distribution of repetitive DNAs did not find evidence of an association between repeat density and numbers of segregating chromosomal rearrangements. For example, the J and E chromosomes, which are about the same size show comparable percentages and distributional patterns of repetitive sequence (~11.0%; Fig. 3), in spite that the former exhibit 4-fold lower number of polymorphic inversions than the latter (5 vs. 22, respectively [9]).

Orthologous group assignment and variation in gene family size

OrthoMCL clustered the 152,068 PCGs in the *Drosophila* pan-genome dataset into 23,394 orthologous groups, of which 8390 (35.9%) formed the core set shared by all 14 species (Additional file 6: Figure S2). Of this core set, 6293 were single-copy gene families. *D. subobscura* contained 10,483 orthologous groups, including 904 (8.6%; 965 genes) lineage-specific, of which 867 were single-copy orphans. These numbers and categories of orthologous groups are similar to those obtained for its



close relative *D. guanche* in a previous comparison of the same 13 *Drosophila*, excluding *D. subobscura* (10,417 orthologous groups, including 838 species-specific, of which 828 were orphans [59]). Also, the number of orphan genes in *D. subobscura* is within the range of those estimated for the other 13 *Drosophila*, which varies from 294 in *D. erecta* to 2341 in *D. persimilis* (Additional file 6: Figure S2).

CAFE analysis (see the Methods section for details) carried out without taking into account variation in genome quality across genomes indicates that the best description of the data is provided by the five λ model, which distinguishes average fast- (λ_F), medium- (λ_M) and slow-evolving (λ_S) branches, in addition to allowing the

terminal branches leading to *D. subobscura* and *D. guanche* to have their own rates (λ_{Ds} and λ_{Dg} ; Additional file 7: Table S5). According to this model, the rate of gene family size evolution in these two lineages would be of the same order of magnitude as the average fast rate ($\lambda_{Ds} = 0.0257$ and $\lambda_{Dg} = 0.0191$, vs. $\lambda_F = 0.0216$). Adding a global error term (ϵ) improves the model fit significantly ($-2\Delta L = 191,98$; $p < 1 \times 10^{-6}$; 1 *df.*), which indicates an effect of variation in quality across genomes. The effect, measured as the ratio $(\lambda - \lambda_\epsilon)/\lambda$ [92], is lowest for slow-evolving lineages (24%) and *D. subobscura* (24%), and largest for fast-evolving lineages (43%) and *D. guanche* (41%). The best score of *D. subobscura* compared to *D. guanche* according to this criterion may be a

reflection of a greater contiguity of the assembly provided by the PacBio long-read sequencing used in the first case, compared to the Illumina short-read sequencing used in the second.

The CAFE five λ model with global error term indicates that, of the 9155 gene families inferred to have been present in the *Drosophila* most recent common ancestor, 567 have increased and 636 decreased in size in the terminal branch leading to *D. subobscura* (Additional file 8: Figure S3). Of them, 62 show significant expansions (43; 272 genes) or contractions (19; 121 genes) relative to the genome-wide average ($P < 0.01$; Additional file 9: Figure S4). Functional enrichment analysis of the rapidly evolving families showed the expanding and contracting families to be significantly enriched for 52 and 77 GO terms, respectively (Additional file 10: Table S6 and Additional file 11: Table S7). Most-encompassing GO terms associated with the families that have expanded include, among others, ‘*thermosensory behavior*’ (GO:0040040) (Additional files 12, 13 and 14: Figures S5-S7), and those associated with the families that have contracted include ‘*sensory perception of sound*’ (GO:0007605) and ‘*response to red light*’ (GO:0010114) (Additional files 15, 16 and 17: Figures S8-S10). These terms appear particularly noteworthy considering the continuing role of *D. subobscura* as a model for research on insect thermal biology, and that, as previous research has shown, the species may be unique within the *obscura* group in not producing courtship auditory cues by wing vibration [8], and being unable to mate in the dark [6, 7, 93]. We hope that these results will stimulate future research on the role that those gene families play in the functional biology of *D. subobscura*.

Evolutionary history of chromosomal rearrangement in the *subobscura* subgroup

Comparative synteny mapping of the genome of *D. subobscura* with those of three increasingly distant relatives, namely *D. guanche*, *D. pseudoobscura* and *D. melanogaster* using SyMAP showed the amount of genome rearrangement to scale up with evolutionary distance. Aggregated across the five Muller elements, *D. subobscura* synteny with each of the aforementioned species is fragmented into an increasingly larger number of increasingly smaller blocks: 31 blocks of 3.952 Mb average size (13 inverted), 333 of 0.345 Mb (164), and 540 of 0.220 Mb (264), respectively (Additional file 18: Table S8 and Additional file 19: Table S9). Chromosome A shows the greatest degree of synteny fragmentation in all three pairwise species comparisons (12, 90, and 125 vs. 5, 61, and 104 blocks, for A vs. the average autosome), in agreement with reported higher rates of rearrangement evolution for this Muller element compared to the autosomes [65, 94].

Identified synteny blocks between the *D. subobscura* and *D. guanche* genomes have associated 28 breakpoints (11, 2, 4, 5 and 6, for the A, J, U, E and O chromosomes, respectively), of which 25 could be ascribed to 13 large-megabase scale paracentric inversions as shown in Fig. 4. To simplify matters, in that figure and henceforth, we used subindex “a” to denote ancestral arrangements of the species subgroup (except for U_{1+2} , because it is shared by the three species), “g” for inversions fixed in the lineage of *D. guanche*, “ms” for inversions fixed in the most recent common ancestor of *D. madeirensis* and *D. subobscura*, and “h” for hypothetical rearrangement steps invoked to interconvert alternative gene arrangements. Of the 13 rearrangement differences, 6 occurred in chromosome A, including 4 overlapping inversions in its proximal half (A_{h11} - A_{h4}), and 2 single inversions in its distal half (A_5 and A_6); and 1, 2, 2 and 2 inversions in autosomes J (J_{ST}), U (U_1 and U_2), E (E_{g1} and E_{ST}) and O (O_{ms} and O_4), respectively. With respect to the proximal half of chromosome A, 4 overlapping inversions is the minimum number of reversals required to interconvert the gene arrangements of the two species in that region [60]. Figure 4 (upper right) depicts one of those hypothetical paths (in fact, the only one consistent with Ah being the newest; see below) inferred using the algorithm implemented in GRIMM (<http://grimm.ucsd.edu/GRIMM/> [95]), taking into account the ordering and orientation of the observed 9 syntenic blocks. Overall these results corroborate previous cytological ideas as to the number of paracentric inversion differences between the two species [14, 60].

Of those 13 rearrangement differences, nine are thought to be fixed between the two species, including A_{h11} - A_{h3} , A_5 and A_6 , J_{ST} , E_{g1} and E_{ST} and O_{ms} ; three are thought to be fixed in *D. guanche* and polymorphic in *D. subobscura*, including A_{h4} (assumed to be the same as *D. subobscura*’s A_1), U_1 and U_2 ; and one, namely O_4 , it is found only as polymorphic in *D. subobscura* [14] (here, it may be helpful to recall that the *ch-cu* homokaryotypic strain used to represent *D. subobscura* is *standard* for all chromosomes except chromosome O, for which it is O_{3+4} ; see below). For none of these 13 inversions, except O_4 [34], the nucleotide sequences of their breakpoints have been molecularly characterized. Yet this knowledge could allow testing current cytology-based ideas about the identities and evolutionary polarities of the rearrangement states, as well as ascertaining their originating mechanisms through assessment of remains of their molecular footprints.

To further validate the high quality of the newly obtained *D. subobscura* genome, we applied it to determining the unknown breakpoint sequences of the aforementioned 12 inversions as follows (Additional file 20: Figure S11). We defined synteny breakpoint as the nucleotide interval between contiguous SyMAP synteny blocks.

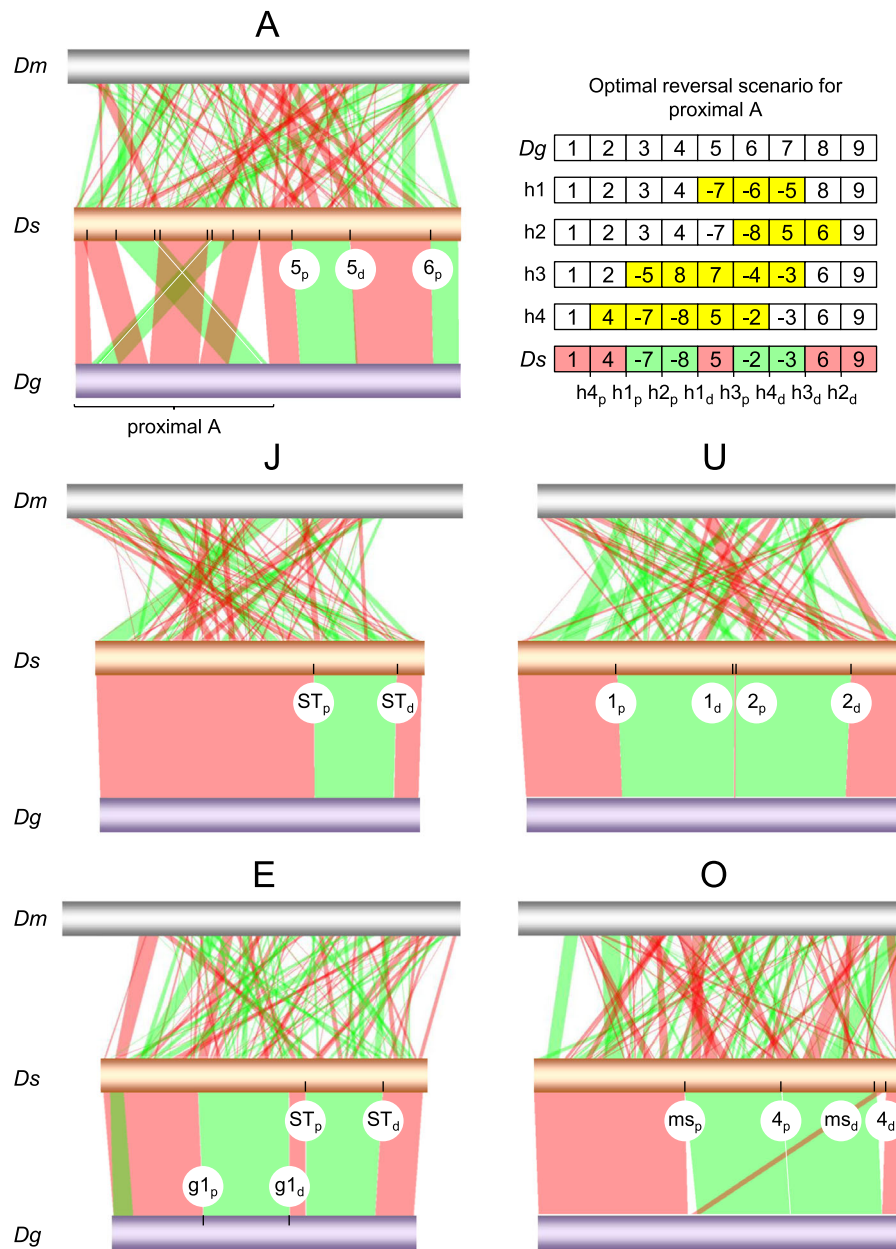


Fig. 4 SyMAP comparative chromosome synteny analysis between *D. subobscura* (*Ds*; central gold horizontal bars) and each of *D. melanogaster* (*Dm*; upper grey) and *D. guanche* (*Dg*; bottom purple). Bands connecting homologous chromosomes denote noninverted (pink) and inverted (green) synteny blocks. Labeled ticks on chromosomes indicate proximal (p) and distal (d) inversion breakpoints. Labels for breakpoints in the proximal region of the A chromosome are provided in the upper right panel of the figure (h1_p to h4_d), along with the optimal reversal scenario for the transition between the standard sequence of *D. subobscura* and the arrangement of *D. guanche* in this region inferred using the GRIMM algorithm. The eight synteny blocks of that transition are designated by positive (noninverted) and negative (inverted) numbers, and the corresponding four intermediate hypothetical inversions (yellow) by letter “h” subscripted 1–4. Cytological map positions and pseudochromosome coordinates of inversions breakpoints are given in Additional file 21: Table S10

Suppose two orthologous gene arrangements A|BC|D and A|CB|D in taxa 1 and 2, respectively, where the second arrangement is identical to the first one, but for the inverted sequence CB, with the vertical lines denoting the inversion breakpoints. If e.g. region A|B, spanning the proximal breakpoint in taxon 1 plus 5 kb towards

the inside of each of its two flanking synteny blocks is BLASTed against the genome of taxon 2, there should produce two hits, one in locus A, and the second one in locus B. In addition, each hit should carry associated an alignment overhang due to lack of homology between B and C, and between A and D, respectively; and the

coordinates of the hits should match the SyMAP coordinates for the breakpoints spanning the inversion. Furthermore, the results from breakpoints of the same inversion must be reciprocally consistent, regardless the taxon used as query.

By the above described approach, we were able to isolate and characterize the putative breakpoint sequences of all the 12 targeted inversions. The results challenge previous cytology-based assumptions about the identity and evolutionary polarity for some of the rearrangement states. Additional file 21: Table S10 and Fig. 4 summarize the main results. The proximal half of chromosome A provides an all-embracing example. In this region, the structural transition between the two genomes requires minimally four inversions (A_{h1} - A_{h4} ; Fig. 4). Cytological evidence for shared breakpoints supporting that A_{h4} is the same inversion as A_1 , led to postulate that A_{h1} - A_{h3} were fixed in the lineage of *D. guanche* [14, 25]. Recently, the proximal and distal breakpoints of inversion A_1 segregating in *D. subobscura* were assessed by a mixed approach combining cytological and molecular methods [96]. Although the attempt was unsuccessful, it managed to narrow them down to within a few kilobases distal to the markers *cm* (CG3035) and *dod* (CG17051), respectively. The coordinates of those markers in the newly obtained *ch-cu*, i.e., A_{ST} genome (chrA:1,206,180 bp and chrA:8,875,126 bp, respectively) lie more than half a megabase proximal to their corresponding nearest breakpoint of the A_{h4} inversion separating the *ch-cu* strain from *D. guanche* (chrA:692,605 bp and chrA:8,198,726 bp; Additional file 21: Table S10). This finding indicates that the previously supposed-to-be same inversion shared by the two species, i.e., A_{h4} equal to A_1 , in fact represents two different inversions that originated separately.

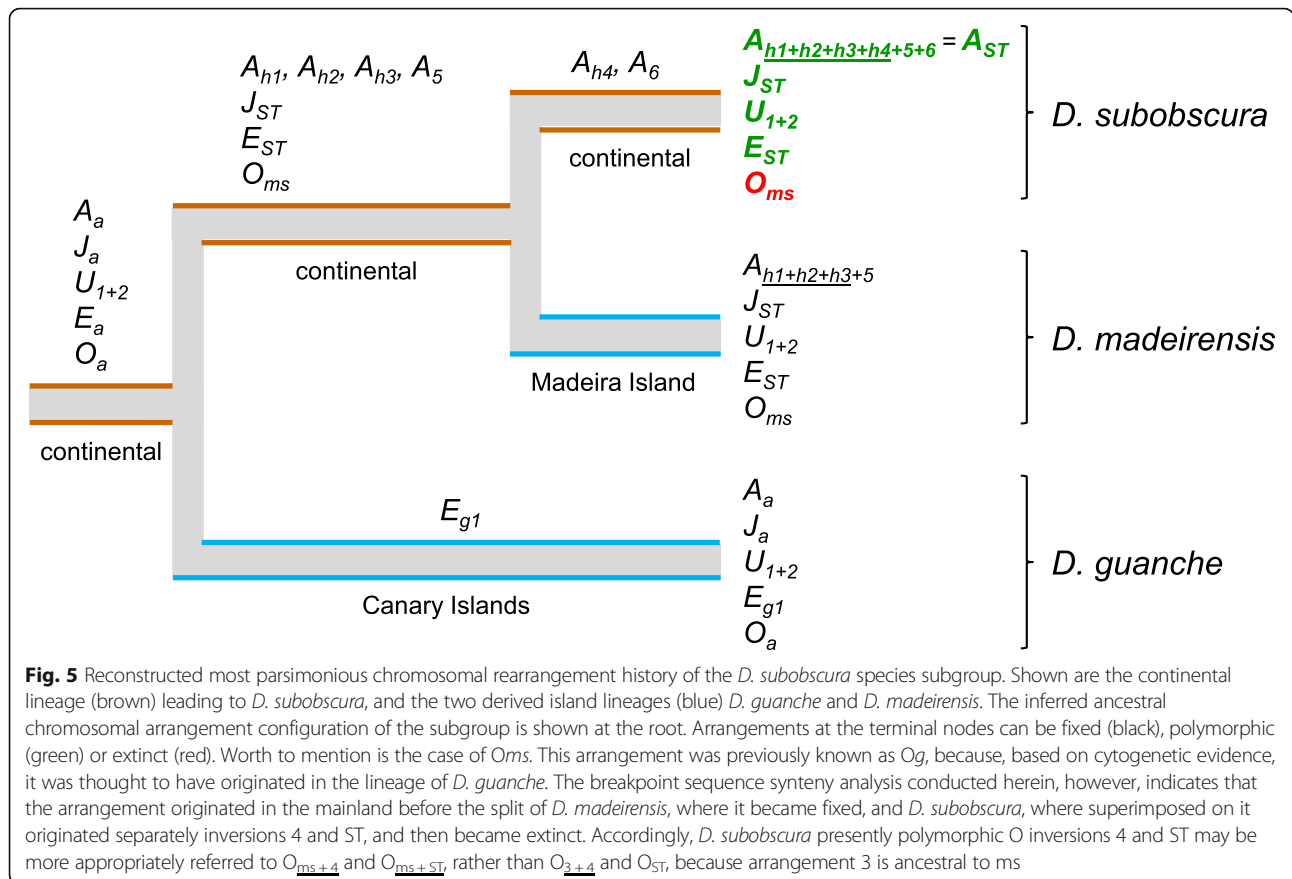
Comparative analysis of gene arrangement of the regions around the breakpoints of A_{h4} in *D. subobscura* and *D. guanche* with those in the outgroup *D. melanogaster* indicates that *D. guanche* shows the ancestral arrangement state (CG2076, CG2081, CG18085, |, CG15203, CG1537, CG1545, and CG32677, CG43347, CG1628, |, CG15302, CG32683, CG2096; for the proximal and distal breakpoints, respectively, in both *D. guanche* and *D. melanogaster*; with the vertical lines denoting the inversion breakpoints; Additional file 21: Table S10), whereas *D. subobscura* shows the derived state (i.e., CG2076, CG2081, CG18085, |, CG1628, CG43347, CG32677, and CG1545, CG1537, CG15203, |, CG15302, CG32683, CG2096; Additional file 21: Table S10). In addition, no evidence was found for duplicated and/or repetitive sequences in the breakpoint regions from reciprocal BLAST searches, which supports that the inversion originated through a chromosomal breakage mechanism, either straight-breaks, or nearly straight-breaks, i.e., staggered-breaks whose resulting

duplications are too short to leave long-lasting traces [61]. Be that as it may, no gene was found to have been directly disrupted by the inversion, suggesting that A_{h4} may have been favored indirectly because of its recombination suppression effects.

Apart from the example of A_{h4} , it is worth pinpointing the cases of A_6 and the pair U_1 and U_2 . The first inversion seems a reversal of the telomeric end of chromosome A. Alternatively, it could be subtelomeric [60], and that the tip of the chromosome not affected by the inversion was not included in the assembly. In any case, the rearrangement produced the peritelomeric peak of transposable element repetitive content shown in Fig. 3. With respect to the pair U_1 and U_2 , available cytological evidence could not distinguish between the distal breakpoint of U_1 and the proximal breakpoint of U_2 , pointing to an instance of breakpoint reuse [9]. However, from the assembly the two breakpoints are clearly distinct, although they are only 31 kb distant from each other (Additional file 21: Table S10).

Figure 5 shows reconstructed most parsimonious evolutionary trajectories for all the 12 targeted inversions. Inclusion of *D. madeirensis* was because it is nearly homosequential with *D. subobscura*, and thought to be karyotypically monomorphic for inversions [15]; and also because, together with *D. guanche*, they are the small oceanic-island endemics of the species subgroup. Of the 12 inversions, all but one would have originated in the continental lineage leading to the presently inversion-rich *D. subobscura* (A_{h1} to A_{h4} , A_5 , A_6 , U_1 , U_2 , J_{ST} , E_{ST} and O_{ms}), whereas only one became fixed in the inversion-poor island lineages (E_{g1} in *D. guanche*). Of note is the case of O_{ms} , previously denoted O_g for it was thought to have originated in *D. guanche*. If it is considered that in *D. subobscura* O_{ms} , rather than O_3 as previously thought, is the immediate ancestor on which presently segregating ST and 4 arose, then it may be pertinent to rename O_{ST} and O_{3+4} to O_{ms+ST} and O_{ms+4} , respectively.

Between lineages, considering only rearrangement replacements, *D. subobscura* has evolved at a rate of 5.6 inversions/Myr (assuming 1.72 Myr to the common ancestor of the species subgroup; Additional file 3: Figure S1), which is over 10 times higher than that for the average island endemic (0.4 inversion/Myr; assuming 0.92 Myr to the split of *D. madeirensis* [27]). The difference is highly significant ($P = 3.3 \times 10^{-5}$, Poisson distribution). The lower rate of rearrangement accumulation in *D. guanche* and *D. madeirensis* compared to that in *D. subobscura* could be a reflection of a lower rate of rearrangement formation in small-sized island species. Another, not mutually exclusive possibility is that the difference could be related to that *D. guanche* and *D. madeirensis* remained localized to the small oceanic islands in which they arose, which have



maintained relatively homogenous conditions [97], whereas, in comparison, *D. subobscura* is vastly distributed across multiple contrasting environments with high dispersion. This latter situation was shown to result in increased rates of structural evolution, when the evolutionary fate of inversions is driven by their effect in keeping sets of positively selected alleles together against maladaptive gene flow [64].

The role of the inversions recombination-suppression effect as one driver of genome structural evolution in the *subobscura* subgroup is further supported by the observed ratios of chromosomal divergence to polymorphism between the A sex chromosome and the autosomes in *D. subobscura*. In this lineage, compared to the average autosome the A chromosome shows 8 times larger inversion fixation rate (0.44 vs. 3.5 inversions/Myr, respectively; $P = 0.006$, Poisson distribution; so-called “faster-X” pattern [98]), while 1.8 fewer presently segregating inversions (14 vs. 8, respectively [9]). These conclusions remain qualitatively the same after accounting for chromosome length. The observation of contrasting ratios of polymorphism to divergence between the A sex chromosome and the autosomes agrees with expectations from positive selection models of inversion evolution as byproduct of their

recombination-suppression effects in the face of gene flow, which explain this pattern as resulting from: (i) the higher efficiency of negative selection against locally recessive maladaptive alleles at A-linked genes, whereby A-linked inversions would be expected to capture higher-fitness genotypes with greater probability of fixation; and (ii) the higher likelihood that recessive deleterious alleles generate associative overdominance on autosomes, which would hinder autosomal inversions from fixation [65].

While we may have identified a signature of indirect inversion effects in driving the observed non-random patterns of genome structure evolution in the *subobscura* subgroup, that would not preclude the contribution of other mechanisms. Two would seem be particularly plausible and better suited to be assessed with the data on hand, including mutational biases in the formation of new inversions and direct inversion effects. Overall, however, no positive evidence for any of these two mechanisms could be obtained in the present study. With respect to the first, the observed accelerated rate of structural evolution of the A chromosome compared to the autosomes in *D. subobscura* could result from a bias in the formation of inversions arising from the comparatively higher repetitive content of the A

chromosome (for example, if inversions tended to originate by ectopic recombination [99]). However, reciprocal BLASTN searches using the inversions breakpoints did not detect evidence for an enhanced repeat-based formation of A-linked relative to autosomal inversions. With respect to direct inversion effects, we provide a discussion of our findings in the context of related results below.

Figure 5 shows that, in all five inversion-rich chromosomes of *D. subobscura*, presently segregating *standard* structural variants arose in the mainland after the split of *D. guanche*. In addition to these finding, all the 12 inversions were inferred to have originated by chromosomal breakage. In 4 of the cases, the presence of duplicated sequences in opposite orientation on the flanks of the derived rearrangement provided clear-cut evidence of an origin by staggered breaks, including U_1 (689 bp-long duplication), U_2 (1007 bp), E_{ST} (513 bp) and O_{ms} (538 bp) (Additional file 21: Table S10). The remaining 8 cases could have originated through straight- or nearly straight-breaks. In no case evidence for gene disruptions at the breakpoints could be found, which does not support direct positive selection on the inversions as a major driver of genome structure evolution in the *subobscura* subgroup (see above). Overall, our results suggest that chromosomal breakage is the dominant originating mechanism for inversions in the subgenus *Sophophora*. This contrasts with the situation in the subgenus *Drosophila*, in which inversions appear to originate mainly via ectopic recombination. Although its causes remain to be understood, this difference supports that inversions can arise by alternative major mechanisms in different lineages [61].

Conclusions

We presented the first high-quality, long read-based nuclear and complete mitochondrial genome for *D. subobscura*, and applied it to a synteny analysis of the evolution of genome structure in the *subobscura* species subgroup. We found the sequenced genome to exhibit a relatively compact size, compared to known values from the *obscura* group. *SGM-sat* and *sat290* represent the first and second most abundant satDNAs classes, conversely to the situation in the close relative *D. guanche*. *D. subobscura* exhibits the highest rate of accumulation of paracentric inversions of its subgroup. All identified inversions originated by chromosomal breakage, which adds to the evidence favoring this as the prevailing mechanism of inversion formation in the *Sophophora* subgenus of *Drosophila*. No evidence for direct gene disruption at the inversions breakpoints was found. This observation, together with the finding of contrasting ratios of inversion fixation to polymorphism between the A sex chromosome and the autosomes, overall suggests

that the evolution of genome structure in the lineage leading to *D. subobscura* has been driven indirectly, through the inversions recombination-suppression effects in keeping sets of adaptive alleles together in the face of the high dispersion ability of the species. We have built a genome browser and a BLAST server (<http://dsubobscura.serveftp.com/>) to facilitate the further use of this resource.

Methods

D. subobscura karyotype and chromosome arrangement designation

D. subobscura has six pairs of chromosomes: five acrocentric and one dot. The five acrocentric chromosomes are symbolized by the alphabet vowels capitalized: A (the sex chromosome; Muller's element A, homologous to X in *D. melanogaster*); I, commonly replaced by J (D, 3L); U (B, 2L); E (C, 2R); and O (E, 3R) [9, 11]. The species karyotype is divided into 100 numbered sections as follows: A (1–16), J (17–35), U (36–53), E (54–74), O (75–99) and *dot* (100). Each section is subdivided into 3–5 lettered subsections (from A to E [12]).

Gene arrangements are denoted by subscripts next to chromosome symbols (ST: *standard*; otherwise: alternative arrangements to ST). Overlapping inversions are denoted by underlines below number subscripts [100]. The O chromosome has been particularly amenable for study of structural variation, for it is the only chromosome for which a balanced lethal strain (namely, the *Varicose/Bare*, or abbreviated *Va/Ba* balancer stock [58]) is available. By convention, the O chromosome is divided into two segments, designated I (sections 91 to 99) and II (sections 75 to 90), which are located distal and proximal to the centromere, respectively. The structural variant of the O chromosome used in this study is designated O_{3+4} , a rearrangement of segment I thought to have originated by superimposition of inversion 4 on the ancestral, and now extinct in *D. subobscura* gene order O_3 . It may be worth noting here, that the findings herein show that the immediate ancestral state to inversion 4 is not O_3 , but arrangement O_{ms} , previously called O_g because it was thought to have originated in *D. guanche* (see the Results and Discussion section, and Fig. 5 legend).

D. subobscura lines

We used one inbred line for de novo complete genome assembly using PacBio long-read data. The inbred line was obtained by 10 generations of full-sib mating of progeny of a single gravid female from our highly homozygous laboratory stock of the *ch-cu* marker strain. The *ch-cu* strain was established by Loukas et al. [16] from flies descended from the "*β-ch-cu-stock*" [101, 102]. Structurally, it is homokaryotypic for the ST arrangements in all

chromosomes except in chromosome O, for which it is homokaryotypic for the O_{3+4} configuration. Crossing schemes and the methods for polytene chromosome staining and identification are described elsewhere [36]. The assayed inbred line was stored frozen at -80°C immediately upon obtention.

Genome size estimation by flow cytometry

Genome size of adult *D. subobscura ch-cu* was quantified from five replicates of brain cell nuclei using propidium-iodide (PI) based flow cytometry [89]. By this method, the size of a target genome is estimated by comparing stain uptake of the target genome (PI-fluor_{target}), with that of a standard genome of known size (PI-fluor_{standard}). A *D. virilis* strain with known 328 Mb genome size [68] was used as the standard.

Nuclei were extracted from samples of $10-80^{\circ}\text{C}$ —frozen heads from four-days-old ice-immobilized females, each including 5 heads from *ch-cu* and 5 heads from the standard. Each sample was transferred to a glass/glass homogenizer (Kontes Dounce Tissue Grinder 7 ml), ground on ice-cold LB Galbraith buffer using the large clearance pestle (pestle A), and the homogenates filtered through nylon mesh (20 μm). The filtrates were stained for 2 h in $50\ \mu\text{g}\ \text{ml}^{-1}$ PI, and subsequently analyzed on a BD Biosciences (BDB) Dual Laser FACSalibur (Becton Dickinson, Franklin Lakes, NJ, USA) flow cytometer, using the forward (FS) and side (SS) scattering, together with the red PI fluorescence ($> 670\ \text{nm}$) detected by the FL3 detector. Data were generated at low flow rate (~ 1000 nuclei/min). Data analysis was performed using the BD FACSDiva 4.0 software (BD Biosciences, San José, CA, USA). Individual nuclei were gated from aggregates and debris by their area (FL3-A) vs. width (FL3-W) fluorescence signal. Measures were obtained from a minimum of 10,000 nuclei per sample. Genome sizes were estimated using the formula: $\text{GS}_{D. subobscura (ch-cu)} = \text{GS}_{D. virilis} \times (\text{PI-fluor}_{D. subobscura (ch-cu)} / (\text{PI-Fluor}_{D. virilis}))$.

High molecular weight genomic DNA isolation and PacBio whole-genome sequencing

High-quality high molecular weight gDNA was obtained from 60 mg mixes of -80°C frozen adult males and females, using a modified version of the phenol/chloroform method of Chen et al., [103] that yields $\sim 25\ \mu\text{g}$ of high quality DNA per assay, as assessed by NanoDrop ND1000 (NanoDrop Technologies Inc., Wilmington, DE, USA) spectrophotometer and standard agarose gel electrophoresis. The genome of the inbred *ch-cu* line was sequenced to nominal 40-fold genome coverage using PacBio (Pacific Biosciences, Menlo Park, CA, USA) RSII single-molecule real-time (SMRT) technology from a 20-kb SMRTbell template library, using P6-C4 chemistry and seven SMRT cells. Libraries construction

and PacBio sequencing were outsourced to MacroGen (MacroGen Inc., Seoul, South Korea).

De novo genome assembly

Raw PacBio reads were assembled using the Canu assembler (Ver. 1.5 [104]) on recommended settings for read error correction, trimming and assembly, and genome size set at 150 Mb (see below). In addition, we also tried HINGE [105], FALCON [106] and MECAT [107]. Compared to Canu, these alternative bioinformatics pipelines produced smaller and less contiguous assemblies on our data. These analyses were performed on a 2.80-GHz 8-CPU Intel Xeon 64-bit 32 GB-RAM computer running Ubuntu 16.04 LTS.

Genome scaffolding

Chromosomal assignment, ordering and orientation of Canu contigs was accomplished in four steps. In step I, the contigs were checked for the presence of inter- and intra-chromosomal chimeras using a semi-automatic recursive approach combining: i) cross-species synteny information inferred using BLAT [108] and BLASTN [109], setting the genome of *D. melanogaster* (release r6.22) and the more closely related, yet not so-well characterized genome of *D. pseudoobscura* (r3.04) as the reference. Here, BLASTN was used in relatively few cases where BLAT either did not return a hit, returned multiple equal score hits, or returned a hit to scaffold unknown from *D. pseudoobscura*. The first of these three cases involved short and fast-evolving contigs and bacterial contigs; the second one involved contigs containing Repbase (Ver. 20,150,897 [110]) identified repetitive sequences, which were re-examined after masking of the repeats; in the third case, BLASTN was used to confirm that the target contig mapped exclusively to one scaffold. Cross-species synteny information obtained in this way was combined with ii) the wealth of available *D. subobscura*'s physical mapping [18, 84, 111] and genetic linkage [13, 112, 113] data. Markers' sequences were retrieved from FlyBase 2.0 (release FB2017_02) using gene names and/or annotation symbols provided by the authors. In step II, Canu contigs that passed step I were scaffolded using SSPACE-LongRead (Ver. 1–1 [73]). In step III, the resulting SSPACE scaffolds were submitted to a second round of quality check as in step I. In step IV, the assembled sequence that passed step III was assigned genomic coordinates based on the physical location of the markers.

Genome annotation

Prediction and annotation of the genome assembly was conducted using MAKER (Ver. 3.01.02. -beta [114, 115]) annotation pipeline with default parameters. Repetitive elements were identified using RepeatMasker (Ver. 4.0.6

[116]) combined with the *Drosophila* genus specific repeat library included in Repbase database. Two previously described satellites, namely *sat290* [90] and *SGM-sat* [91] were absent from the Repbase database, thereby they were ascertained separately by BLAST search using already available sequences from the *D. guanche sat290* [90] and the *D. subobscura SGM-sat* (GenBank accession AF043638.1 [91]) as queries. SNAP [117], AUGUSTUS [118], GeneMark-ES [119], and geneid [120] were selected for ab initio gene model prediction on the repeat masked genome sequence. Proteomes from 12 *Drosophila* species from Flybase database (FB2017_05, released October 25, 2017 [121]), and additional 491 *D. subobscura* protein sequences from UniProt database (release 2017_12 [122]) were used in the analysis.

The quality of the annotation was controlled using the Annotation Edit Distance (AED) metric [123]. AED values are bounded between 0 and 1; an AED value of 0 indicates perfect agreement of the annotation to aligned evidence. Conversely, a value of 1 indicated no evidence support.

Functional annotation of MAKER-predicted proteins was made by BLASTP (Ver. 2.6.0+) searches against the *Drosophila* UniProt-SwissProt manually curated datasets [124]. Prediction of protein functional domains was accomplished using InterProScan (Ver. 5.29–68.0 [125]) on the Pfam [126], InterPro [127], and Gene Ontology (GO) [128, 129] domain databases. UniProt-SwissProt BLAST and InterProScan functional assignments were extracted using the ANNOTATION INFORMATION EXTRACTOR (ANNIE [130]), which assigns gene names and products by database cross-referencing. InterProScan functional assignments were mapped to Gene Ontology (GO) terms using Blast2GO (Ver. 5.0.13 [131]). The combined graph function of Blast2GO was used to generate gene ontology graphs and pie charts from the GO terms.

Genome assembly and annotation completeness was gauged using the Benchmarking Universal Single-Copy Orthologs (BUSCO) tool (BUSCO, Ver. 3 [132]) analysis against the diptera_odb9 dataset, which contains 2799 highly-conserved, single-copy genes likely to be present in any dipteran genome. The dipteran set was selected, because being the most narrowly defined *Drosophila*-including set, it is also the largest, therefore the one expected to provide the best resolution.

Mitogenome assembly and annotation

Annotation of the *D. subobscura* mitogenome was conducted using the MITOS online tool (<http://mitos.bioinf.uni-lipzig.de/index.py> [133]), with default settings, meta-zoan reference, and invertebrate genetic code, and was further adjusted manually according to its alignment with available mitogenomes from other *Drosophila* species.

Orthologous group assignment and gene family expansion/contraction analyses

The complete set of *D. subobscura* annotated proteins were clustered into orthologous groups by comparison with the 12 *Drosophila* genomes (FlyBase releases *dana_R1.06*, *dere_R1.05*, *dgri_R1.05*, *dmel_R6.22*, *dmoj_R1.04*, *dper_R1.3*, *dpse_R3.04*, *dsec_R1.3*, *dsim_R2.02*, *dvir_R1.07*, *dwil_R1.05*, and *dyak_R1.05*), plus that of its close relative *D. guanche* (*dgua_R1.0* [59]). Orthologous group assignment was conducted using OrthoMCL (Ver. 5 [134]) on default settings. OrthoMCL generates orthologous groups via all-to-all BLASTP comparison followed by Markov clustering of the reciprocal best similarity pairs.

Analysis of gene family expansion and contraction was conducted using the Computational Analysis of gene Family Evolution (CAFE Ver. 3.1 [92]) tool. For a specified ultrametric phylogenetic tree, and given the gene family sizes in the extant species, CAFE uses a maximum likelihood stochastic birth-and-death process to model the rate and direction of change in gene family size (in number of gene births and deaths per gene per million years; symbolized λ) over the tree. CAFE was run on default parameters using a 14 species ultrametric tree built by grafting *D. subobscura* and *D. guanche* onto the 12 *Drosophila* tree used by Hahn et al. [135] at positions obtained from the *TimeTree* database (<http://www.timetree.org/> [136]):

(((((Dsim:2.1,Dsec:2.1):3.2,Dmel:5.3):5.9,(Dere:8.5,-Dyak:8.5):2.7):42.1,Dana:53.3):2.3,((Dpse:1.4,Dper:1.4):13.1,(Dsub:3.1,Dgua:3.1):11.4):41.1):6.8,Dwil:62.4):0.8,((Dvir:32.7,Dmoj:32.7):4.3,Dgri:37):26.2); with branch lengths given in million years.

Model-fitting considered three nested likelihood models of gene family size evolution. The first model assumes a single global λ_G for all lineages. The second model allows for three λ to accommodate for fast- ($\lambda_F \geq 0.010$), medium- ($0.010 > \lambda_M > 0.002$), and slow-evolving ($\lambda_S \leq 0.002$) branches. Assignment of each branch to its corresponding λ category (i.e., λ_F , λ_M or λ_S) in this model was made a priori, based on the best results for a fully 26 λ -parameters model (i.e., one for each branch of the phylogeny), as in Hahn et al. [135]. The third model is a five λ generalization of the second model to allow for the terminal branches leading to *D. subobscura* and *D. guanche* having their own rates (i.e., λ_{Ds} and λ_{Dg} , respectively). Estimates of λ obtained using this approach are sensitive to suboptimal genome assembly and/or annotation. Therefore, the obtained best-fitting model was refined by adding to it a term of error (ϵ) in genome quality. The effect of the error term on λ provides an indirect measure of genome assembly and/or annotation completeness [92]. For each model, at least five CAFE runs were performed and those runs with the highest likelihood score per model were included. To meet the

CAFE assumption that gene families must have been present at the root of the tree, only families found in at least one species of both the *Sophophora* and *Drosophila* subgenera, were considered. Both OrthoMCL and CAFE analyses were conducted considering only the longest splice forms.

Functional enrichment analyses of gene families uncovered to have been rapidly evolving along the terminal branch of this species by CAFE were carried out using the Blast2GO [131] implementation of the one-sided Fisher's exact test, with false discovery rate (FDR) < 0.001. Enriched GO terms were summarized and visualized using the online version of REVIGO (<http://revigo.irb.hr/> [137]). This tool identifies representative GO terms by semantic similarity.

Whole-genome synteny analysis

The genome of *D. subobscura* was analyzed for conservation of synteny against those of three increasingly distant species, namely *D. guanche* (*dgua_R1.01* [59]), *D. pseudoobscura* (*dpse_R3.04*), and *D. melanogaster* (*dmel_R6.22*), using the Synteny Mapping and Analysis Program (SyMAP, Ver. 4.2. [138, 139]) tool on default options. SyMAP is a long-range whole-genome synteny mapping tool devised to accommodate for intervening micro-rearrangements which could result from misassembling, but also from real structural changes. Therefore, SyMAP seemed especially suited for investigating large, cytologically visible recent chromosomal rearrangement events that are the focus of the present study.

Additional files

Additional file 1: Table S1. Genome sequencing and assembly statistics (coverages based on a 150Mb genome size; lengths are in bp). (DOCX 43 kb)

Additional file 2: Table S2. Genetic markers used for validation, and physical anchoring, ordering and orientation of scaffolds. In total, 683 markers were considered, of which 621 were used. 62 markers were not used because they showed inconsistencies as to their localization with respect to markers from other studies and our own data. The MS Excel file contains eight spreadsheets, including one for this title, one for each of the five major pseudo-chromosomes (i.e., A, J, U, E and O), one with a summary, and one with the references for the marker data. For each pseudo-chromosome, markers are listed in column "A", including used cytological (numbered, black), used linkage (numbered, blue), and nonused (nonnumbered, red) markers. For each marker, information relative to its name, cytological localization, authors, corresponding *D. pseudoobscura* "GA" gene model name, inconsistency where it applies, coordinates in the pseudo-chromosome, scaffold name, scaffold orientation and cytological span, and BLASTn statistics is provided in subsequent columns, from "B" to "Y". Column "X" provides the number of used marker per scaffold. Alternating color in the background denotes different scaffolds. Cytological coordinates are always relative to the Kunze-Mühl and Müller [12] standard reference map. From the summary spreadsheet, most of the inconsistencies (72%) come from one (Laayouni et al. 2007) out of the total 26 cited works. Excluding that study, the total percent of inconsistencies is only 2.85% (i.e., 17 out of 638 markers). (XLSX 168 kb)

Additional file 3: Figure S1. RelTime timetree of 14 *Drosophila* species obtained using the maximum-likelihood tree-topology that results after GTR + G + I best-fit modeling of a 50 concatenated nuclear low-codon bias orthologous gene alignment dataset. Blue diamonds indicate Obbard et al. [78] mutation-based calibrated nodes, and orange boxes 95% confidence intervals for target divergences. (PDF 10 kb)

Additional file 4: Table S3. *D. subobscura* mitogenome gene content and order (lengths in bp). (DOCX 45 kb)

Additional file 5: Table S4. Repetitive content of the *D. subobscura* genome. (DOCX 43 kb)

Additional file 6: Figure S2. OrthoMCL analysis of gene families in *D. subobscura*. Numbers of orthoMCL clusters and of genes within those clusters on each node are given in black and white rectangles, respectively. (PDF 15 kb)

Additional file 7: Table S5. Optimal CAFE model selection for the evolution of gene family size along the 14 *Drosophila* ultrametric tree in Figures S3-S4. Shown are the four assayed increasingly complex models, including the 1- λ and 3- λ models, and the 5- λ model without and with global assembling error term (ϵ); and their corresponding parameter estimates, including global (λ_G), slow (λ_S), medium (λ_M), fast (λ_F), *D. subobscura* (λ_{DS}) and *D. guanche* (λ_{DG}) lambdas, and global error, and maximum-likelihood scores (-lnL). (DOCX 42 kb)

Additional file 8: Figure S3. CAFE analysis of the evolution of gene family size in *D. subobscura*. Shown on each branch are its corresponding numbers of expanded (left) and contracted (right) gene families. Circled numbers on nodes are identifiers for internal branches of the phylogeny leading to those nodes. The colors of the circles indicate estimated rates of gene gain and loss according to the legend on the upper left (blue: slow, grey: medium, red: fast). (PDF 33 kb)

Additional file 9: Figure S4. CAFE analysis of the evolution of gene family size in *D. subobscura*. Shown on each branch are its corresponding numbers of significantly expanded (green) and contracted (orange) gene families. Circled numbers on nodes are identifiers for internal branches of the phylogeny leading to those nodes. The colors of the circles indicate estimated rates of gene gain and loss according to the legend on the upper left (blue: slow, grey: medium, red: fast). (PDF 33 kb)

Additional file 10: Table S6. Over represented GO Terms among CAFE significantly expanded gene families in *D. subobscura* inferred using one-sided Fisher exact test (FDR < 0.001) implemented in Blast2Go (BP: Biological Process; MF: Molecular Function; CC: Cellular Component). (DOCX 46 kb)

Additional file 11: Table S7. Over represented GO Terms among CAFE significantly contracted gene families in *D. subobscura* inferred using one-sided Fisher exact test (FDR < 0.001) implemented in Blast2Go (BP: Biological Process; MF: Molecular Function; CC: Cellular Component). (DOCX 47 kb)

Additional file 12: Figure S5. REVIGO summary scatterplot for 27 over-represented Biological Process GO terms in CAFE-expanded gene families. Shown GO term names denote cluster representatives centered on their corresponding GO term. Distances between GO terms are in units of semantic similarity. Circle color indicates FDR values, and circle size generality of the GO term (the lower, the greater the uniqueness of the term). (PDF 88 kb)

Additional file 13: Figure S6. REVIGO summary scatterplot for 17 over-represented Molecular Function GO terms in CAFE-expanded gene families. Shown GO term names denote cluster representatives centered on their corresponding GO term. Distances between GO terms are in units of semantic similarity. Circle color indicates FDR values, and circle size generality of the GO term (the lower, the greater the uniqueness of the term). (PDF 90 kb)

Additional file 14: Figure S7. REVIGO summary scatterplot for 9 over-represented Cellular Component GO terms in CAFE-expanded gene families. Shown GO term names denote cluster representatives centered on their corresponding GO term. Distances between GO terms are in units of semantic similarity. Circle color indicates FDR values, and circle size generality of the GO term (the lower, the greater the uniqueness of the term). (PDF 49 kb)

Additional file 15: Figure S8. REVIGO summary scatterplot for 51 over-represented Biological Process GO terms in CAFE-contracted gene families. Shown GO term names denote cluster representatives centered on their corresponding GO term. Distances between GO terms are in units of semantic similarity. Circle color indicates FDR values, and circle size generality of the GO term (the lower, the greater the uniqueness of the term). (PDF 126 kb)

Additional file 16: Figure S9. REVIGO summary scatterplot for 12 over-represented Molecular Function GO terms in CAFE-contracted gene families. Shown GO term names denote cluster representatives centered on their corresponding GO term. Distances between GO terms are in units of semantic similarity. Circle color indicates FDR values, and circle size generality of the GO term (the lower, the greater the uniqueness of the term). (PDF 90 kb)

Additional file 17: Figure S10. REVIGO summary scatterplot for 8 over-represented Cellular Component GO terms in CAFE-contracted gene families. Shown GO term names denote cluster representatives centered on their corresponding GO term. Distances between GO terms are in units of semantic similarity. Circle color indicates FDR values, and circle size generality of the GO term (the lower, the greater the uniqueness of the term). (PDF 63 kb)

Additional file 18: Table S8. Number of syntenic blocks between *D. subobscura* and increasingly distant relatives. (DOCX 41 kb)

Additional file 19: Table S9. Average size of the syntenic block (in Mb) between *D. subobscura* and increasingly distant relatives. (DOCX 41 kb)

Additional file 20: Figure S11. Schematic of the strategy used for inversion breakpoint detection. From top to bottom: shown are (a) two noninverted (SB1 and SB3; pink) and one inverted (SB2; green) hypothetical SyMAP synteny blocks between two taxa (1 and 2). The regions flanking the points of broken synteny (vertical dotted lines) are labelled A-D correspondingly; (b) BLASTing regions AB and CD from taxon 1 against the genome of taxon 2 each produces two hits (c) at opposite ends of the inverted synteny block with associated overhangs; (d) steps b-c are repeated using taxon 2 for the BLAST queries to test for reciprocal consistency (see main text for more detail). (PDF 97 kb)

Additional file 21: Table S10. Synteny analysis of inversion breakpoints. Provided is breakpoint information for 12 inversions, including six from pseudo-chromosome A (h1, h2, h3, h4, 5 and 6), one from J (ST), two from U (1 and 2), two from E (g1 and ST), and one from O (ms). The MS Excel file contains six spreadsheets, including one for this title, and one for each of the five major pseudo-chromosomes (i.e., A, J, U, E and O). For each pseudo-chromosome, inversions are listed in column "A". For each inversion, information about the three protein coding genes flanking each side of each breakpoint in three species, including *D. melanogaster*, *D. guanche* and *D. subobscura* is provided in subsequent columns, from "B" to "Q". This information includes species names, names and pseudo-chromosome coordinates of the three coding gene markers on both sides of each distal and proximal breakpoint, and the size of the pseudo-chromosome segment spanned by the breakpoints in Mb. Also provided is, for each breakpoint, its cytological and estimated pseudo-chromosome coordinates, and its hypothetical originating mechanism with the length of the associated duplication where it applies. Cells color background indicate contiguity (brown) or altered (yellow) order of the markers relative to the outgroup (*D. melanogaster*/*D. pseudoobscura*). For example, in the case of hypothetical inversion 1 of the A chromosome (i.e., h1) in *D. subobscura*, the three markers downstream the proximal breakpoint and upstream the distal breakpoint are in reverse order relative to *D. guanche*, which shows the markers ordered as in *D. melanogaster*. Reciprocal BLASTn searches with each breakpoint did not detect evidence of duplication, suggesting that the most likely originating mechanism of inversion A_{h1} (depicted in yellow) is simple, or nearly straight breaks. (XLSX 31 kb)

Acknowledgements

We thank Manuela Costa from the Cytometry Unit at Universitat Autònoma de Barcelona for her technical support and advice in flow cytometry measures.

Funding

This study was supported by the Spanish Ministerio de Ciencia e Innovación grant CGL2017-89160P; and Generalitat de Catalunya grant 2017SGR 1379 to the Grup de Genòmica, Bioinformàtica i Biologia Evolutiva, Universitat Autònoma de Barcelona (Spain). CK was supported by a PIF PhD fellowship from the Universitat Autònoma de Barcelona (Spain). Note that the funding agencies were not involved in the design of the study or in any aspect of the collection, analysis and interpretation of the data or paper writing.

Availability of data and materials

All the *D. subobscura* sequencing data obtained in this work are available in the European Nucleotide Archive (ENA) under the project ID PRJEB31081. Furthermore, all the scaffolds and the pseudomolecules, along with their annotations, are available in our local server in the form of a genome browser (<http://dsubobscura.servvefp.com/>).

Authors' contributions

CK, RT, FR-T conceived and designed the research. CK performed most of the analyses with support from RT and FR-T. VG-V carried out the flow cytometry measures with supervision from RT and FR-T. CK, RT, and FR-T wrote the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 18 December 2018 Accepted: 6 March 2019

Published online: 18 March 2019

References

- Collin JE. Note: *Drosophila subobscura* sp. n. male, female. *J Genet.* 1936;33:60.
- Buzzati-Traverso AA, Scossiroli RE. The "obscura group" of the genus *Drosophila*. *Adv Genet.* 1955;7:47–92.
- Maynard-Smith J. Fertility, mating behaviour and sexual selection in *Drosophila subobscura*. *J Genet.* 1956;54:261–79.
- Holman L, Freckleton RP, Snook RR. What use is an infertile sperm? A comparative study of sperm-heteromorphic *Drosophila*. *Evolution.* 2007;62:374–85.
- Fisher DN, Doff RJ, Price TA. True polyandry and pseudopolyandry: why does a monandrous fly remate? *BMC Evol Biol.* 2013;13:157.
- Philip D, Rendel JM, Spurway H, Haldane JBS. Genetics and karyology of *Drosophila subobscura*. *Nature.* 1944;154:260–2.
- Rendel JM. Genetics and cytology of *D. subobscura*. II. Normal and selective matings in *Drosophila subobscura*. *J Genet.* 1945;46:287–302.
- Ewing AW, Bennet-Clark HC. The courtship songs of *Drosophila*. *Behaviour.* 1968;31:288–301.
- Krimbas CB. The inversion polymorphism of *Drosophila subobscura*. In: Krimbas CB, Powell JR, editors. *Drosophila inversion polymorphism*: Boca Raton: CRC Press; 1992. p. 127–220.
- Emmens CW. The morphology of the nucleus in the salivary glands of four species of *Drosophila* (*D. melanogaster*, *D. immigrans*, *D. funebris* and *D. subobscura*). *Z Zellforsch u mikroskop Anal.* 1937;26:1–20.
- Mainx F, Koske T, Smital E. Untersuchungen über dier chromosomale Struktur europaischer Vertreter der *Drosophila obscura* Gruppe. *Z Indukt Abstamm Vererbungsl.* 1953;85:354–72.
- Kunze-Mühl E, Müller E. Weitere Untersuchungen über die chromosomale Struktur und die natürlichen Strukturtypen von *Drosophila subobscura* Coll. *Chromosoma.* 1958;9:559–70.
- Loukas M, Krimbas CB, Mavragani-Tsipidou P, Kastritsis CD. Genetics of *Drosophila subobscura* populations VIII. Allozyme loci and their chromosome maps. *J Hered.* 1979;70:17–26.

14. Moltó MD, De Frutos R, Martínez-Sebastián MJ. The banding pattern of polytene chromosomes of *Drosophila guanche* compared with that of *D. subobscura*. *Genetica*. 1987;75:55–70.
15. Papaceit M, Prevosti A. A photographic map of *Drosophila madeirensis* polytene chromosomes. *J Hered*. 1991;82:471–8.
16. Loukas M, Krimbas CB, Vergini Y. The genetics of *Drosophila subobscura* populations. IX. Studies on linkage disequilibrium in four natural populations. *Genetics*. 1979;93:497–523.
17. Santos J, Serra L, Solé E, Pascual M. FISH mapping of microsatellite loci from *Drosophila subobscura* and its comparison to related species. *Chromosom Res*. 2010;18:213–26.
18. Orengo DJ, Puerma E, Papaceit M, Segarra C, Aguadé M. Dense gene physical maps of the non-model species *Drosophila subobscura*. *Chromosom Res*. 2017;25:145–54.
19. Latorre A, Moya A, Ayala FJ. Evolution of mitochondrial DNA in *Drosophila subobscura*. *Proc Natl Acad Sci U S A*. 1986;83:8649–53.
20. Christie JS, Picornell A, Moya A, Ramon MM, Castro JA. Mitochondrial DNA effects on fitness in *Drosophila subobscura*. *Heredity*. 2011;107:239–45.
21. Kurbalija Novičić Z, Immonen E, Jelić M, Anđelković M, Stamenković-Radak M, Arnqvist G. Within-population genetic effects of mtDNA on metabolic rate in *Drosophila subobscura*. *J Evol Biol*. 2015;28:338–46.
22. Monclús M. Distribución y ecología de drosophilidos en España. II. Especies de *Drosophila* de las Islas Canarias, con la descripción de una nueva especie. *Bol R Soc Esp His Nat Secc Biol*. 1976;74:197–213.
23. Monclús M. *Drosophilidae* of Madeira, with the description of *Drosophila madeirensis*-n. sp. *Z Zool Syst Evolutionsforsch*. 1984;22:94–103.
24. Taxodros: a taxonomic database of the *Drosophilidae*. Bächli G. University of Zürich. <https://www.taxodros.uzh.ch/>. Accessed 17 October 2018.
25. Krimbas CB, Loukas M. Evolution of the *obscura* group *Drosophila* species. I. Salivary chromosomes and quantitative characters in *D. subobscura* and two closely related species. *Heredity*. 1984;53:469–82.
26. Rego C, Santos M, Matos M. Quantitative genetics of speciation: additive and non-additive genetic differentiation between *Drosophila madeirensis* and *Drosophila subobscura*. *Genetica*. 2007;131:167–74.
27. Herrig DK, Modrick AJ, Brud E, Llopart A. Introgression in the *Drosophila subobscura*-*D. madeirensis* sister species: evidence of gene flow in nuclear genes despite mitochondrial differentiation. *Evolution*. 2014;68:705–19.
28. Papaceit M, San Antonio J, Prevosti A. Genetic analysis of extra sex combs in the hybrids between *Drosophila subobscura* and *D. madeirensis*. *Genetica*. 1991;84:107–14.
29. Khadem M, Krimbas CB. Studies of the species barrier between *Drosophila subobscura* and *D. madeirensis*. III. How universal are the rules of speciation? *Heredity*. 1993;70:353–61.
30. Mittleman BE, Manzano-Winkler B, Hall JB, Korunes KL, Noor MAF. The large X-effect on secondary sexual characters and the genetics of variation in sex comb tooth number in *Drosophila subobscura*. *Ecol Evol*. 2017;7:533–40.
31. Powell JR. Progress and prospects in evolutionary biology: the *Drosophila* model. New York: Oxford University Press; 1997.
32. Zivanovic G, Arenas C, Mestres F. Individual inversions or their combinations: which is the main selective target in a natural population of *Drosophila subobscura*? *J Evol Biol*. 2016;29:657–64.
33. Orengo DJ, Puerma E, Papaceit M, Segarra C, Aguadé M. A molecular perspective on a complex polymorphic inversion system with cytological evidence of multiply reused breakpoints. *Heredity*. 2015;114:610–08.
34. Puerma E, Orengo DJ, Aguadé M. Multiple and diverse structural changes affect the breakpoint regions of polymorphic inversions across the *Drosophila* genus. *Sci Rep*. 2016;6:36248.
35. Menozzi P, Krimbas CB. The inversion polymorphism of *D. subobscura* revisited: synthetic maps of gene arrangement frequencies and their interpretation. *J Evol Biol*. 1992;5:625–41.
36. Rodríguez-Trelles F, Alvarez G, Zapata C. Time-series analysis of seasonal changes of the O inversion polymorphism of *Drosophila subobscura*. *Genetics*. 1996;142:179–87.
37. Brncic D, Budnik M. Colonization of *D. subobscura* Collin in Chile. *Dros Inform Serv*. 1980;55:20.
38. Beckenbach AT, Prevosti A. Colonization of North America by the European species, *D. subobscura* and *D. ambigua*. *Am Midl Nat*. 1986;115:10–8.
39. Pascual M, Aquadro CF, Soto V, Serra L. Microsatellite variation in colonizing and paleartic populations of *Drosophila subobscura*. *Mol Biol Evol*. 2001;18:731–40.
40. Prevosti A, Ribo G, Serra L, Aguade M, Balaña J, Monclus M, Mestres F. Colonization of America by *Drosophila subobscura*: experiment in natural populations that supports the adaptive role of chromosomal-inversion polymorphism. *Proc Natl Acad Sci U S A*. 1988;85:5597–600.
41. Huey RB, Gilchrist GW, Carlson ML, Berrigan D, Serra L. Rapid evolution of a geographic cline in size in an introduced fly. *Science*. 2000;287:308–9.
42. Ayala FJ, Serra L, Prevosti A. A grand experiment in evolution: the *Drosophila subobscura* colonization of the Americas. *Genome*. 1989;31:246–55.
43. Jungen H. Abnormal sex ratio, linked with inverted gene sequence, in populations of *D. subobscura* from Tunisia. *Dros Inform Serv*. 1967;42:109.
44. Verspoor RL, Smith JML, Mannion NML, Hurst GDD, Price TAR. Strong hybrid male incompatibilities impede the spread of a selfish chromosome between populations of a fly. *Evol Lett*. 2018;2:169–79.
45. Rodríguez-Trelles F, Rodríguez MA. Rapid micro-evolution and loss of chromosomal diversity in *Drosophila* in response to climate warming. *Evol Ecol*. 1998;12:829–38.
46. Rodríguez-Trelles F, Rodríguez MA. Comment on 'Global genetic change tracks global climate warming in *Drosophila subobscura*'. *Science*. 2007;315:1497.
47. Balanyà J, Oller JM, Huey RB, Gilchrist GW, Serra L. Global genetic change tracks global climate warming in *Drosophila subobscura*. *Science*. 2006;313:1773–5.
48. Balanyà J, Huey RB, Gilchrist GW, Serra L. The chromosomal polymorphism of *Drosophila subobscura*: a microevolutionary weapon to monitor global change. *Heredity*. 2009;103:364–7.
49. Rodríguez-Trelles F, Rodríguez MA. Measuring evolutionary responses to global warming: cautionary lessons from *Drosophila*. *Insect Conserv Div*. 2010;3:44–50.
50. Rezende EL, Balanyà J, Rodríguez-Trelles F, Rego C, Fragata I, Matos M, Serra L, Santos M. Climate change and chromosomal inversions in *Drosophila subobscura*. *Clim Res*. 2010;43:103–14.
51. Rodríguez-Trelles F, Tarrío R, Santos M. Genome-wide evolutionary response to a heat wave in *Drosophila*. *Biol Lett*. 2013;9:e20130228.
52. Rozas J, Aguadé M. Gene conversion is involved in the transfer of genetic information between naturally occurring inversions of *Drosophila*. *Proc Natl Acad Sci U S A*. 1994;91:11517–21.
53. Betrán E, Rozas J, Navarro A, Barbadilla A. The estimation of the number and the length distribution of gene conversion tracts from population DNA sequence data. *Genetics*. 1997;146:89–99.
54. Munté A, Rozas J, Aguadé M, Segarra C. Chromosomal inversion polymorphism leads to extensive genetic structure: a multilocus survey in *Drosophila subobscura*. *Genetics*. 2005;169:1573–81.
55. Puig Giribets M, García Guerreiro MP, Santos M, Ayala FJ, Tarrío R, Rodríguez-Trelles F. Chromosomal inversions promote genomic islands of concerted evolution of *Hsp70* genes in the *Drosophila subobscura* species subgroup. *Mol Ecol*. 2019. <https://doi.org/10.1111/mec.14511>.
56. Santos M, Céspedes W, Balanyà J, Trotta V, Calboli FC, Fontdevila A, Serra L. Temperature-related genetic changes in laboratory populations of *Drosophila subobscura*: evidence against simple climatic-based explanations for latitudinal clines. *Am Nat*. 2005;165:258–73.
57. Fragata I, Lopes-Cunha M, Bárbaro M, Kellen B, Lima M, Santos MA, Faria GS, Santos M, Matos M, Simões P. How much can history constrain adaptive evolution? A real-time evolutionary approach of inversion polymorphisms in *Drosophila subobscura*. *J Evol Biol*. 2014;27:2727–38.
58. Sperlich D, Feuerbach-Mravlag H, Lange P, Michaelidis A, Pentzós-Daponte A. Genetic load and viability distribution in central and marginal populations of *Drosophila subobscura*. *Genetics*. 1977;86:835–48.
59. Puerma E, Orengo DJ, Cruz F, Gómez-Garrido J, Librado P, Salguero D, Papaceit M, Gut M, Segarra C, Alioto TS, et al. The high-quality genome sequence of the oceanic island endemic species *Drosophila guanche* reveals signals of adaptive evolution in genes related to flight and genome stability. *Genome Biol Evol*. 2018;10:1956–69.
60. Brehm A, Krimbas CB. Evolution of the *obscura* group *Drosophila* species. III. Phylogenetic relationships in the *subobscura* cluster based on homologies of chromosome A. *Heredity*. 1990;65:269–75.
61. Ranz JM, Maurin D, Chan YS, von Grothuss M, Hillier LW, Roote J, Ashburner M, Bergman CM. Principles of genome evolution in the *Drosophila melanogaster* species group. *PLoS Biol*. 2007;5:e152.
62. Kirkpatrick M. How and why chromosome inversions evolve. *PLoS Biol*. 2010;8:e1000501.

63. Dobzhansky T. Genetics of natural populations. XIV. A response of certain gene arrangements in the third chromosome of *Drosophila pseudoobscura* to natural selection. *Genetics*. 1947;32:142–60.
64. Kirkpatrick M, Barton N. Chromosome inversions, local adaptation and speciation. *Genetics*. 2006;176:419–34.
65. Connallon T, Olito C, Dutoit L, Papoli H, Ruzicka F, Yong L. Local adaptation and the evolution of inversions on sex chromosomes and autosomes. *Phil Trans R Soc B*. 2018;373:20170423.
66. Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, Schatz MC. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics*. 2017;33:2202–4.
67. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*. 2011;27:764–70.
68. Hare EE, Johnston JS. Genome size determination using flow cytometry of propidium iodide-stained nuclei. *Methods Mol Biol*. 2011;772:3–12.
69. Nardon C, Deceliere G, Loevenbruck C, Weiss M, Vieira C, Biémont C. Is genome size influenced by colonization of new environments in dipteran species? *Mol Ecol*. 2005;14:869–78.
70. Animal Genome Size Database. Gregory TR. <http://www.genomesize.com>. Accessed 19 June 2018.
71. Rhoads A, Au KF. PacBio sequencing and its applications. *Genomics Proteomics Bioinformatics*. 2015;13:278–89.
72. Chandler JA, Lang JM, Bhatnagar S, Eisen JA, Kopp A. Bacterial communities of diverse *Drosophila* species: ecological context of a host-microbe model system. *PLoS Genet*. 2011;7:e1002272.
73. Boetzer M, Pirovano W. SPADe-LongRead: scaffolding bacterial draft genomes using long read sequence information. *BMC Bioinf*. 2014;15:211.
74. English AC, Richards S, Han Y, Wang M, Vee V, Qu J, Qin X, Muzny DM, Reid JG, Worley KC, et al. Mind the gap: upgrading genomes with Pacific biosciences RS long-read sequencing technology. *PLoS One*. 2012;7:e47768.
75. Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA, Kaufman TC, Kellis M, Gelbart W, Iyer VN, et al. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature*. 2007;450:203–18.
76. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, et al. The genome sequence of *Drosophila melanogaster*. *Science*. 2000;287:2185–95.
77. Steinemann M, Pinsker W, Sperlich D. Chromosome homologies within the *Drosophila obscura* group probed by in situ hybridization. *Chromosoma*. 1984;91:46–53.
78. Obbard DJ, Maclennan J, Kim KW, Rambaut A, O'Grady PM, Jiggins FM. Estimating divergence dates and substitution rates in the *Drosophila* phylogeny. *Mol Biol Evol*. 2012;29:3459–73.
79. Kumar S, Stecher G, Tamura K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol Biol Evol*. 2016;33:1870–674.
80. Mello B, Tao Q, Tamura K, Kumar S. Fast and accurate estimates of divergence times from big data. *Mol Biol Evol*. 2017;34:45–50.
81. Ramos-Onsins S, Segarra C, Rozas J, Aguadé M. Molecular and chromosomal phylogeny in the obscura group of *Drosophila* inferred from sequences of the *rp49* gene region. *Mol Phylogenet Evol*. 1998;9:33–41.
82. Gao JJ, Watabe HA, Aotsuka T, Pang JF, Zhang YP. Molecular phylogeny of the *Drosophila obscura* species group, with emphasis on the Old World species. *BMC Evol Biol*. 2007;7:87.
83. Russo CAM, Mello B, Frazão A, Voloch CM. Phylogenetic analysis and a time tree for a large drosophilid data set (Diptera: *Drosophilidae*). *Zool J Linn Soc*. 2013;169:765–75.
84. Pratedsaba R, Segarra C, Aguadé M. Inferring the demographic history of *Drosophila subobscura* from nucleotide variation at regions not affected by chromosomal inversions. *Mol Ecol*. 2015;24:1729–41.
85. Boore JL. Animal mitochondrial genomes. *Nucleic Acids Res*. 1999;27:1767–80.
86. De Ré FC, Wallau GL, Robe LJ, Loreto EL. Characterization of the complete mitochondrial genome of flower-breeding *Drosophila incompta* (Diptera, *Drosophilidae*). *Genetica*. 2014;142:525–35.
87. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 2013;30:772–80.
88. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol*. 2000;17:540–52.
89. Gregory TR, Johnston JS. Genome size diversity in the family *Drosophilidae*. *Heredity*. 2008;101:228–38.
90. Bachmann L, Raab M, Sperlich D. Satellite DNA and speciation: a species specific satellite DNA of *Drosophila guanache*. *Z Zool Syst Evol-Forsch*. 1989;27:84–93.
91. Miller WJ, Nagel A, Bachmann J, Bachmann L. Evolutionary dynamics of the SGM transposon family in the *Drosophila obscura* species group. *Mol Biol Evol*. 2000;17:1597–609.
92. Han MV, Thomas GW, Lugo-Martinez J, Hahn MW. Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. *Mol Biol Evol*. 2013;30:1987–97.
93. Tanaka R, Higuchi T, Kohatsu S, Sato K, Yamamoto D. Optogenetic activation of the *fruitless*-labeled circuitry in *Drosophila subobscura* males induces mating motor acts. *J Neurosci*. 2017;37:11662–74.
94. Bhutkar A, Schaeffer SW, Russo SM, Xu M, Smith TF, Gelbart WM. Chromosomal rearrangement inferred from comparisons of 12 *Drosophila* genomes. *Genetics*. 2008;179:1657–80.
95. Tesler G. GRIMM: genome rearrangements web server. *Bioinformatics*. 2002;18:492–3.
96. Puerma E, Orengo DJ, Aguadé M. Inversion evolutionary rates might limit the experimental identification of inversion breakpoints in non-model species. *Sci Rep*. 2017;7:17281.
97. de Nascimento L, Willis KJ, Fernández-Palacios JM, Criado C, Whittaker RJ. The long-term ecology of the lost forest of La Laguna, Tenerife (Canary Islands). *J Biogeogr*. 2009;36:499–514.
98. Charlesworth B, Coyne JA, Barton NH. The relative rates of evolution of sex-chromosomes and autosomes. *Am Nat*. 1987;130:113–46.
99. Delprat A, Guillén Y, Ruiz A. Computational sequence analysis of inversion breakpoint regions in the cactophilic *Drosophila mojavensis* lineage. *J Hered*. 2019;110:102–17.
100. Zouros E, Krimbas CB, Tsakas S, Loukas M. Genic versus chromosomal variation in natural populations of *D. subobscura*. *Genetics*. 1974;78:1223–44.
101. Koske T, Maynard-Smith J. Genetics and cytology of *Drosophila subobscura*. X. The fifth linkage group. *J Genet*. 1954;52:521–41.
102. Lankinen P, Pinsker W. Allozyme constitution of two standard strains of *Drosophila subobscura*. *Experientia*. 1977;33:1301–2.
103. Chen H, Rangasamy M, Tan SY, Wang H, Siegfried BD. Evaluation of five methods for total DNA extraction from Western corn rootworm beetles. *PLoS One*. 2010;5:e11963.
104. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res*. 2017;27:722–36.
105. Kamath GM, Shomorony I, Xia F, Courtade TA, Tse DN. HINGE: long-read assembly achieves optimal repeat resolution. *Genome Res*. 2017;27:747–56.
106. Chin CS, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, Dunn C, O'Malley R, Figueroa-Balderas R, Morales-Cruz A, et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat Methods*. 2016;13:1050–4.
107. Xiao C-L, Chen Y, Xie S-Q, Chen K-N, Wang Y, Han Y, Luo F, Xie Z. MECAT: fast mapping, error correction, and de novo assembly for single-molecule sequencing reads. *Nat Methods*. 2017;14:1072–4.
108. Kent WJ. BLAT-the BLAST-like alignment tool. *Genome Res*. 2002;12:656–64.
109. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990;215:403–10.
110. Bao W, Kojima KK, Kohany O. Repbase update, a database of repetitive elements in eukaryotic genomes. *Mob DNA*. 2015;6:11.
111. Papaceit M, Aguadé M, Segarra C, Hey J. Chromosomal evolution of elements B and C in the *Sophophora* subgenus of *Drosophila*: evolutionary rate and polymorphism. *Evolution*. 2006;60:768–81.
112. Pinsker W, Sperlich D. Cytogenetic mapping of enzyme loci on chromosomes J and U of *Drosophila subobscura*. *Genetics*. 1984;108:913–26.
113. Böhm I, Pinsker W, Sperlich D. Cytogenetic mapping of marker genes on the chromosome elements C and E of *Drosophila pseudoobscura* and *D. subobscura*. *Genetica*. 1987;75:89–101.
114. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinf*. 2011;12:491.
115. Campbell MS, Holt C, Moore B, Yandell M. Genome annotation and curation using MAKER and MAKER-P. *Curr Protoc Bioinformatics*. 2014;48:4.11.1–39.
116. Smit AFA, Hubley R, Green P. RepeatMasker Open-4.0. 2013-2015; <http://www.repeatmasker.org>.
117. Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, de Bakker PIW. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics*. 2008;24:2938–9.

118. Stanke M, Morgenstern B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* 2005;33:W465–7.
119. Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res.* 2005;33:6494–506.
120. Blanco E, Parra G, Guigó R. Using geneid to identify genes. *Curr Protoc Bioinformatics.* 2007;18:4.3.1–4.3.28.
121. Gramates LS, Marygold SJ, Santos GD, Urbano JM, Antonazzo G, Matthews BB, Rey AJ, Tabone CJ, Crosby MA, Emmert DB, et al. FlyBase at 25: looking to the future. *Nucleic Acids Res.* 2017;45:D663–71.
122. UniProt CT. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* 2018;46:2699.
123. Eilbeck K, Lewis S, Mungall C, Yandell M, Stein L, Durbin R, Ashburner M. The sequence ontology: a tool for the unification of genome annotations. *Genome Biol.* 2005;6:R44.
124. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, et al. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* 2004;32:D115–9.
125. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics.* 2014;30:1236–40.
126. Finn RD, Coghill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, et al. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* 2016;44:D279–85.
127. Finn RD, Attwood TK, Babbitt PC, Bateman A, Bork P, Bridge AJ, Chang HY, Dosztányi Z, El-Gebali S, Fraser M, et al. InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Res.* 2017;45:D190–9.
128. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. Gene ontology: tool for the unification of biology. *The Gene Ontology Consortium Nat Genet.* 2000;25:25–9.
129. The Gene Ontology Consortium. Expansion of the gene ontology knowledgebase and resources. *Nucleic Acids Res.* 2017;45:D331–8.
130. Tate R, Hall B, DeRego T, Geib S. Annie: the ANnotation Information Extractor (Version 1.0). 2014; <http://genomeannotation.github.io/annie>.
131. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics.* 2005;21:3674–6.
132. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics.* 2015;31:3210–2.
133. Bernt M, Donath A, Jühling F, Externbrink F, Florentz C, Fritzschn G, Pütz J, Middendorf M, Stadler PF. MITOS: improved de novo metazoan mitochondrial genome annotation. *Mol Phylogenet Evol.* 2013;69:313–9.
134. Li L, Stoeckert CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* 2003;13:2178–89.
135. Hahn MW, Han MV, Han S-G. Gene family evolution across 12 *Drosophila* genomes. *PLoS Genet.* 2007;3:e197.
136. Kumar S, Stecher G, Suleski M, Hedges SB. TimeTree: a resource for timelines, timetrees, and divergence times. *Mol Biol Evol.* 2017;34:1812–9.
137. Supek F, Bosnjak M, Skunca N, Smuc T. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One.* 2011;6:e21800.
138. Soderlund C, Nelson W, Shoemaker A, Paterson A. SyMAP: a system for discovering and viewing syntenic regions of FPC maps. *Genome Res.* 2006;16:1159–68.
139. Soderlund C, Bomhoff M, Nelson WM. SyMAP v3.4: a turnkey synteny system with application to plant genomes. *Nucleic Acids Res.* 2011;39:e68.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions



3.2 Chapter 2: The cyclically seasonal *Drosophila subobscura* inversion O₇ originated from fragile genomic sites and relocated immunity and metabolic genes



The Cyclically Seasonal *Drosophila subobscura* Inversion O₇ Originated From Fragile Genomic Sites and Relocated Immunity and Metabolic Genes

Charikleia Karageorgiou*, Rosa Tarrío* and Francisco Rodríguez-Trelles*

Grup de Genòmica, Bioinformàtica i Biologia Evolutiva (GGBE), Departament de Genètica i de Microbiologia, Universitat Autònoma de Barcelona, Barcelona, Spain

OPEN ACCESS

Edited by:

Pedro Simões,
University of Lisbon, Portugal

Reviewed by:

Stephen Wade Schaeffer,
Pennsylvania State University (PSU),
United States
Priscilla Erickson,
University of Virginia, United States
Emma Berdan,
University of Gothenburg, Sweden

*Correspondence:

Charikleia Karageorgiou
charikleia.karageorgiou@uab.cat
Rosa Tarrío
rosamaria.tarrío@uab.cat
Francisco Rodríguez-Trelles
franciscojose.rodruigueztrrelles@uab.cat

Specialty section:

This article was submitted to
Evolutionary and Population Genetics,
a section of the journal
Frontiers in Genetics

Received: 26 May 2020

Accepted: 09 September 2020

Published: 09 October 2020

Citation:

Karageorgiou C, Tarrío R and
Rodríguez-Trelles F (2020) The
Cyclically Seasonal
Drosophila subobscura Inversion O₇
Originated From Fragile Genomic
Sites and Relocated Immunity
and Metabolic Genes.
Front. Genet. 11:565836.
doi: 10.3389/fgene.2020.565836

Chromosome inversions are important contributors to standing genetic variation in *Drosophila subobscura*. Presently, the species is experiencing a rapid replacement of high-latitude by low-latitude inversions associated with global warming. Yet not all low-latitude inversions are correlated with the ongoing warming trend. This is particularly unexpected in the case of O₇ because it shows a regular seasonal cycle that peaks in summer and rose with a heatwave. The inconsistent behavior of O₇ across components of the ambient temperature suggests that is causally more complex than simply due to temperature alone. In order to understand the dynamics of O₇, high-quality genomic data are needed to determine both the breakpoints and the genetic content. To fill this gap, here we generated a PacBio long read-based chromosome-scale genome assembly, from a highly homozygous line made isogenic for an O₃₊₄₊₇ chromosome. Then we isolated the complete continuous sequence of O₇ by conserved synteny analysis with the available reference genome. Main findings include the following: (i) the assembled O₇ inversion stretches 9.936 Mb, containing > 1,000 annotated genes; (ii) O₇ had a complex origin, involving multiple breaks associated with non-B DNA-forming motifs, formation of a microinversion, and ectopic repair in *trans* with the two homologous chromosomes; (iii) the O₇ breakpoints carry a pre-inversion record of fragility, including a sequence insertion, and transposition with later inverted duplication of an *Attacin* immunity gene; and (iv) the O₇ inversion relocated the major insulin signaling *forkhead box subgroup O (foxo)* gene in tight linkage with its antagonistic regulatory partner *serine/threonine-protein kinase B (Akt1)* and disrupted concerted evolution of the two inverted *Attacin* duplicates, reattaching them to dFOXO metabolic enhancers. Our findings suggest that O₇ exerts antagonistic pleiotropic effects on reproduction and immunity, setting a framework to understand its relationship with climate change. Furthermore, they are relevant for fragility in genome rearrangement evolution and for current views on the contribution of breakage versus repair in shaping inversion-breakpoint junctions.

Keywords: non-B DNA, genome fragility, *foxo* (forkhead box subgroup O), *Akt1* (serine/threonine-protein kinase B), *Attacin* antibacterial genes, immunometabolism, thermal adaptation, seasonal selection

INTRODUCTION

Chromosome inversions are arguably the genetic traits with the earliest and richest record of associations with climate (Hoffmann and Rieseberg, 2008). Research into evolutionary responses to contemporary global warming (Hughes, 2000; Parmesan, 2006) is therefore faced with the challenge of understanding how inversions originate and spread in populations (Kirkpatrick, 2010), while trying to determine their roles in climatic adaptation (Gienapp et al., 2008; Messer et al., 2016).

Chromosome inversions are ubiquitous chromosomal mutations consisting in the reversal of the orientation of a chromosome segment. They originate through either of two major mechanisms, each with its associated distinctive footprints. The first mechanism is intrachromatid non-allelic homologous recombination (NAHR) between inversely oriented repeats. This mechanism generates inversions with duplications at their ends in both the inverted and uninverted states (Cáceres et al., 1999). The second mechanism is chromosomal breakage and ectopic repair via non-homologous end joining (NHEJ). This mechanism either does not generate duplications or generates them but at the ends of the inverted state only. These two types of NHEJ footprints have been explained in terms of differences in the mode of breakage. Two modes of breakage have been advanced: “cut-and-paste” via clean double-strand breaks (DSBs) that generate blunt ends and staggered. NHEJ inversions without duplications at their ends would originate via cut-and-paste (Wesley and Eanes, 1994), whereas those with inverted duplications at their ends would originate via staggered breaks in one or the two breakpoints. Two staggering models for the origin of the inverted duplications have been proposed (Kehrer-Sawatzki et al., 2005; Matzkin et al., 2005; Ranz et al., 2007): according to the isochromatid model, the duplications would be the filled-in single-stranded overhangs that would result from paired single strand breaks (SSBs) located staggered with each other on opposite strands of the same chromatid (Kehrer-Sawatzki et al., 2005), whereas according to the chromatid model, the duplications would result from unequal exchange between paired sister chromatids, each with one of two paired staggered DSBs at each breakpoint (Matzkin et al., 2005). Note that here the terms *isochromatid* and *chromatid* have switched meanings relative to how they are used in cytogenetics (Savage, 1976). The two staggering models are chromatid models because they assume that inversions originate from either single chromatids during premeiotic mitosis (isochromatid), or paired sister chromatids from the same chromosome during meiotic prophase (chromatid) (Ranz et al., 2007). The models cannot be distinguished based on the pattern of inverted duplications. Yet the chromatid model has been favored over the isochromatid model, because of the length of DNA that would need to be unwound by enzymatic activity in the latter model (Ranz et al., 2007). The chromatid model is also not without potential caveats because NHEJ was found to be suppressed during the meiotic prophase in *Drosophila* (Joyce et al., 2012; Hughes et al., 2018). The prevalence and distribution of the NAHR and NEHJ mechanisms of inversion formation within and across lineages are currently under debate (Ranz et al., 2007; Delprat et al.,

2019). The NEHJ mechanism rests upon the occurrence of two or more DSBs. But the source of the DSBs (whether environmental, such as ionizing radiation, or spontaneous, such as non-B DNA-associated sequence instability, where non-B DNA denotes any DNA conformation that is not in the canonical right-handed B form; Lobachev et al., 2007; Zhao et al., 2010; Farré et al., 2015), the relative contributions of breakage versus repair to shaping breakpoint junctions (Ranz et al., 2007; Kramara et al., 2018; Scully et al., 2019), and the relative frequency with which the joined broken ends are from the same chromatid (isochromatid model) versus two distinct sisters (chromatid model) (Ranz et al., 2007) or even, as has been more recently suggested by Orengo et al. (2019), non-sister chromatids (chromosome model) are additional open questions.

Inversions can have direct or/and indirect functional effects (Kirkpatrick, 2010). Direct effects are those ascribable to the mutation *per se*, as it altered the structure or expression of functional sequences at the breakpoints, or the functional neighborhood of genes in the cell nucleus (McBroome et al., 2020). Indirect effects emanate from their associated recombination-suppression effects when in heterozygous condition, whereby they can bind together into close linkage association particular combinations of alleles at genetically distant loci. The evolutionary significance of polymorphic inversions is often thought to chiefly stem from their indirect effects (Dobzhansky, 1947; Wasserman, 1968; Kirkpatrick and Barton, 2006). Although data have been lacking on the relative importance of the two types of effects, there has been renewed interest in using genomics to determine mechanisms for the spread, establishment, and maintenance or fixation of inversions (Corbett-Detig and Hartl, 2012; Corbett-Detig, 2016; Fuller et al., 2016, 2017, 2019; Cheng et al., 2018; Said et al., 2018; Lowry et al., 2019). Because they usually involve many genes, chromosome inversions have enhanced potential for affecting multiple traits, which should expand the opportunities for their maintenance via balancing selection. The extent to which that is the case and the types and transience of the balancing selection mechanisms involved are only beginning to be elucidated (Kapun and Flatt, 2018; Wellenreuther and Bernatchez, 2018; Faria et al., 2019). Amid these unknowns, the inversion polymorphisms of *Drosophila subobscura* emerged among the first genetic traits identified as involved in a species' adaptation to contemporary climate warming (Rodríguez-Trelles and Rodríguez, 1998, 2007; Balanyà et al., 2006; Rezende et al., 2010).

Drosophila subobscura is a native Palearctic species broadly distributed in Europe and the newly invaded areas of North and South America (reviewed in Krimbas, 1992), where it is found generally associated with woodland habitats. It belongs in the *obscura* group, within which it clusters with the recently derived small-island endemics *Drosophila guanche* and *Drosophila madeirensis*, forming the *subobscura* three-species subgroup (Bächli, 2020). *D. subobscura* has one of the smallest and least repetitive *Drosophila* reference genomes obtained thus far, which is distributed among five large telocentric chromosomes (A, J, U, E, and O) and one small dot (Karageorgiou et al., 2019). In stark contrast with its two insular relatives, the species has evolved highly rearranged chromosome sequences, which is due

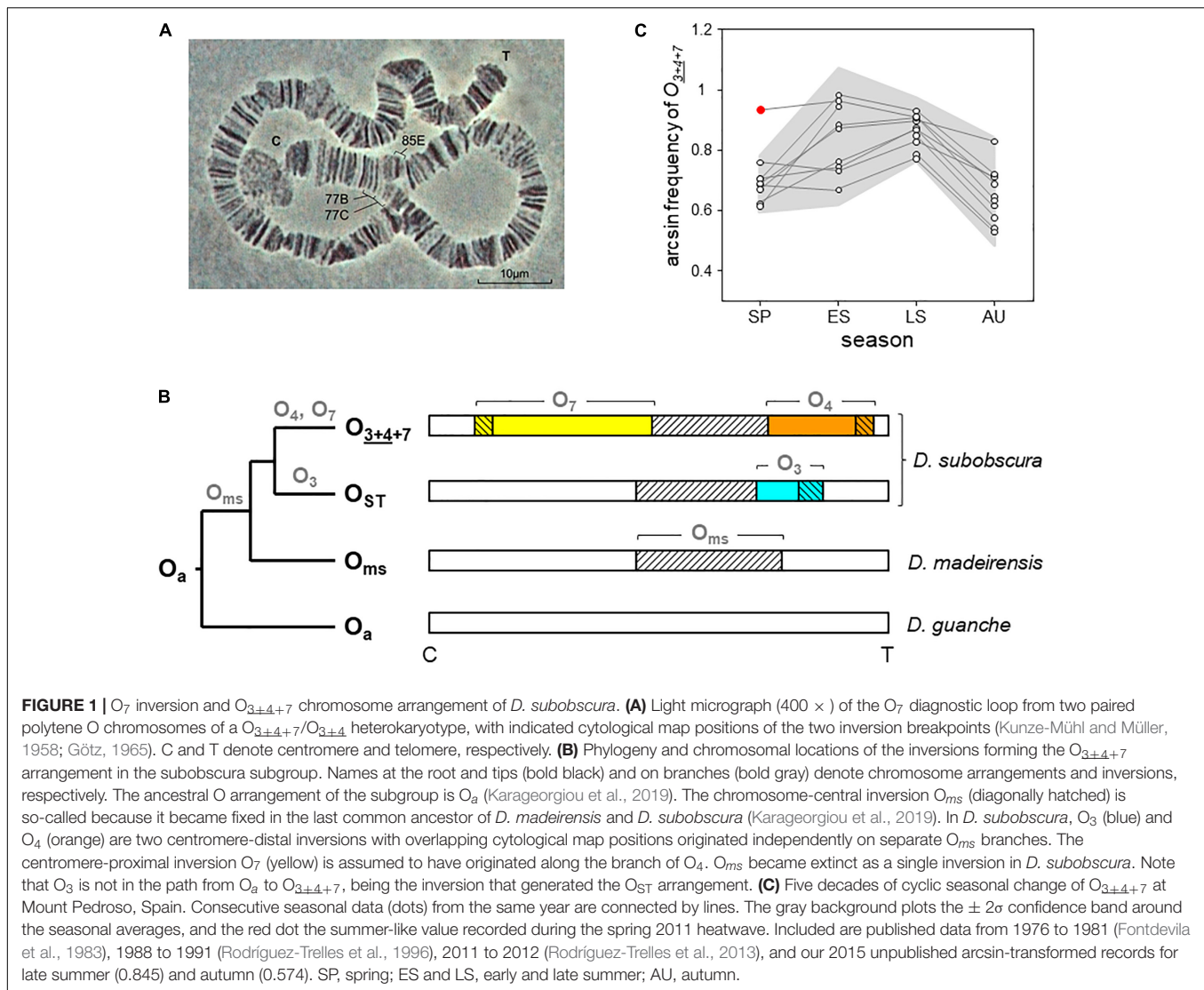
to having experienced accelerated fixation rates of paracentric inversions, especially the A sex chromosome. This situation has been interpreted as indicative of the inversions' role in binding together adaptive alleles in the face of the species' intense continent-wide gene flow (Karageorgiou et al., 2019). Presently, *D. subobscura* harbors a rich inversion polymorphism, with its five major chromosomes showing parallel adaptive variation patterns across latitude (Ayala et al., 1989), seasons (Rodríguez-Trelles et al., 1996, 2013), and through a heatwave (Rodríguez-Trelles et al., 2013), while rapidly shifting in close association with the ongoing rise in global temperatures (Rodríguez-Trelles and Rodríguez, 1998, 2010; Balanyà et al., 2006). Laboratory attempts to establish the causal nature of this association have, however, largely been inconclusive (Santos et al., 2005; Fragata et al., 2014). Ultimately, a complete understanding of the role of inversions in adaptation to contemporary climate warming in *D. subobscura* will necessarily include the identities and functional properties of the genome sequences affected by them. Advances along this line include the isolation and characterization of breakpoint sequences for 11 of the more than 65 large cytologically visible inversions known for the species, including A₂ (Puerma et al., 2017), O₃ (Papacit et al., 2012), O₄ and O₈ (Puerma et al., 2016a), E₁ and E₂ (Puerma et al., 2014), E₃ and E₉ (Orengo et al., 2015), E₁₂ (Puerma et al., 2016b), and U₁ and U₂ (Karageorgiou et al., 2019). An overall conclusion is that none of these inversion breakpoints disrupted any obvious candidate gene for direct adaptation to temperature, despite the fact that all but the E₃ inversion are supposed to be involved in adaptation to climate (e.g., Menozzi and Krimbas, 1992; Rego et al., 2010; Arenas et al., 2018). Apart from the fact that thermal traits are genetically complex and that many of the genes that impinge on them are still unknown, the above conclusion supports that those inversions' role in thermal adaptation would be through either position effects, indirect linkage generation effects, or both.

As part of a wider effort to develop a high-quality reference genome for *D. subobscura* encompassing the species' rich chromosomal polymorphisms, here we focus on the O₇ inversion. The breakpoints of this inversion were located cytologically at subsections 77B/C and 85E on the Kunze-Mühl and Müller *standard* map (Figure 1A; Kunze-Mühl and Müller, 1958; Götz, 1965). O₇ is among the top 10% known largest *D. subobscura* inversions, stretching most of the centromere-proximal half of the O chromosome (Krimbas, 1992). In nature, it attains significant frequencies only in combination with the non-overlapping centromere-distal complex of two overlapping inversions O₃₊₄, forming the chromosome arrangement O₃₊₄₊₇ (Figure 1B). The tight association between O₇ and O₃₊₄ is likely maintained by an interaction between selection and the strongly reduced recombination between them (Pegueroles et al., 2010a).

O₇ could be initially classified as a warm-climate inversion. In the Palearctic, it shows a southern distribution. In northwest Spain, where it has been longitudinally monitored starting in mid-1970s (Fontdevila et al., 1983; Rodríguez-Trelles et al., 1996, 2013), it shows a pronounced regular seasonal cycle (estimated to account for more than 60% of the inversion's

temporal variation; Rodríguez-Trelles et al., 1996) that peaks in summer and drops in winter (Figure 1C). In 2011, it rose to summer-like levels in spring during a heatwave, with the magnitude of the increase closely matching that of the thermal anomaly (Figure 1C; Rodríguez-Trelles et al., 2013). However, (i) the average annual frequency of O₇ in northwest Spain remains unchanged after decades of sustained climate warming experienced by the region (Rodríguez-Trelles et al., 2013; our unpublished records). (ii) Following the 2011 heatwave, the inversion reached summer-like frequencies in April, but did not continue rising through the ensuing summer (Figure 1C), perhaps hampered by recessive deleterious alleles (Rodríguez-Trelles et al., 2013). (iii) The Palearctic distribution of O₇ is disjointed between the peninsulas of Iberia and Turkey (Götz, 1967). These are similar latitude areas separated by ~2,500 km within the continuous species' range. Assuming that the inversion is molecularly the same in the two areas, this spatial pattern can hardly be explained on the sole basis of a postglacial expansion scenario (Menozzi and Krimbas, 1992), considering how rapidly it spread through the recently invaded areas of the New World (Prevosti et al., 1988). And (iv) in the more studied Iberian Peninsula, the distribution of the inversion has negative or no correlations with the geographical variation in temperature. For example, the average annual frequency of the inversion declines from ~50% to near-zero values along the > 1,000-km stretching from the northwestern-most to the northeastern-most territories, despite the latter having a warmer climate than the former (de Frutos, 1972; Solé et al., 2002; Rodríguez-Trelles et al., 2013). The same is true for the West Atlantic fringe of the peninsula along which the inversion levels remain basically the same despite the fact that it stretches seven latitudinal degrees of steep thermal gradient (Brehm and Krimbas, 1988; Solé et al., 2002; Rodríguez-Trelles et al., 2013). The inconsistent patterns of O₇ between components of the ambient temperature suggest that it is influenced by selective factors other than temperature alone.

The O chromosome offers the methodological advantage over the other *D. subobscura* chromosomes that there is an available balancer-strain called *Varicose/Bare* (*Va/Ba*) (Sperlich et al., 1977). In this study, we first used the *Va/Ba* strain to develop an isogenic line with two identical copies of a wild O chromosome carrying the O₃₊₄₊₇ arrangement. Second, we used PacBio long-read technology to generate a high-quality annotated chromosome-scale genome sequence for the line. Third, we isolated the complete continuous nucleotide sequence of the inversion O₇ by conserved synteny analysis of the obtained O₃₊₄₊₇ chromosome with the available O chromosome from the species' reference genome, which is structurally O₃₊₄ (Karageorgiou et al., 2019). In addition, we also considered two other published sequences of the O chromosome, including a high-quality long-read-based sequence from *D. subobscura* (Bracewell et al., 2019), and an Illumina-based sequence from *D. guanche* (Puerma et al., 2018). We give an account of O₇ main features, together with a detailed description of its mechanism of formation. Our findings provide clues to the mixed evidence for this inversion's role in thermal adaptation.



MATERIALS AND METHODS

Species Karyotype and Inversion Nomenclature

Drosophila subobscura shows the ancestral karyotype configuration of the genus *Drosophila*, consisting of five large telocentric rods (Muller elements A-E) and one dot (Muller F) (Powell, 1997). The five rods include the sex chromosome (Muller A) and four autosomes of which the O chromosome (Muller E; homologous to chromosome arm 3R from *D. melanogaster*) is the largest (~30 Mb), comprising around 25% of the species' nuclear euchromatic genome (~125 Mb; Karageorgiou et al., 2019).

An early landmark in the study of chromosomal inversion polymorphisms of *D. subobscura* was the development of structurally homozygous strains, as tools to identify new inversions by the location and shape of the loops formed in inversion heterozygotes (Zollinger, 1950; Maynard-Smith and Maynard-Smith, 1954; Zouros et al., 1974; Loukas et al., 1979). The “Küsnacht” strain, named after the Swiss locality of collection

of the flies (Zollinger, 1950), became the first established (Koske and Maynard-Smith, 1954). The chromosomal arrangements of the strain, which happened to be those most common in Central Europe, were subscripted ST (for “standard”) and from them new inversions were designated with numeral subindices following their order of discovery (Kunze-Mühl and Sperlich, 1955). This naming system was not intended to convey polarity of evolutionary change. Accordingly, O_{3±4+7} is the arrangement that can be interconverted with O_{ST} by the two centromere-distal overlapping inversions O₃ and O₄ (denoted by the underline joining the subscripts; Zouros et al., 1974) and the centromere-proximal inversion O₇. The ancestor-descendant relationships of these inversions are shown in **Figure 1B**.

Drosophila Lines

O chromosome conserved synteny analysis was based on data from four whole-genome *de novo* assemblies, including three PacBio long-read-based assemblies from *D. subobscura* and one Illumina short-read-based assembly from *D. guanche*. Of the

three *D. subobscura* assemblies, one was used as reference for inversion O₇ and was newly generated in this study. The other two were used as references for the *standard* configuration [note that the distal breakpoint of O₇ maps within inversion O_{ms} (Karageorgiou et al., 2019), whereby is expected to exhibit opposite orientation in *D. subobscura* relative to *D. guanche*; **Figure 1B**] and were already available (Karageorgiou et al., 2019; Bracewell et al., 2019). Also available was the assembly from *D. guanche* (Puerma et al., 2018), which was used as an outgroup. Henceforth, we will refer to these four assemblies as Ds_7, Ds_ch-cu, Ds_B, and Dg, respectively.

To generate the Ds_7 assembly, we developed a line that is isogenic for an O₃₊₄₊₇ arrangement from the wild and homokaryotypic and highly homozygous for the ST arrangements of the rest of the chromosomes (i.e., A_{ST}, J_{ST}, U_{ST}, E_{ST}, and O₃₊₄₊₇). The O arrangement was first isolated by crossing wild males to virgin females from the *cherry-curved* (*ch-cu*) recessive marker stock; they were then submitted to nine generations of backcrossing with *ch-cu* females and finally isogenized using the *Va/Ba* balancer stock (Sperlich et al., 1977). The expression of the *Ba* gene is highly variable. Therefore, to prevent potential errors at sorting out phenotypically O₃₊₄₊₇ homokaryotypes, the *Va/Ba* stock was previously selected for zero macrobristles on the scutum and scutellum. Crossing schemes and the methods for polytene chromosome staining and identification are described elsewhere (Rodríguez-Trelles et al., 1996). The assayed line was stored frozen at -80°C immediately upon obtention. The wild flies used to develop the line were derived from our survey of the natural population of Berbiz (Spain; Lat.: 43,18949, Long.: -3,09025, Datum: WGS84, elevation: 219 m a.s.l) conducted in July 7, 2012 (Rodríguez-Trelles et al., 2013).

The remaining three assemblies were derived from strains homokaryotypic for all chromosomes. The Ds_ch-cu assembly was generated from the *ch-cu* strain of our laboratory (A_{ST}, J_{ST}, U_{ST}, E_{ST}, and O₃₊₄; Karageorgiou et al., 2019) and the Ds_B assembly from an isofemale laboratory stock derived from a natural population from Eugene, Oregon, in 2006 (A_{ST}, J_{ST}, U₁₊₂, E_{ST}, and O₃₊₄; Bracewell et al., 2019). The Dg assembly was generated from an isofemale laboratory stock derived from a natural population from the Canary Islands, Spain, in winter 1999 (Puerma et al., 2018); it shows the chromosome configuration of the last common ancestor of the *subobscura* subgroup except for chromosome E, which carries the arrangement E_{g1} (A_a, J_a, U₁₊₂, E_{g1}, and O_a; Puerma et al., 2018; Karageorgiou et al., 2019; Bracewell et al., 2019).

High Molecular Weight Genomic DNA Isolation and PacBio Whole-Genome Sequencing

High-quality high-molecular-weight gDNA was obtained from 60 mg of -80°C frozen adult females, using a modified version of the phenol/chloroform method of Chen et al. (2010) that yields ~25 µg of high-quality DNA per assay, as assessed by NanoDrop ND1000 (NanoDrop Technologies Inc., Wilmington, DE, United States) spectrophotometer and standard agarose

gel electrophoresis. The genome of the Ds_7 isogenic line was sequenced to nominal 66-fold genome coverage using PacBio (Pacific Biosciences, Menlo Park, CA, United States) Sequel single-molecule real-time (SMRT) technology from a 20-kb SMRTbell template library, using Polymerase 3.0 chemistry and two SMRT cells. Libraries construction and PacBio sequencing were outsourced to MacroGen (MacroGen Inc., Seoul, South Korea).

Chromosome-Scale Assembly and Scaffolding

Raw PacBio reads were assembled using the Canu assembler (version 1.8; Koren et al., 2017) on recommended settings for read error correction, trimming and assembly, and genome size set at 150Mb based on previously published flow cytometry data (Karageorgiou et al., 2019). These analyses were performed on a 2.80-GHz 8-CPU Intel Xeon 64-bit 32GB-RAM computer running Ubuntu 18.04 LTS.

Chromosome-scale assembly and scaffolding followed the four steps outlined in Karageorgiou et al. (2019) as well as a fifth step, to improve genome completeness and contiguity, consisting of merging the Ds_7 assembly with a preselected set of segments from the reference Ds_ch-cu assembly using one round of quickmerge (Chakraborty et al., 2016), as follows: first, the CANU contigs that could be certainly anchored, ordered, and oriented on the nuclear chromosomes were aligned against the Ds_ch-cu reference using NUCmer (Kurtz et al., 2004). Second, the segments of Ds_ch-cu not overlapped by the CANU contigs, each extended 10 kb outward from each of its two ends, were extracted. Finally, separately for each chromosome, the extracted Ds_ch-cu segments, together with the CANU contigs set as the backbone, were fed into quickmerge. This approach was found to reduce the chances of misassembly and chimerism, while making it straightforward to trace the non-backbone sequence in the assembly. Dot plots of the merged assembly against the reference Ds_ch-cu assembly were used as a further step of misassembling correction. The obtained Ds_7 assembly was polished with 26 × mean coverage of 150-base-pair (bp) MP Illumina reads from the O₃₊₄₊₇ isogenic line using two rounds of Pilon (version 1.23; Walker et al., 2014).

Genome Annotation

Gene prediction and annotation of the assembled genome were conducted using the MAKER (version 3.01.02.-beta; Holt and Yandell, 2011; Campbell et al., 2014) annotation pipeline. Repetitive elements were identified using RepeatMasker (version 4.0.6; Smit et al., 2013/2015, at¹) combined with three repeat libraries, including (i) the *Drosophila* genus-specific repeat library contained in the Repbase database (release 20170127; Bao et al., 2015); (ii) a library of *subobscura* subgroup specific satellites, sat290 and SGC-sat (Karageorgiou et al., 2019); and (iii) a library of *de novo* identified repeats generated using RepeatModeler (version 1.0.11) on the assembly masked for the first two libraries. Novel long terminal repeats (LTRs), miniature inverted-repeat transposable elements (MITEs), tandem repeats,

¹<http://repeatmasker.org>

and rDNA and tDNA genes were identified using LTRharvest (GenomeTools version 1.5.10; Ellinghaus et al., 2008), MITE Tracker (version 2.7.1; Crescente et al., 2018), Tandem Repeat Finder (TRF; version 4.09; Benson, 1999), RNAmmer (version 1.2; Lagesen et al., 2007), and tRNAscan-SE (version 2.0; Lowe and Chan, 2016), respectively. All tools were run on default settings, except LTRharvest, for which we set -seed 100, -similar 90.0, and -mintsd 5, following Hill and Betancourt (2018). The quality of the annotation was controlled using the Annotation Edit Distance (AED) metric (Eilbeck et al., 2005). AED values are bounded between 0 and 1. An AED value of 0 indicates perfect agreement of the annotation to aligned evidence, and conversely, a value of 1 indicates no evidence support.

Functional annotation of MAKER-predicted proteins was made by BLASTP (version 2.6.0+) searches against the *Drosophila* UniProt-SwissProt manually curated datasets (Apweiler et al., 2004). Prediction of protein functional domains was accomplished using InterProScan (version 5.29–68.0; Jones et al., 2014) on the Pfam (Finn et al., 2016), InterPro (Finn et al., 2017), and Gene Ontology (Ashburner et al., 2000; The Gene Ontology Consortium, 2017) domain databases. Genome assembly and annotation completeness were gauged using the Benchmarking Universal Single-Copy Orthologs (BUSCO) tool [BUSCO, version 4 (Seppy et al., 2019)], with the latest update of the dipteran gene set (diptera_odb10), which contains 3,285 highly conserved, single-copy genes expected to be present in any dipteran genome.

Isolation and Characterization of the O₇ Breakpoints

Suppose that +A|+B+C|+D and +A|–C–B|+D represent two chromosome arrangements whose gene orders differ only by the orientation of the segment between A and D (with symbols denoting A and D, the segments upstream from the centromere-proximal breakpoint and downstream from the centromere-distal breakpoint, respectively; vertical bars, breakpoint junctions; and plus/minus signs, orientation of the segment relative to the uninverted sequence). We proceeded in two steps. First, we isolated the regions containing the breakpoint junctions by chromosome conserved synteny analysis between the uninverted and inverted states using the Synteny Mapping and Analysis Program (SyMAP, version 4.2.; Soderlund et al., 2011) tool on default options, and NUCmer (see Karageorgiou et al., 2019). The O₇ breakpoints were identified as the loci of interrupted synteny whose locations and distance from each other agree with the cytogenetic mapping data of the inversion (Karageorgiou et al., 2019). Second, we localized the breakpoint junctions at base-pair resolution and performed comparative analyses of their flanking sequences using the progressive guide tree-based MAFFT algorithm (version 7²) with the accuracy-oriented method “L-INS-i” (Katoh et al., 2019). Each of the regions +A|+B and +C|+D from the uninverted state was aligned separately, first with +A|–C and then with –B|+D from the inverted state. From each of the four resulting alignments, we used the regions showing positional homology between the

uninverted and inverted states to isolate segments A, B, C, and D, correspondingly. The remaining sequences of the uninverted state were submitted to a second round of comparative analysis among them, and with segments A to D to identify the homologies missed in the first round. As representatives of the uninverted state, we used Ds_ch-cu together with the previously published assemblies Ds_B and Dg, and this last one was set as the outgroup.

Phylogenetic Inferences

MAFFT-based tree reconstruction of the *Attacin* gene family in *Drosophila* was performed via maximum likelihood. Model selection and tree inference were conducted using IQ-Tree (Kalyaanamoorthy et al., 2017; Nguyen et al., 2015). Tree searches were conducted starting from sets of 100 initial maximum parsimony trees using nearest neighbor interchange with default perturbation strength and a stopping rule settings. Branch support was assessed using the ultrafast bootstrap approximation (UFboot; 1,000 replicates) (Hoang et al., 2018), and two single-branch tests including the Shimodaira–Hasegawa-like approximate likelihood ratio test (SH-aLRT; 1,000 replicates) (Guindon et al., 2010) and the approximate Bayes parametric test (Anisimova et al., 2011).

Non-B DNA Sequence and Transcription Factor Binding Site Scans

Scans for potential non-B DNA-forming sequences considered the following features: inverted repeats (IRs) (capable of forming hairpin and/or cruciform DNA), direct/tandem repeats (slipped/hairpin structures), mirror repeats (triplexes), alternate purine-pyrimidine tracts (left-handed Z-DNA), G4 motifs (tetraplex and G–quadruplex DNA), and A–phased repeats (static bending). Searches were conducted online using for IRs Palindrome Analyzer (Brázda et al., 2016³; accessed January 24, 2020) with repeat length of 6–20 nt, spacer length ≤ 10 nt, and number of mismatches ≤ 1; for tandem repeats Tandem Repeat Finder (TRF version 4.09; Benson, 1999⁴; accessed Jan 24, 2020) in basic mode; and for the remaining features nBMST (Cer et al., 2012⁵; accessed January 24, 2020) with prefixed default settings. The propensity of IRs to adopt non-B conformation was assessed using the difference in free energy between the DNA sequence in the linear and cruciform structures, as implemented in Palindrome Analyser (Brázda et al., 2016).

Transcription start site (TSS) prediction was conducted using the NNPP method (Reese, 2001⁶). Searches for putative binding sites for Relish (*Rel*), the heterodimer Dif/Rel, dFOXO, Dorsal (*dl*), and Serpent (*srp*) transcription factors in the 1-kb upstream region of the *Attacin* predicted TSSs were performed using the FIMO tool (Grant et al., 2011) from the MEME suite (Bailey et al., 2015). For *Rel* and Dif/Rel, and for dFOXO,

³<http://bioinformatics.ibp.cz:9999/#/en/palindrome>

⁴<https://tandem.bu.edu/trf/trf.html>

⁵<https://nonb-abcc.ncifcrf.gov/apps/nBMST/default>

⁶https://www.fruitfly.org/seq_tools/promoter.html

²<http://mafft.cbrc.jp/alignment/software/>

we used the FootprintDB database (Sebastián and Contreras-Moreira, 2014⁷) *Drosophila melanogaster* Major Position Matrix Motifs (DMMPMM) identified, respectively, by Senger et al. (2004) and Weirauch et al. (2014). For *dl* and *srp*, we used the REDfly database (version 5.5.3; Rivera et al., 2019⁸) improved iDMMPMM motifs developed by Kulakovskiy and Makeev (2009). Searches were performed using a *p* value cutoff of 10⁻³.

RESULTS

Chromosome-Scale Assembly and Annotation of Chromosome Arrangement O₃₊₄₊₇

The PacBio Sequel sequencing of the O₃₊₄₊₇ isogenic line genome generated 2,457,493 reads, with mean and longest lengths of 11,257 bp and 117,750 bp, respectively. Canu correction and trimming retained a 42-fold genome coverage for the assembly. Of the 385 Canu-generated contigs, the 14 that could be confidently anchored, ordered, and oriented covered the complete reference genome, with an added length of 126.770 Mb and N50 of 10.587 Mb. Quickmerge of those 14 CANU contigs resulted in six chromosome-scale scaffolds, one per each of the major *D. subobscura* chromosomes (Table 1). Of note, chromosome O was built from two contigs only, with the centromere-proximal contig (tig00026085; 29.679 Mb) spanning almost all the chromosome length (96.9%) (Figure 2A). The Ds₇ assembly contained 13,459 MAKER-annotated genes, nearly all with well-supported predictions (AED₅₀ = 99.3%). Only 2.6% (87) of the BUSCO genes were missing, indicating that the assembly is almost complete. The O chromosome contained 3,220 (23.9%) of the annotations of the assembly.

Identification of Inversion O₇ Using Chromosome Conserved Synteny Analysis

The structural transition between the O chromosomes of the Ds₇ and Ds_{ch-cu} assemblies called for one large megabase-sized inversion (Figures 2B,C), whose breakpoints located

⁷<http://floresta.eead.csic.es/footprintdb>

⁸<http://redfly.ccr.buffalo.edu/>

TABLE 1 | Ds₇ assembly summary statistics (Muller elements are given in parenthesis, and lengths are given in megabases of sequence).

Component	Length	Scaffolds	Canu contigs	Largest Canu contig	Gene models
Nuclear genome	126.770	6	14	29.679	13,459
Dot (F)	1.412	1	1	1.412	96
A (A)	22.941	1	2	17.229	2,323
J (D)	25.018	1	3	10.587	2,652
U (B)	26.010	1	3	13.133	2,561
E (C)	20.783	1	3	9.524	2,607
O (E)	30.629	1	2	29.679	3,220

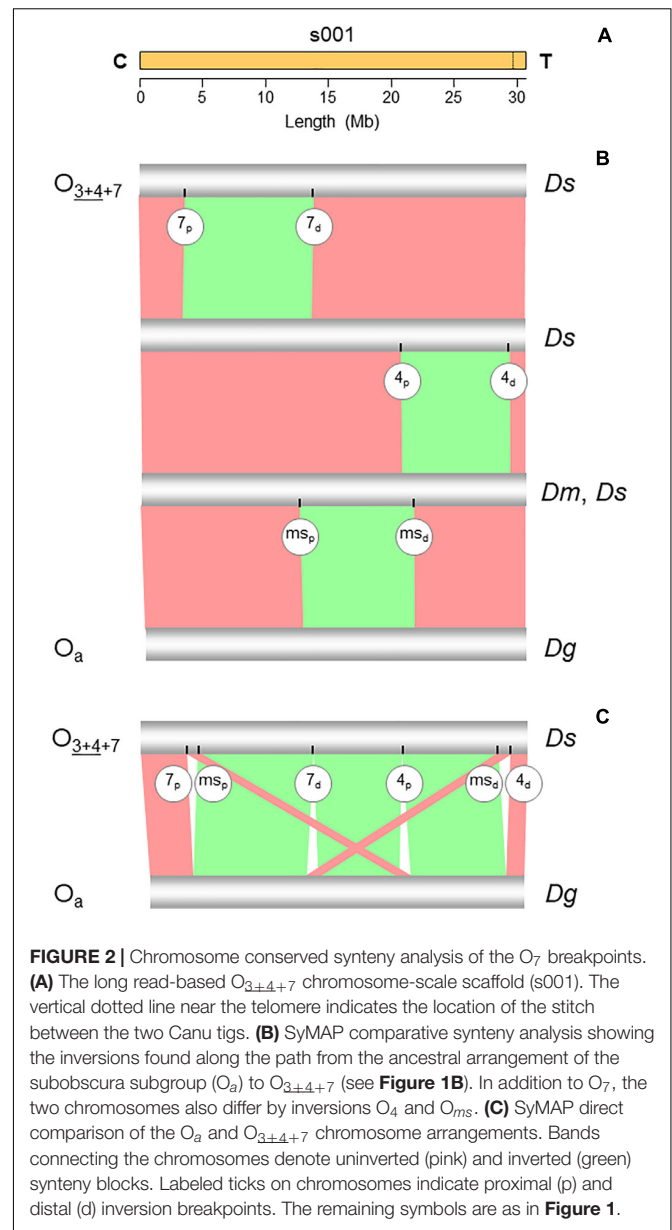
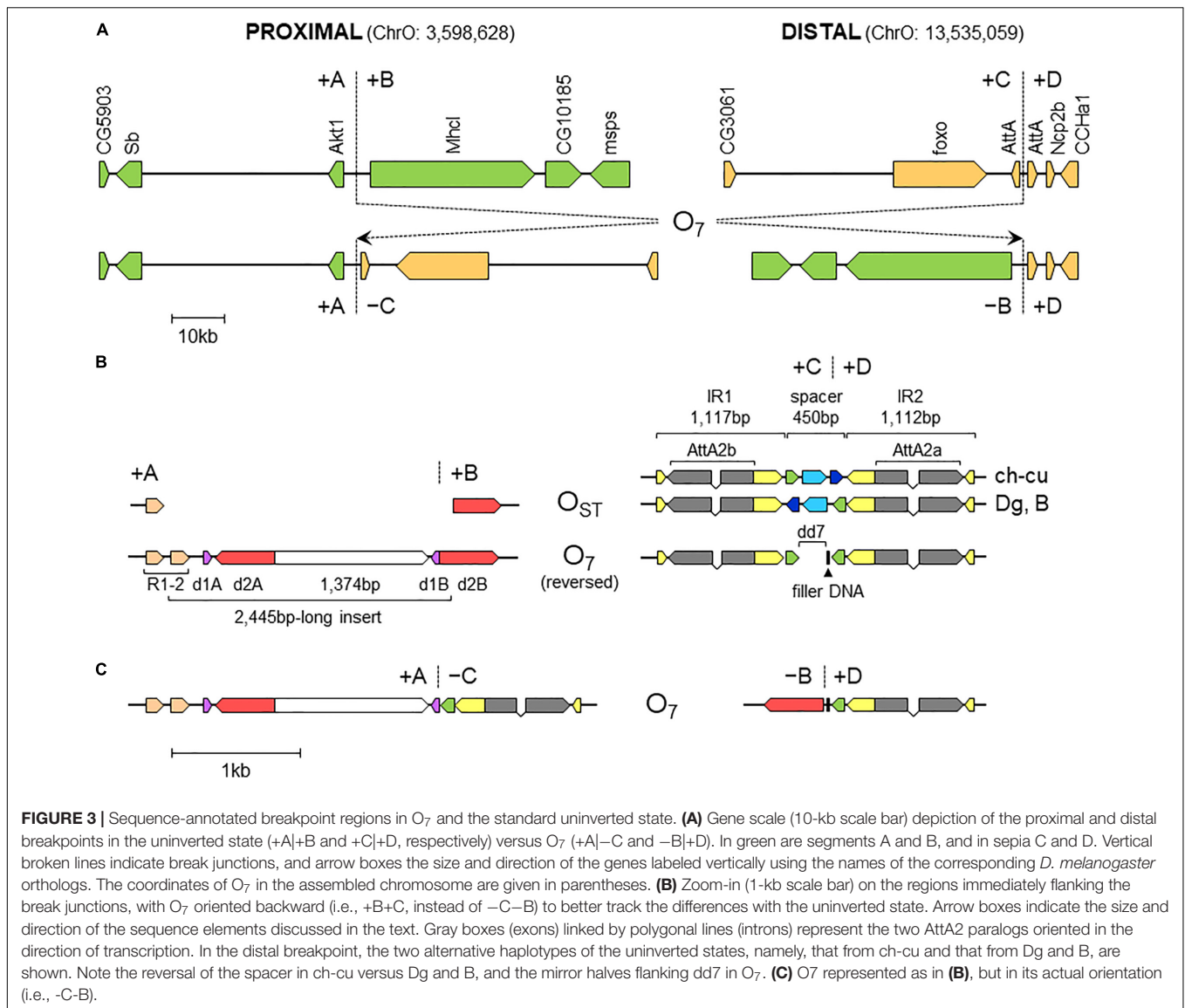


FIGURE 2 | Chromosome conserved synteny analysis of the O₇ breakpoints. **(A)** The long read-based O₃₊₄₊₇ chromosome-scale scaffold (s001). The vertical dotted line near the telomere indicates the location of the stitch between the two Canu tigs. **(B)** SyMAP comparative synteny analysis showing the inversions found along the path from the ancestral arrangement of the subobscura subgroup (O_a) to O₃₊₄₊₇ (see Figure 1B). In addition to O₇, the two chromosomes also differ by inversions O₄ and O_{ms}. **(C)** SyMAP direct comparison of the O_a and O₃₊₄₊₇ chromosome arrangements. Bands connecting the chromosomes denote uninverted (pink) and inverted (green) synteny blocks. Labeled ticks on chromosomes indicate proximal (p) and distal (d) inversion breakpoints. The remaining symbols are as in Figure 1.

cytologically precisely as it would be expected if they were from O₇. Relative to the nearest of the available 140 cytologically mapped markers of the O chromosome (see Karageorgiou et al., 2019), the proximal breakpoint was located 44.5 kb downstream from *Sb* (Dmel\CG4316) and 117.4-kb upstream from microsatellite *dsub02*, and the distal breakpoint 111.8 kb downstream from *rdx* (Dmel\CG12537) and 29.3kb upstream from *Abi* (Dmel\CG9749). *Sb* and *dsub02* have been respectively mapped to subsections 77B (Dolgova, 2013) and 77C (Santos et al., 2010), and *rdx* and *Abi* to subsection 85E (Dolgova, 2013; Pegueroles et al., 2013) of the Kunze-Mühl and Müller (1958) standard cytological map. Other than O₇, no *D. subobscura* inversion maps to those positions.

Comparative analysis of the genes annotated in the regions immediately flanking the breakpoints in Ds₇, Ds_{ch-cu} and



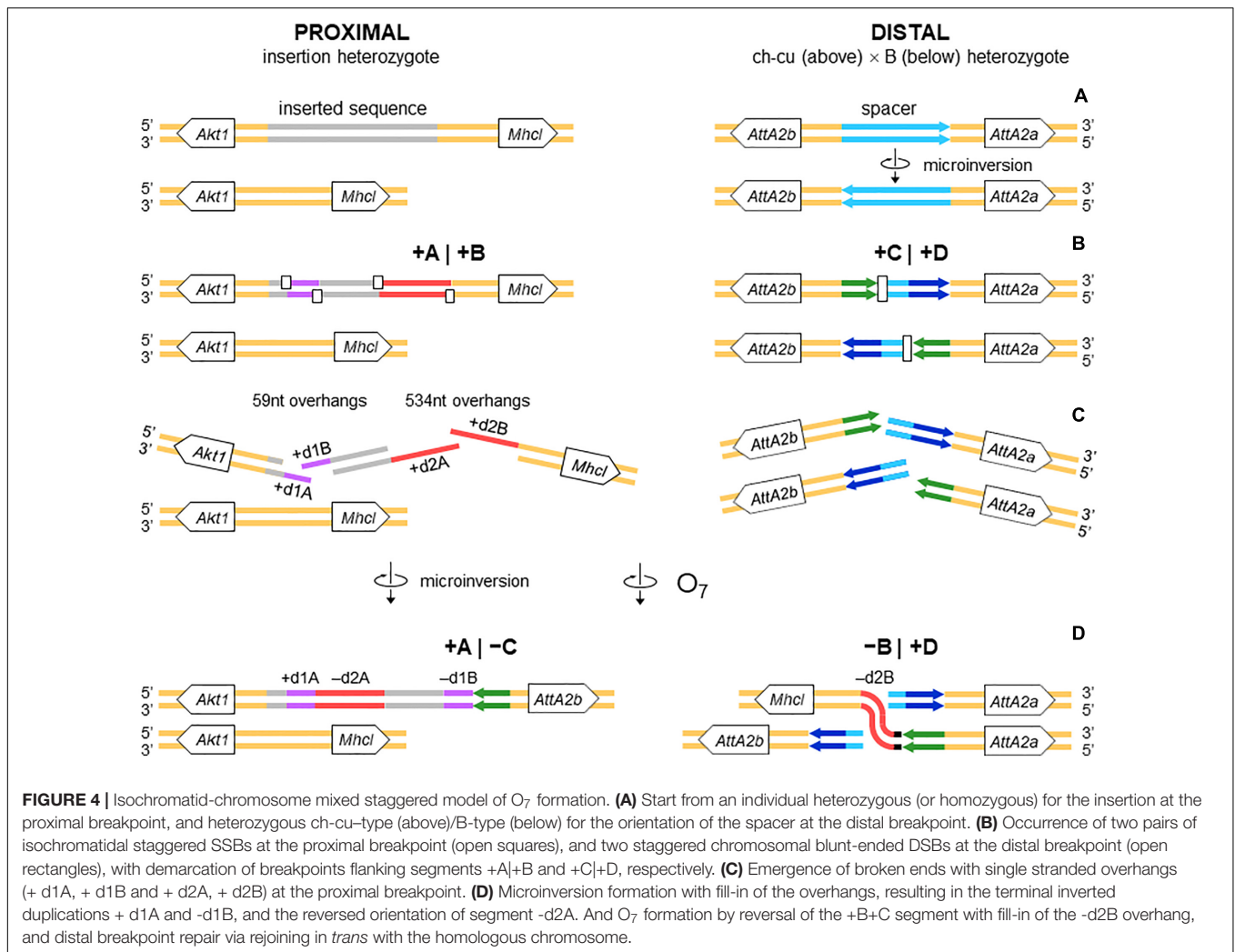
Ds_B with those in the outgroup Dg (**Figure 3A**) corroborated that Ds_{ch-cu} and Ds_B carried the uninverted state, whereas Ds₇ carried the inverted state. The assembled O₇ has a size of 9,936,431 bp, totaling 32.4% of the chromosome (30,629,152 bp). It has a GC content (43.8%) below that of the O chromosome (44.9%) since it is located in the chromosome centromere-proximal half, which is relatively AT-rich (Karageorgiou et al., 2019). O₇ was predicted to have 1,028 protein-coding genes, or 31.9% of the gene models of the O chromosome, in close agreement with its percent of chromosome length.

Nature and Properties of the DNA Sequences Surrounding O₇ Breakpoint Junctions

Proximal Breakpoint of O₇

The alignments used for isolation of the breakpoint junctions and their corresponding flanking regions A, B, C, and D are shown

in **Supplementary Figures 1, 2**. **Figure 3B** provides a schematic representation of the +A|+B region based on the alignment of **Supplementary Figure 3**. In the case of O₇, the region was reconstructed using the reverse complement of segment -B. The breakpoint junction is located within a 2,445-bp-long sequence stretch present only in the inverted state. The site of the insertion is flanked by multiple indels, which suggests that the insertion occurred in a region of prior sequence instability. Of the insertion length, 2,317 bp are on the +A segment and 128 bp on the +B segment. The insertion begins with a 153-bp-long direct repetition (R1-2) of the upstream flank. Proceeding downstream from this repeat, there are two inverted duplications named d1 and d2, each with copies A and B, with d1 shorter (59 bp long each of d1A and d1B) than d2 (534 and 540 bp for copies d2A and d2B, respectively). The two A copies (i.e., d1A and d2A) are separated from the two B copies (i.e., d1B and d2B) by an intervening sequence of 1,374 bp. The junction between +A and +B is precisely located between d1B and d2B. d2B extends

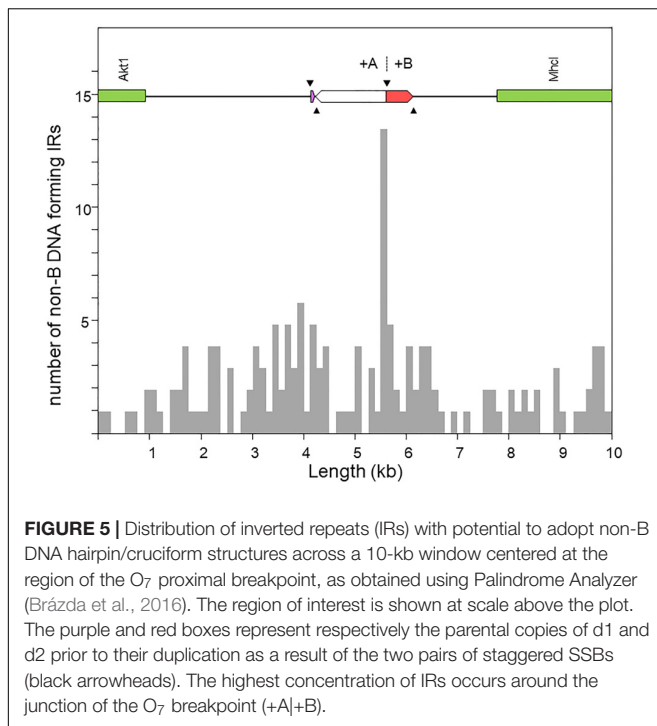


409 bp downward from the downstream end of the insertion into the region of resumed homology between O₇ and the uninverted state, indicating the orientation of the parent copy.

The above pattern of sequence copy number, order, and orientation most parsimoniously indicates that the proximal breakpoint of O₇ was formed on an insertion region that experienced two pairs of staggered SSBs, which resulted in two DSBs (Figure 4; but see section “DISCUSSION” for alternative models). The upstream-most DSB generated the proximal breakpoint of a 2,026-bp-long microinversion and the downstream-most DSB generated a junction flanked upstream by the distal end of the microinversion and downstream by the proximal end of O₇. Accordingly, duplications d1 and d2 would, respectively, represent the filled-in staggered SSB-induced terminal single-stranded overhangs of the microinversion and inversion O₇. Figure 3C shows O₇'s segments A and B such as they are found in the inversion. That d2A and d2B show direct instead of reverse relative orientation as it would be expected if paired-staggered SSBs generate inversions with inverted repeated ends (Ranz et al., 2007) would be explained by the reversal in the orientation of d2A as a result of the microinversion.

Relative to the predicted nearest gene TSSs, the events took place in an intergenic region. Specifically, the upstream-most SSB occurred 1,364 bp downstream from *Akt1* (CG4006; *serine/threonine-protein kinase B*) gene, and the downstream-most one 2,047 nt upstream from *Mhcl* (CG31045; *myosin heavy chain-like*) gene (Figure 3A). From our repeat annotation pipeline, the region around the breakages is a composite of repetitive sequences [16 in total, ranging in length from 21 bp of a (TTG)_n simple-repeat to 532 bp of satellite rnd-4_family-179], interspersed with traces of transposable elements [84 bp from an LTR and 72 bp from a long interspersed nuclear element (LINE)]. Overall, no evidence of open reading frames and/or specific motifs could be found pointing to the observed breakages as directly caused by the insertion/excision of other sequences.

The role of non-B DNA as source of DSBs is well-established. Generally, DSBs are expected to colocalize with their causal non-B DNA motifs (e.g., Kolb et al., 2009; Lu et al., 2015). We used this prediction to investigate whether the local DNA conformational environment of the ancestral sequence could have acted as trigger or mediator of the complex rearrangement of the proximal breakpoint region. We proceeded in two steps:



first, we reconstructed the region of the rearrangement before the breakages. It should be recalled that most of the rearranged sequence is embedded in an insertion that is absent in the ancestral non-rearranged state. Therefore, we reconstructed the prebreakages state by undoing the hypothetical rearrangement steps that generated the present sequence state. Specifically, we reversed the orientation of the microinversion (**Supplementary Figure 4**) and deleted one copy of each DSB-induced duplication (**Supplementary Figure 5**). The resulting sequence had the form: + d1, (+ 1,374 bp), |, + d2 (**Figure 4B**). Which copy of each of the two duplications to eliminate was inconsequential, because they are nearly identical to each other in the two cases (98.3% and 95.6%, for the identities between copies A and B of dup1 and dup2, respectively). Furthermore, the observed high level of identity (97.3%) between d2 and its homologous region in Ds_{ch-cu} and Ds_B suggested that the rearrangement is recent enough to allow assuming that the original conformational sequence features that could have mediated it are still observable. After establishing the prebreakage sequence, we next looked for sequences with the potential to form non-B DNA structures along a 10-kb window centered on it.

Figure 5 shows the distribution of the number of IRs capable of forming hairpin and cruciform structures along the target sequence. The highest density occurs immediately around the junction between the microinversion and inversion O₇. In particular, the breakpoint is located within a ~150-nt-long stretch of AT-rich sequence [simple repeat (ATTT)_n, from our genome annotation pipeline] containing 15 IRs, of which one located 68 nt downstream the breakpoint junction ranked in the top 5% with highest likelihood of intrastrand annealing to form a hairpin (AATTTT AAAATT; $\Delta G_S - \Delta G_L = 2.64$). In addition, embedded in the IR cluster, there is one tandem repeat

of 8.7 copies of the consensus heptanucleotide AATAAAT, and one mirror repeat of two 11 nt-long repeats separated by a 30-nt spacer, indicating that the proximal breakpoint of O₇ occurred on an unstable sequence with potential for adopting multiple alternative non-B DNA conformations.

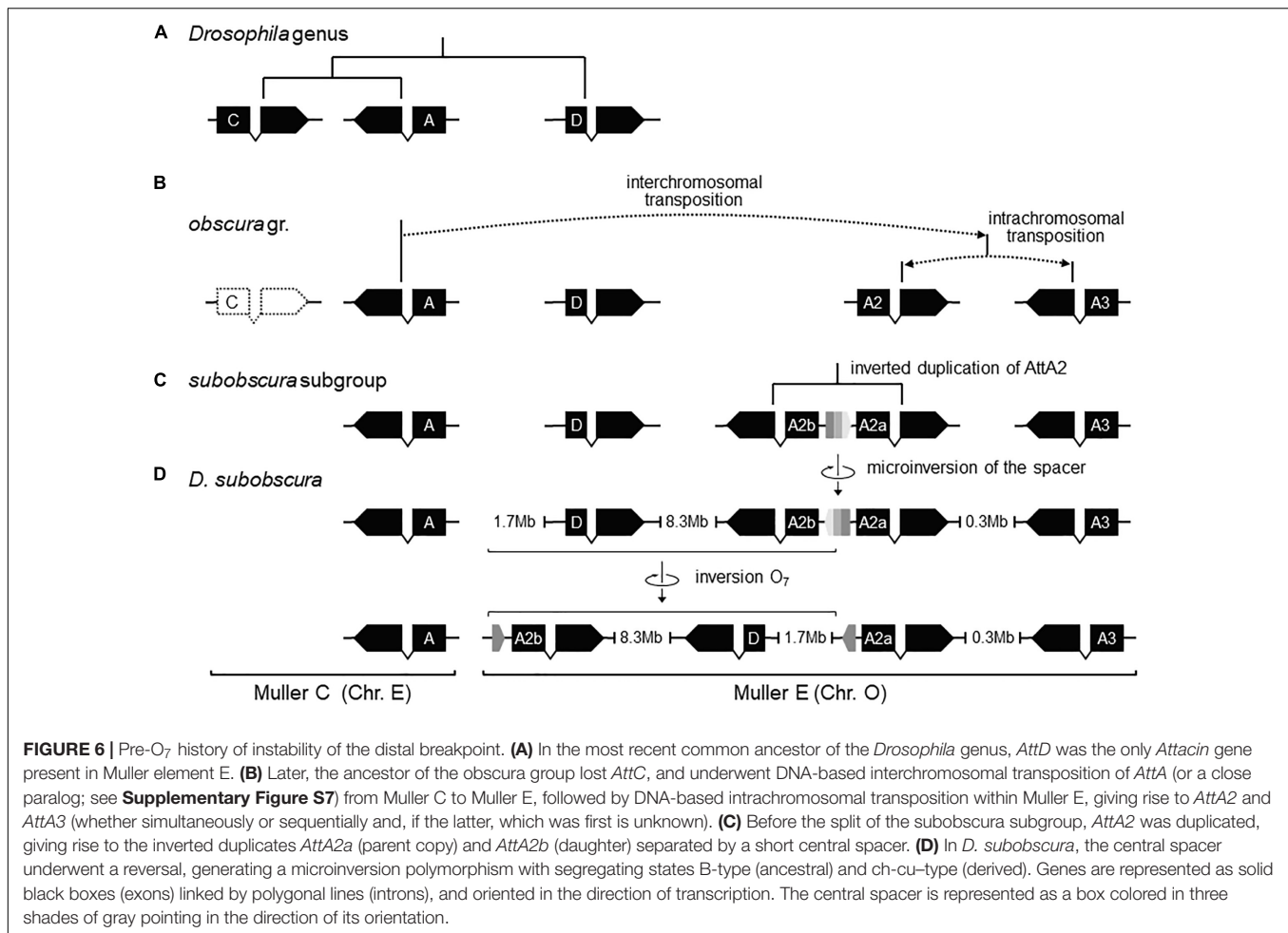
Distal Breakpoint of Inversion O₇

Figure 3B provides a schematic representation of the +C|+D region based on the alignment of **Supplementary Figure 6**. In the case of O₇, the region was reconstructed using the reverse complement of segment -C. From up to downstream, the breakpoint junction is located within a 450-aligned-sites-long gap-rich spacer region, spanning between two highly identical long IRs, IR1 and IR2, of 1,117 and 1,112 sites of alignment length, respectively. There is no evidence of duplicated sequence in Ds₇ relative to the other assemblies, indicating that the DSB either was a clean cut or did not involve significantly staggered SSBs. On the other hand, the spacer of Ds₇ was the shortest (250 nt) of all four lines (407, 317, and 343 nt for Ds_{ch-cu}, Ds_B, and Dg, respectively) because of a single deletion located precisely at the center of the region (hereon called dd7, for distal deletion of O₇). A closer look at the pattern of pairwise sequence similarities along the spacer revealed two findings: (i) dd7 split the Ds₇ spacer in two mirror halves. For the upstream half, Ds₇ is almost identical (96.8%) to Ds_{ch-cu} while bearing no detectable homology to Ds_B, whereas for the downstream half, Ds₇ is almost identical (97.6%) to Ds_B while bearing no detectable homology to Ds_{ch-cu}; and (ii) the spacer of Ds_{ch-cu} is almost identical (95.4%; excluding indels) to that of Ds_B but in reversed orientation. The reversal occurred in Ds_{ch-cu}, because in Ds_B the spacer is oriented as in the outgroup Dg.

Altogether, the above observations can be understood as follows (**Figure 4**). Prior to the origination of the distal breakpoint of O₇, a carrier of an uninverted chromosome of B-type experienced a reversal of the spacer region between the IRs, giving rise to the uninverted chromosome of ch-cu-type. Later on, a homokaryotype for the uninverted chromosome that was heterozygous for the microinversion of the spacer underwent at least two DSBs, one in each of two homologous non-sister chromatids, such that the DSB in the ch-cu-type chromatid occurred immediately before the first site of the dd7 and that in the B-type chromatid immediately after the last site of the dd7. Finally, the reversed + B end generated by the proximal staggered DSB in the ch-cu-type chromatid illegitimately joined with the + D end generated by the distal DSB in its homologous non-sister B-type chromatid, which resulted in a recombinant chromosome carrying the inversion O₇ with the exact observed dd7 deletion.

Like the proximal breakpoint, the distal breakpoint occurred in an intergenic region yet at comparatively much shorter distance (~390 bp) to the nearest genes. Specifically, the breakage separated two copies of an *Attacin* gene (CG10146; *AttA*) located opposite to each other on each of the two arms of the long IR. Our repeat annotation pipeline did not identify repetitive sequences in the vicinity of the distal breakpoint in Ds_{ch-cu} or Ds_B.

We searched the region of the spacer for potential non-B DNA-forming sequences in the vicinity of the breakpoint



junctions in Ds_{ch-cu} and Ds_B. In both cases, we found that the IR with the highest propensity to form a hairpin was a perfect 14-bp-long palindromic sequence located next to the breakpoint junctions (ATGAACT AGTTCAT; $\Delta G_S - \Delta G_L = 2.05$; located 13 and 2 bp upstream and downstream the junction in Ds_{ch-cu} and Ds_B, respectively). Apart from IRs, we did not detect additional potential non-B DNA sequences around the distal breakpoint.

All nucleotides in the +A|C region of Ds₇ could be unambiguously ascribed to segment A or C. However, in the -B|+D region -B and +D are separated by 21 extra inserted nucleotides (i.e., GAGCACTCTCCACAGCAAAGT). We decided to ascribe this sequence to the distal breakpoint junction, because it contains an 8-bp substring (underlined) that resembles the beginning of the +D end (CATCAAAG), and hence it likely represents filler DNA generated by a microhomology-templated repair mechanism.

Pre-inversion Record of Rearrangement of O₇ Breakpoints

Previously, it was shown that the proximal breakage of O₇ was preceded by an insertion. Likewise, the region of the distal breakage had a pre-inversion history of rearrangement, which

run closely associated with a highly dynamic evolution of the *Attacin* immunity gene family in the *obscura* species group. This conclusion is based on phylogenetic analysis of the *Attacin* family in *Drosophila* (**Supplementary Figure 7**) using synteny to distinguish orthologous from paralogous copies (**Supplementary Table 1**). The results are summarized in **Figures 6A–D**. The most recent common ancestor of the *Drosophila* genus (**Figure 6A**) carried three copies of the gene with relationships [(A,C),D], of which the more distant D was located in Muller element E, and the closer to each other A and C in Muller element C. After it split from the *melanogaster* group (**Figure 6B**), the branch leading to the *obscura* group lost copy C and underwent an interchromosomal transposition of copy A from Muller element C to E. The daughter copy then underwent another, in this case intrachromosomal, transposition, which originated two new *Attacin* copies that we called *AttA2* and *AttA3*, with *AttA2* located between *foxo* and *Npc2b*, and *AttA3* located ~300 kb downstream from *AttA2*, between *Cul5* and *Sirt7*. The two transpositions were genome-based duplications rather than retroposition events, because the new copies conserved the intron position of their parental gene. Before the split of the *subobscura* subgroup (**Figure 6C**), copy *AttA2* underwent an inverted duplication that generated the two closely

spaced copies *AttA2b* and *AttA2a* in head-to-head orientation, and transcribed in opposite directions. In *D. subobscura* (Figure 6D), the spacer between the IRs experienced a reversal of orientation generating the microinversion polymorphism of the distal breakpoint. Subsequently, a heterozygote for the microinversion underwent distal DSBs that allowed the formation of the recombinant O₇ inversion via ectopic repair of non-sister chromatids.

Potentially Functional Effects of the O₇ Mutation

The distal break of O₇ disrupted concerted evolution between two *subobscura* subgroup-specific *AttA2* duplicates. This conclusion is based on the previous section's results, together with the phylonetwork of coding sequences shown in Figure 7. Accordingly, right after the duplication of *AttA2*, the two paralogs began to evolve in concert, converting each other to generate their present characteristic phylogenetic pattern of greater resemblance between paralogs from the same species (i.e., *D. guanche* and *D. subobscura*) than between orthologs from different species (e.g., Puig-Giribets et al., 2019). At one end of the resemblance, it is the ch-cu strain, whose two *AttA2* copies are identical to each other, and at the other end O₇, where the copy relocated by the inversion evolved significantly faster than the one that remained in place, owing exclusively to an acceleration of the synonymous substitution rate [$P < 0.05$; Tajima's relative rate test (Tajima, 1993) using either of the remaining six sequences as outgroup], as the two copies are identical at the amino acid level. The acceleration took place in the direction of a slight decrease in codon bias in the relocated copy ($N_c = 51.2$ vs. 50.7, for the comparison *AttA2b* vs. *AttA2a*, respectively; where N_c is the improved effective number of codons index; Sun et al., 2013). The increased synonymous rate can be understood, in part because the inversion released the two *Attacin* copies from evolving in concert; and in part assuming that the expression of the paralogs shifted as a result of changes in regulatory environment associated with their relocation.

Considering the short spacing between the two *AttA2* paralogs in the uninverted chromosome (~390 bp), it appeared likely that the inversion would have detached them from part of their promoters, binding them to new potentially *cis*-acting elements. To assess this possibility, we searched 1 kb upstream of the predicted TSS of each gene for putative transcription factor binding sites (TFBSs) for five transcription factors (TFs), including the nuclear factor κ B factors dorsal (dl), dorsal-immunity related factor Dif and Relish (Rel), the GATA factor Serpent (*srp*), and the forkhead factor dFOXO. The first four TFs are under control of the Toll and immune deficiency (IMD) immunity pathways and regulate *Attacin* inducible expression in response to bacterial infection (Senger et al., 2004). dFOXO TF is controlled by the insulin/insulin-like growth factor signaling (IIS) metabolic pathway and regulates constitutive *Attacin* expression in non-infected flies suffering from energy shortage or stress (Becker et al., 2010). The results are shown in Figure 8. The *AttA2* genes had predicted TFBSs for the immunity related factors in both uninverted and inverted chromosome states, but only

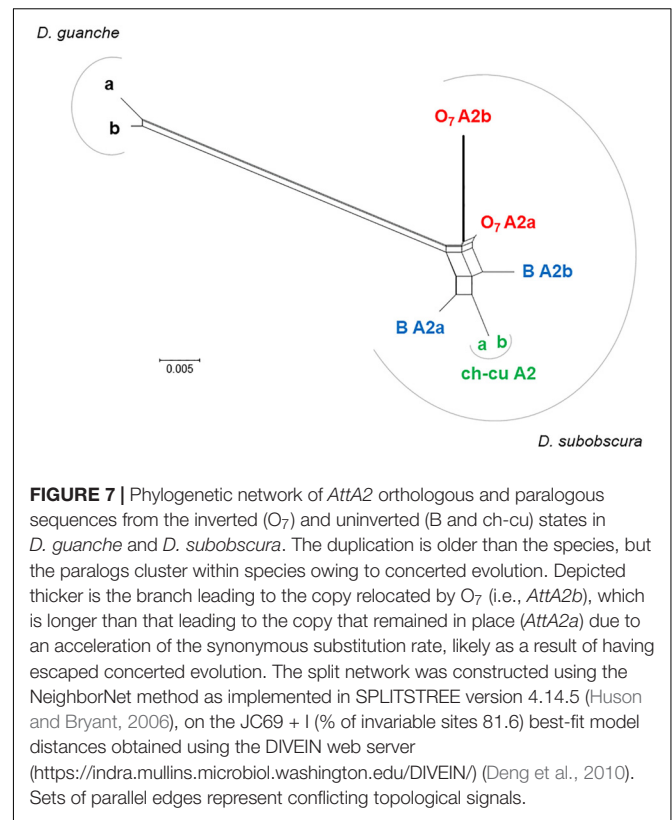


FIGURE 7 | Phylogenetic network of *AttA2* orthologous and paralogous sequences from the inverted (O₇) and uninverted (B and ch-cu) states in *D. guanche* and *D. subobscura*. The duplication is older than the species, but the paralogs cluster within species owing to concerted evolution. Depicted thicker is the branch leading to the copy relocated by O₇ (i.e., *AttA2b*), which is longer than that leading to the copy that remained in place (*AttA2a*) due to an acceleration of the synonymous substitution rate, likely as a result of having escaped concerted evolution. The split network was constructed using the NeighborNet method as implemented in SPLITSTREE version 4.14.5 (Huson and Bryant, 2006), on the JC69 + I (% of invariable sites 81.6) best-fit model distances obtained using the DIVEIN web server (<https://indra.mullins.microbiol.washington.edu/DIVEIN/>) (Deng et al., 2010). Sets of parallel edges represent conflicting topological signals.

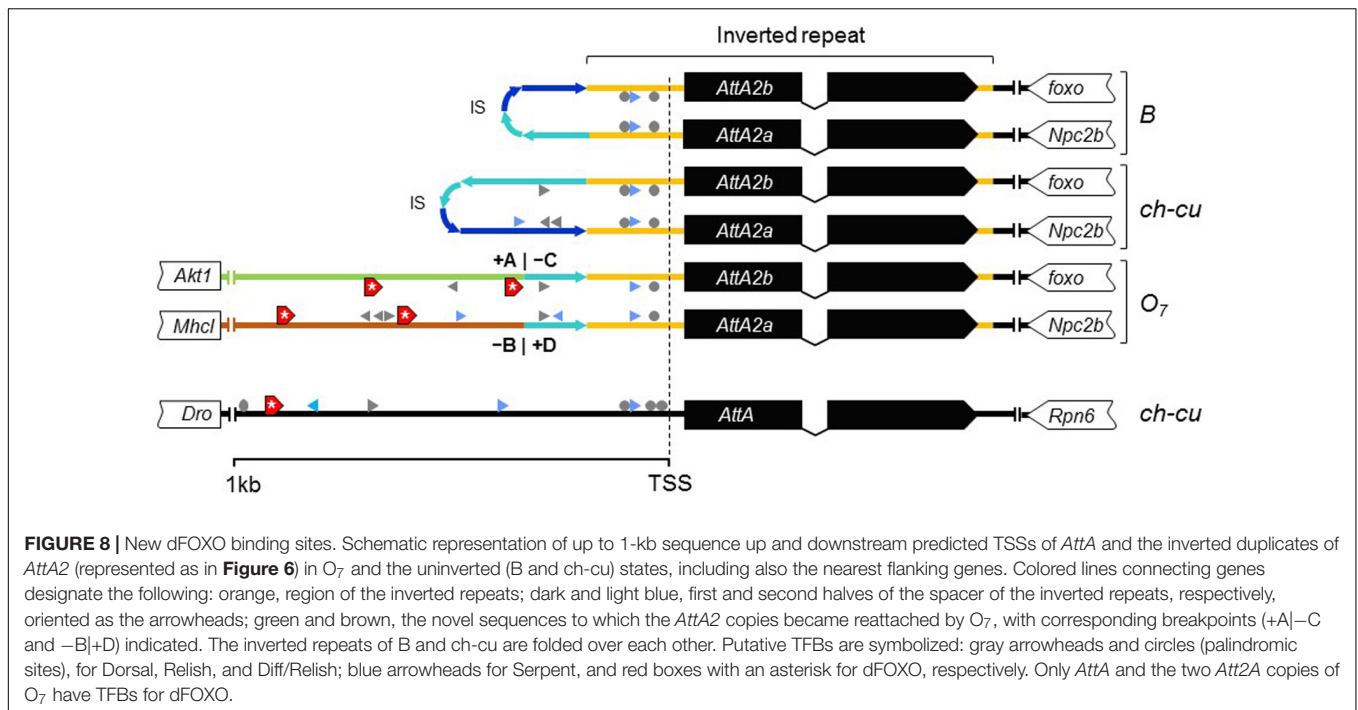
the *AttA2* genes of the inverted chromosome had TFBSs for the metabolic factor dFOXO. Furthermore, the dFOXO TFBSs were all contributed by the newly attached sequence. The fact that the *AttA2* genes were conserved at the amino acid level in *D. subobscura*, together with the observed qualitative difference in predicted *cis*-acting sequence between uninverted and inverted chromosomes, suggests that the inversion O₇ brought the *AttA2* genes under the influence of the IIS metabolic pathway.

In addition to the *Attacin* immunity genes, the breakpoint regions include *Akt1* and *foxo*, two interacting core components of the IIS metabolic pathway identified by other studies as candidate for climate adaptation (Fabian et al., 2012; Paaby et al., 2014; Kapun et al., 2016; Durmaz et al., 2019). The roles of these genes and the potential impact of O₇ on them are dealt with in the *Discussion*.

DISCUSSION

Molecular Mechanism of O₇ Formation O₇ Is a Complex Multibreak Inversion Formed via Rejoining in *trans* With the Two Homologous Chromosomes

Sequence data on inversion formation in *Drosophila* have been interpreted in terms of two major mechanisms with associated distinctive footprints. The first mechanism is intrachromatid NAHR between inversely oriented repeats. This mechanism generates inversions with duplications at their ends in both the



inverted and uninverted states (Cáceres et al., 1999), which is not the case of *O*₇.

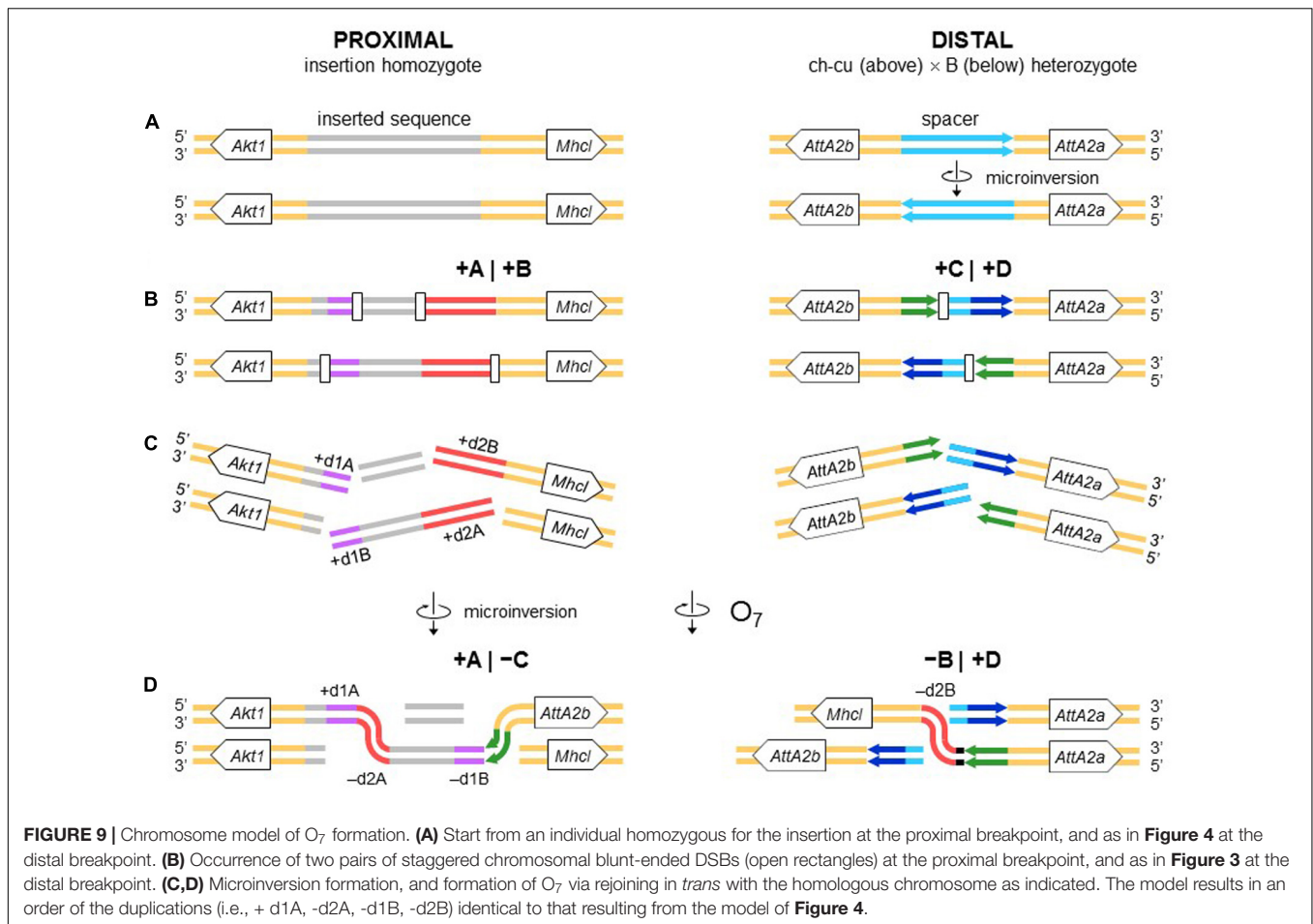
The second mechanism is chromosomal breakage and ectopic repair via NHEJ. This mechanism either does not generate duplications or generates them but at the ends of the inverted state only. These two types of NHEJ footprints have been explained in terms of two alternative modes of breakage: cut-and-paste via clean DSBs that generate blunt ends and staggered on the same (isochromatidal) or different (chromatidal) sister chromatids (see *Introduction*). In the case of *O*₇, it is not a cut-and-paste inversion, but neither is it a typical staggered breaks inversion. Thus, while the inversion proximal breakpoint could be either isochromatidal (**Figure 4**) or chromatidal (**Figure 9**), the distal breakpoint has to involve the two homologous chromosomes (**Figures 4, 9**). This latter pattern could be deduced because of the chanceful circumstance that our two representatives of the uninverted state (i.e., *Ds_ch-cu* and *Ds_B*) segregated for the microinversion of the spacer between the IRs flanking the distal breakpoint. Alternatively, the distal breakage could have occurred in a recombinant between chromosome types ch-cu and B. This, however, appears unlikely because crossover within microinversions should be extremely rare (Greig, 2007). Our conclusion agrees with a study of the genealogical relationships between inversions of the E chromosome in *D. subobscura*, which proposed that *E*₉ arose in a heterokaryotype *E*_{ST}/*E*₁₊₂ to accommodate a conflict between molecular and cytological data (Orengo et al., 2019). This and our results indicate that NHEJ inversions form through mechanisms that can incorporate information from the two homologous chromosomes (chromosome model), in

addition to the previously proposed intrasister and intersister chromatidal exchanges.

The Breaks of the *O*₇ Inversion Were Likely Induced by Non-B DNA Secondary Structures

Inversion *O*₇ provides, to our knowledge, the first compelling evidence for a role of non-B DNA in inversion formation in *Drosophila*. Previous studies had reported the presence of AT-rich sequences around the breakpoints of some fixed (Cirera et al., 1995; Richards et al., 2005) and polymorphic (Prazeres da Costa et al., 2009) inversions. In no instance, however, were particular sequences susceptible to adopt secondary structures identified. In the case of *O*₇, the proximal break junction occurred just within a palindromic AT-rich repeat capable of adopting hairpin/cruciform, slipped and triplex DNA conformations. Likewise, the distal junctions are located next to perfect 14-bp-long hairpin/cruciform-forming palindromes.

The role of non-B DNA-forming sequences in causing genome instability is well-established (Wang and Vasquez, 2006; Lobachev et al., 2007; Aguilera and Gómez-González, 2008; Zhao et al., 2010). The shift from B to non-B DNA conformation occurs while DNA is in single-stranded form, e.g., behind replication forks, between Okazaki fragments, or in actively transcribed genes (Voineagu et al., 2008). Non B-DNA structures induce DSBs through, e.g., stalling replication and transcription (Mani and Chinnaiyan, 2010; Kaushal and Freudenreich, 2019). There are no specific predictions as to the type, number, and location of the DSBs generated by any given structure in any particular situation. Still, a single structure can induce multiple DSBs across hundreds of base pairs around it (Wang et al., 2006; McKinney et al., 2020), and stalled replication forks can accumulate up to 3 kb of single-stranded DNA (Sogo et al., 2002; Lopes et al.,

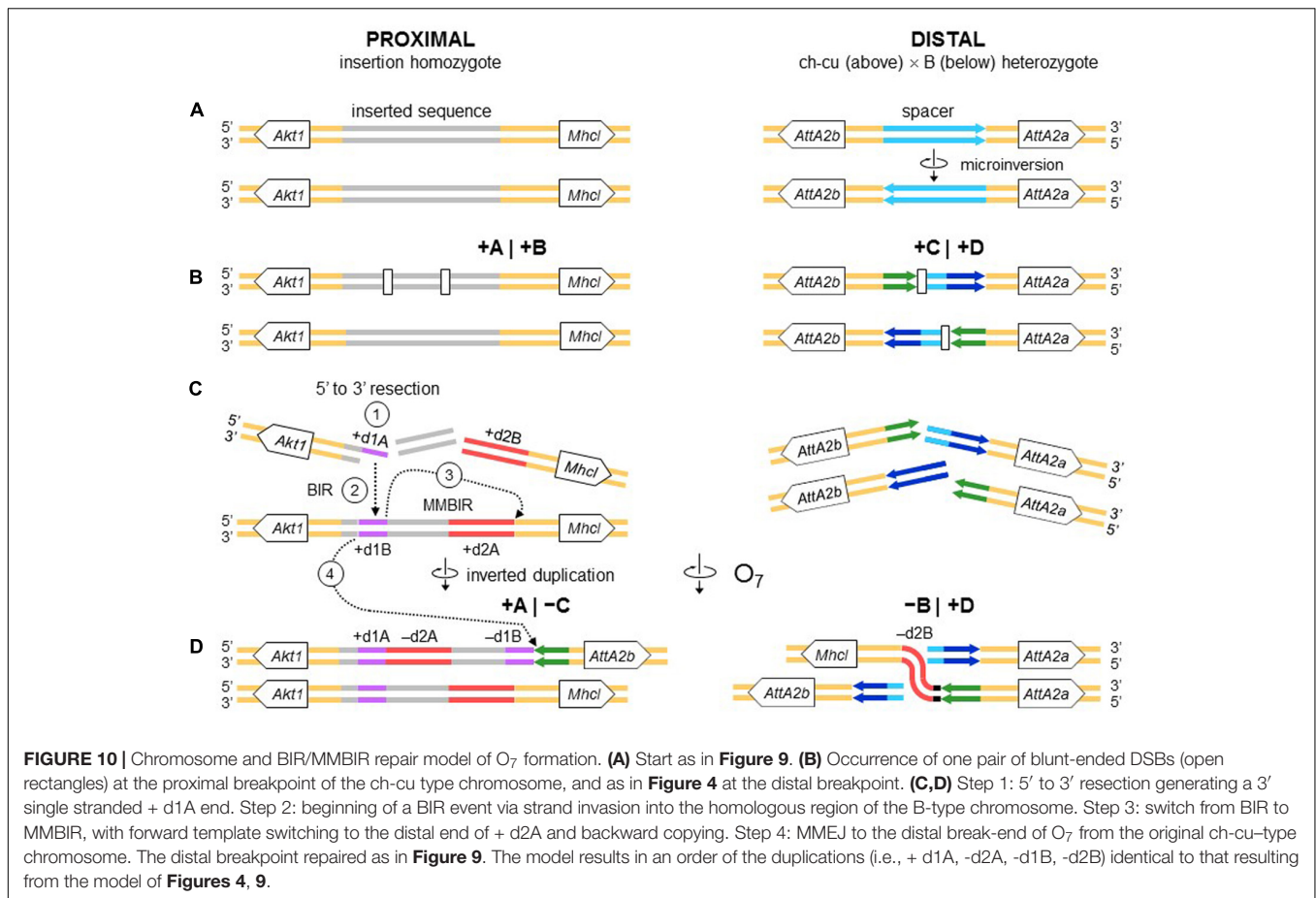


2006). In the case of O₇, this length is well over the size of the overhangs that would be generated by an isochromatid model of the proximal breakpoint (58 and 534 nt; see **Figure 3** and **Supplementary Figure 3**).

The Inverted Duplications at the O₇ Breakpoints Could Be Footprints of Repair Instead of Staggered Breakage

All the aforementioned inverted duplication-generating NHEJ models are predicated upon the role of DNA breakage (Ranz et al., 2007). However, the inverted duplications at the ends of O₇ could also be explained as a result exclusively of repair, with no need for invoking staggering of the breaks. DNA repair has emerged as a key factor capable of generating extremely complex breakpoint sequence rearrangements (reviewed in Scully et al., 2019). The spectrum of known error-prone repair mechanisms can be grossly classified as recombination-based, such as microhomology-mediated end-joining (MMEJ), and replication-based, such as break-induced replication (BIR) and microhomology-mediated BIR (MMBIR) (Lee et al., 2007; Zhang et al., 2009; Hastings et al., 2009). Here, the term *microhomology* is used to mean a short tract (~1 – 25 bp) of chance similarity, rather than common descent. In the case of O₇, three features suggest that what appear to be footprints of breakage by the

staggering models could in fact be footprints of a replication-based mode of repair (reviewed in Kramara et al., 2018; Scully et al., 2019), including (i) presence of non-B DNA-forming sequences just in, or adjacent to breakpoint junctions (see below); (ii) spatial proximity of the breakpoint regions in the nucleus, as evinced by the fact that the genes flanking the junctions are closely related functionally (Farré et al., 2015; but see Sunder and Wilson, 2019); and (iii) multiple breaks concentrated in a short sequence segment. A fourth feature, namely, presence of microhomology at the distal breakpoint junction, would be also consistent with a recombination-based mechanism such as MMEJ. Overall, these features suggest that O₇ arose as result of a non-B DNA-induced replication impairment, affecting at least its proximal breakpoint. It is known that this type of events can trigger BIR and MMBIR repair (Sakofsky et al., 2015). Of the two pathways, the second pathway has yet to be identified in *Drosophila* (Alexander et al., 2016; Bhandari et al., 2019). A possible scenario is detailed in **Figure 10**: first, non-B DNA-induced stalling of a replication fork at the proximal breakpoint of a ch-cu-type chromosome led to two DSBs generating three fragments. Second, the centromere-proximal fragment engaged in a BIR event using the homologous region of a B-type chromosome. Third, a second fork stalling triggered a switch from BIR to MMBIR with template switching to



a downstream microhomology. Copying backward from the new template resulted in the rearrangement of the proximal breakpoint, including the inverted duplication of the O₇ end (e.g., Lee et al., 2007; Smith et al., 2007; Carvalho et al., 2015; Tremblay-Belzile et al., 2015). Finally, the event was terminated by an MMEJ to the distal break-end of O₇ from the original ch-cu-type chromosome (e.g., Scully et al., 2019).

The O₇ Breakpoints Carry a Pre-inversion Record of Fragility

The breakpoint sequences of O₇ had a record of instability prior to the origin of the inversion, as evinced by the fact that they are located within sequences inserted from elsewhere in the genome. This suggests that the regions that gained those insertions were relatively exposed in the nucleus (reviewed in Farré et al., 2015). In the case of the proximal breakpoint, that could be associated with high levels of transcriptional activity at the broadly expressed *Akt1* gene (Andjelković et al., 1995; Slade and Staveley, 2016).

That the O₇ junctions arose in fragile regions, beyond the proximate effects of their associated non-B DNA (see above), may be most apparent from the pre-inversion record of recurrent rearrangement of the IR at the distal breakpoint (**Figure 6**). This record is particularly amenable to reconstruction because the IR largely consists of two copies of the *Attacin*

A gene that are highly conserved. It includes at least three rearrangements that occurred in the lineage of *D. subobscura* after its separation from that of the *melanogaster* group (see section “RESULTS”; **Figure 6**), namely, (i) insertion of *AttA2* between the *foxo* and *Npc2b* genes; (ii) emergence of the IR by inverted duplication of the parental *AttA2* (**Figure 6B**), which could have occurred through an event of forward template switching and backward copying by the DNA polymerase (Smith et al., 2007; Lee et al., 2007), as discussed above; and (iii) emergence of the ch-cu-type chromosome via inversion of the spacer between the IRs in a B-type chromosome (**Figure 6D**), which could be explained as an outcome of a stem-loop formation by the IR, followed by resolution of the strand-exchange junctions between the IR arms (see **Figure 4** in Leigh Brown and Ish-Horowicz, 1981; **Figure 3** in Kolb et al., 2009 and Zhao et al., 2010).

The pre-O₇ insertion in the proximal breakpoint is specific to *D. subobscura* and is therefore much more recent than that of *AttA2* in the distal breakpoint. Preliminary analyses indicate that it is internally rearranged relative to other paralogous copies, supporting that it carries recombinogenic potential. The origin and evolution of this inserted sequence, as well as its possible implication in the formation of other *D. subobscura* inversions, warrant further investigation (CK, RT, and FR-T; manuscript in preparation).

O₇ Breakpoints Potentially Functional Effects

O₇ Relocated *foxo* in Tight Linkage Association With Its Antagonistic Regulatory Partner of the IIS Metabolic Pathway *Akt1*

O₇ changed *foxo* from being megabases (~10 Mb) away from *Akt1* to being tightly linked to it, with only the short *AttA2b* gene sandwiched between them. *Akt1* and *foxo* are functionally conserved genes, which, in *Drosophila*, encode the serine/threonine-protein kinase B AKT/PKB, and the forkhead-box DNA-binding domain-containing TF dFOXO, respectively. The two genes are key antagonistic regulators of the IIS pathway (Teleman, 2010; Slade and Staveley, 2016), a major trigger of shifts in anabolic versus catabolic cellular activity in response to nutritional status (de Jong and Bochdanovits, 2003) and multiple other cues (Regan et al., 2020). In abundant nutrient conditions, AKT/PKB inactivates dFOXO, thus shifting food energy allocation toward reproduction and growth (the IIS pathway). Conversely, scarce nutrient conditions prevent AKT/PKB from inactivating dFOXO, which redirects metabolism toward mobilization of energy stores for somatic maintenance (FOXO pathway). Laboratory research using large effect mutants has shown that the IIS/FOXO pathway is extensively pleiotropic, with major evolutionary conserved effects on fitness-related life-history traits, including growth, size, reproduction, lifespan, and stress resistance (reviewed in Flatt and Partridge, 2018). Research from the field found IIS loci to harbor substantial genetic variation, which frequently exhibits spatiotemporal patterns that look as if they were shaped by selection on the associated IIS traits (Fabian et al., 2012; Paaby et al., 2014; Kapun et al., 2016). In a recent laboratory assay, two *foxo* alleles showing opposite latitudinal clines in *D. melanogaster* were compared on an otherwise homogeneous genetic background. The alleles showed contrasting effects on viability, size-related traits, starvation resistance, and fat content, whose directions were overall consistent with predictions from the clinal variation of the characters (Durmaz et al., 2019).

The O₇ mutation could have altered *Akt1* and/or *foxo* function via multiple non-mutually exclusive mechanisms, such as mutual regulatory interference, considering that they are antagonistic effectors; relocation to the sides of an immunity gene (i.e., *AttA2b*) expected to be under intense purifying selection on expression (see below); and alteration of the genes' functional neighborhood at higher-order levels of chromatin organization (Farré et al., 2015; McBroome et al., 2020). It could be argued that the nuclear environment of the genes remained basically unaltered, if the reason why they became involved in the rearrangement was that they already were in close spatial proximity to each other in the nucleus. This, however, did not necessarily have to be the case, considering recent findings in yeast that rejoining of DNA break ends is not determined by the predamage spatial proximity of the DSBs (Sunder and Wilson, 2019). Be that as it may, bearing in mind that the seasonal increase of O₇ occurs from early spring to

midsummer, coinciding with the growth season, it seems more likely that whatever the effect of the inversion mutation on *Akt1* and/or *foxo*, it occurred in the direction of an enhanced basal IIS versus dFOXO activity relative to the O_{ST} ancestral state. This would raise the question of why the O₇ frequencies decrease (and those of O_{ST} increase) every year from late summer to winter.

O₇ Disrupted the Concerted Evolution of Two *AttA2* Immunity Genes and Reattached Them to Putative dFOXO Metabolic Enhancers

The immune function is highly energy demanding in terms of both maintenance and, especially, rapid deployment upon infection (reviewed in Dolezal et al., 2019). Therefore, within a limited energy budget, a trade-off is expected between reproduction and immunity (Schwenke et al., 2016). The *Drosophila* innate immune response consists of a cellular and a humoral component. The humoral component involves the production of antimicrobial peptides, among which Attacins are active against gram-negative bacteria (Hanson and Lemaitre, 2020). The two main modes of Attacin production, including the induced (by a factor of even > 100) upon infection mode, and the basal in absence-of-infection mode link immunity with the *Akt1/foxo* IIS metabolic signaling pathway (Becker et al., 2010; Dolezal et al., 2019). The inducible mode is regulated primarily by the immunodeficiency *Imd* signaling pathway and to a lesser extent by the *Toll* signaling pathway. The two signaling pathways have the same effect of activating dFOXO, thus mobilizing resources toward the production of Attacins (Dionne et al., 2006; Dolezal et al., 2019). The basal mode is regulated directly by dFOXO activity when induced by starvation (Becker et al., 2010; Buchon et al., 2014). Immunity genes, including *Attacins*, are among the known most rapidly evolving genes and have frequently shown evidence of local adaptation in *Drosophila* (Lazzaro and Clark, 2001, 2003).

There would be a number of mechanisms by which the O₇ mutation could have reduced *Attacin* genes' expression. For example, the breakage of the invertedly transcribed *AttA2* tandem duplicates could have impaired the inducibility of one or the two paralogs, or their separation could have made them lose gene expression coregulation, as might be surmised from the observations that they halted or slowed down evolving in concert, and that *AttA2b* shows decreased codon bias. These mechanisms could have acted synergistically with each other and with those already discussed in connection with *Akt1* and *foxo*. Although this scenario could be partially offset by the increase in basal *AttA2* transcript levels that may be expected from the duplicates having been reattached to dFOXO enhancers (Becker et al., 2010), all in all, the evidence suggests that (i) at its inception, O₇ caused a rearrangement with partial disruption of a set of functionally related loci with overlapping pleiotropic effects on immunometabolic traits. If, in addition to these direct effects, there concurred indirect effects of linkage between locally, and given the functional relationship, likely epistatically interacting alleles warrant further investigation; and (ii) the resulting haplotype imparted a shifted

pattern of resource allocation toward reproduction at a cost to immunity, compared to the O_{ST} ancestor. Such an opposing antagonistic pleiotropy would result in a seasonal frequency cycle qualitatively similar to that shown by the inversions, if reproduction is favored from early spring to midsummer, when O₇ rises (and O_{ST} wanes), and immunity from late summer to winter, when it wanes (and O_{ST} rises). There is ample evidence that the qualitative and quantitative composition of temperate bacterial communities cycles seasonally (Lazzaro et al., 2015; Shigyo et al., 2019). Recently, a study using *D. melanogaster* from the eastern United States (Behrman et al., 2018) found a seasonal shift in immunocompetence, with the trait value declining every spring to autumn. The shift was interpreted as resulting from relaxed selection for immune response during the warm season, much like what we propose here for the O₇/O_{ST} inversion polymorphism. Prior data on temporal genetic variation within and between O inversions point to additional loci that would be consistent with the seasonal cycle of O₇ being mediated by immunometabolic selection (Rodríguez-Trelles, 2003). The case of the *Mpi* gene encoding the key glycolytic enzyme mannose-6-phosphate isomerase (MPI) is noteworthy. From our assembly, *Mpi* is located 2.15 Mb outward from the distal breakpoint of O₇, which is within the estimated region of the inversion-associated strong recombination-suppression effect (3.5 Mb; Pegueroles et al., 2010b). The MPI fast/slow electrophoretic polymorphism was found to be only moderately associated with the O₇/O_{ST} polymorphism. Yet (i) the magnitude of the locus-by-inversion disequilibrium cycled seasonally, and (ii) the cycling occurred because the Fast allele increased in frequency every winter only within the O₇ chromosomal class, but not within the O_{ST} class (Rodríguez-Trelles, 2003). The behavior of *Mpi* could be in part an outcome of hitch-hiking with other linked loci involved in seasonal adaptation. One such candidate could be the *Na pumpα subunit (Atpα)* gene, located only 0.13 Mb farther away from O₇ than *Mpi*, and recently found to be under positive selection for defense against plant secondary compounds in *D. subobscura* (Pegueroles et al., 2016). Still, immune elicitation in *Drosophila* relies upon massive upregulation of glycolysis (Dolezal et al., 2019), which should place a strong demand on MPI activity (Shtraizent et al., 2017). In addition to the evidence from *D. subobscura* just discussed, **Supplementary Table 2** provides additional loci found to exhibit seasonal variation in a genomic survey from other *Drosophila*, which may be candidates for being involved in the seasonal cycling of O₇.

CONCLUSION AND OUTLOOK

Previous work on the spatiotemporal distribution patterns of the inversion polymorphisms of *D. subobscura* indicated that O₇ is driven by selective factors other than temperature alone. Here, we addressed this issue using a genome-based approach to isolate and characterize the O₇ breakpoints. Our findings have general implications for current theories on the molecular mechanisms of formation of this common type of structural genomic change. Furthermore, they suggest that

O₇ may have altered fly's immunometabolism through at least direct effects on core immunity and metabolism genes. This result could help to explain the inversion's conflicting correlations with the seasonal and decadal climate changes, taking into account recent findings from microbial ecology, which indicate that microbial community responses to short- and long-term climate changes can be largely uncorrelated (Romero-Olivares et al., 2017). Considering its large size, it seems likely that O₇'s evolution is also shaped by additional direct or/and indirect effects on genes other than those near its breakpoints. Further progress along this line will include development of functional tests of the identified genes on inverted versus uninverted chromosome backgrounds and use of the obtained assembly for building a SNP panel for O chromosome-wide scans of selection. We have incorporated the chromosome-scale sequence of O₃₊₄₊₇ obtained here into our reference genome browser⁹ to facilitate the further use of this resource.

DATA AVAILABILITY STATEMENT

Datasets presented in this article are available at the European Nucleotide Archive (ENA) under the project ID: PRJEB38585.

AUTHOR CONTRIBUTIONS

CK, RT, and FR-T contributed to the design and implementation of the research, to the analysis of the results, and to the writing of the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by the Spanish Ministerio de Ciencia e Innovación grant CGL2017-89160P; and Generalitat de Catalunya grant 2017SGR 1379 to the Grup de Genòmica, Bioinformàtica i Biologia Evolutiva, Universitat Autònoma de Barcelona (Spain). CK was supported by a PIF Ph.D. fellowship from the Universitat Autònoma de Barcelona (Spain). Note that the funding agencies were not involved in the design of the study or in any aspect of the collection, analysis and interpretation of the data or paper writing.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2020.565836/full#supplementary-material>

⁹<http://dsubobscura.serveftp.com/>

REFERENCES

- Aguilera, A., and Gómez-González, B. (2008). Genome instability: a mechanistic view of its causes and consequences. *Nat. Rev. Genet.* 9, 204–217. doi: 10.1038/nrg2268
- Alexander, J. L., Beagan, K., Orr-Weaver, T. L., and McVey, M. (2016). Multiple mechanisms contribute to double-strand break repair at rereplication forks in *Drosophila* follicle cells. *Proc. Natl. Acad. Sci. U.S.A.* 113, 13809–13814. doi: 10.1073/pnas.1617110113
- Andjelković, M., Jones, P. F., Grossniklaus, U., Cron, P., Schier, A. F., Dick, M., et al. (1995). Developmental regulation of expression and activity of multiple forms of the *Drosophila* RAC protein kinase. *J. Biol. Chem.* 270, 4066–4075. doi: 10.1074/jbc.270.8.4066
- Anisimova, M., Gil, M., Dufayard, J. F., Dessimoz, C., and Gascuel, O. (2011). Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. *Syst. Biol.* 60, 685–699. doi: 10.1093/sysbio/syr041
- Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., et al. (2004). UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* 32, D115–D119. doi: 10.1093/nar/gkw1099
- Arenas, C., Zivanovic, G., and Mestres, F. (2018). Chromosomal thermal index: a comprehensive way to integrate the thermal adaptation of *Drosophila subobscura* whole karyotype. *Genome* 61, 73–78.
- Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., and Cherry, J. M. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25, 25–29. doi: 10.1038/75556
- Ayala, F. J., Serra, L., and Prevosti, A. (1989). A grand experiment in evolution: the *Drosophila subobscura* colonization of the Americas. *Genome* 31, 246–255. doi: 10.1139/g89-042
- Bächli, G. (2020). *TaxoDros: The Database on Taxonomy of Drosophilidae*. Available online at: <https://www.taxodros.uzh.ch> (accessed February 21, 2020).
- Bailey, T. L., Johnson, J., Grant, C. E., and Noble, W. S. (2015). The MEME Suite. *Nucleic Acids Res.* 43, W39–W49. doi: 10.1093/nar/gkv416
- Balanyà, J., Oller, J. M., Huey, R. B., Gilchrist, G. W., and Serra, L. (2006). Global genetic change tracks global climate warming in *Drosophila subobscura*. *Science* 313, 1773–1775. doi: 10.1126/science.1131002
- Bao, W., Kojima, K. K., and Kohany, O. (2015). Repbase update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* 6:11.
- Becker, T., Loch, G., Beyer, M., Zinke, I., Aschenbrenner, A. C., Carrera, P., et al. (2010). FOXO-dependent regulation of innate immune homeostasis. *Nature* 463, 369–373. doi: 10.1038/nature08698
- Behrman, E. L., Howick, V. M., Kapun, M., Staubach, F., Bergland, A. O., Petrov, D. A., et al. (2018). Rapid seasonal evolution in innate immunity of wild *Drosophila melanogaster*. *Proc. Biol. Sci.* 285:20172599. doi: 10.1098/rspb.2017.2599
- Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 27, 573–580. doi: 10.1093/nar/27.2.573
- Bhandari, J., Karg, T., and Golic, K. G. (2019). Homolog-dependent repair following dicentric chromosome breakage in *Drosophila melanogaster*. *Genetics* 212, 615–630. doi: 10.1534/genetics.119.302247
- Bracewell, R., Chatla, K., Nalley, M. J., and Bachtrog, D. (2019). Dynamic turnover of centromeres drives karyotype evolution in *Drosophila*. *eLife* 8:e49002. doi: 10.7554/eLife.49002
- Brázda, V., Kolomazník, J., Lýsek, J., Hároníková, L., Coufal, J., and Šťastný, J. (2016). Palindrome analyser - A new web-based server for predicting and evaluating inverted repeats in nucleotide sequences. *Biochem. Biophys. Res. Commun.* 478, 1739–1745. doi: 10.1016/j.bbrc.2016.09.015
- Brehm, A., and Krimbas, C. B. (1988). The inversion polymorphism of *Drosophila subobscura* natural populations from Portugal. *Genét. Ibér.* 39, 235–248.
- Buchon, N., Silverman, N., and Cherry, S. (2014). Immunity in *Drosophila melanogaster* – from microbial recognition to whole-organism physiology. *Nat. Rev. Immunol.* 14, 796–810. doi: 10.1038/nri3763
- Cáceres, M., Ranz, J. M., Barbadilla, A., Long, M., and Ruiz, A. (1999). Generation of a widespread *Drosophila* inversion by a transposable element. *Science* 285, 415–418. doi: 10.1126/science.285.5426.415
- Campbell, M. S., Holt, C., Moore, B., and Yandell, M. (2014). Genome annotation and curation using MAKER and MAKER-P. *Curr. Protoc. Bioinformatics* 48, 4.11.1–4.11.39. doi: 10.1002/0471250953.bi0411s48
- Carvalho, C. M., Pfundt, R., King, D. A., Lindsay, S. J., Zuccherato, L. W., and Macville, M. V. (2015). Absence of heterozygosity due to template switching during replicative rearrangements. *Am. J. Hum. Genet.* 96, 555–564. doi: 10.1016/j.ajhg.2015.01.021
- Cer, R. Z., Bruce, K. H., Donohue, D. E., Temiz, N. A., Mudunuri, U. S., Yi, M., et al. (2012). Searching for non-B DNA-forming motifs using nBMST (non-B DNA motif search tool). *Curr. Protoc. Hum. Genet.* 18, 11–22. doi: 10.1002/0471142905.hg1807s73
- Chakraborty, M., Baldwin-Brown, J. G., Long, A. D., and Emerson, J. J. (2016). Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. *Nucleic Acids Res.* 44:e147. doi: 10.1093/nar/gkw654
- Chen, H., Rangasamy, M., Tan, S. Y., Wang, H., and Siegfried, B. D. (2010). Evaluation of five methods for total DNA extraction from Western corn rootworm beetles. *PLoS One* 5:e11963. doi: 10.1371/journal.pone.0011963
- Cheng, C., Tan, J. C., Hahn, M. W., and Besansky, N. J. (2018). Systems genetic analysis of inversion polymorphisms in the malaria mosquito *Anopheles gambiae*. *Proc. Natl. Acad. Sci. U.S.A.* 115, E7005–E7014. doi: 10.1073/pnas.1806760115
- Cirera, S., Martín-Campos, J. M., Segarra, C., and Aguadé, M. (1995). Molecular characterization of the breakpoints of an inversion fixed between *Drosophila melanogaster* and *D. subobscura*. *Genetics* 139, 321–326.
- Corbett-Detig, R. B. (2016). Selection on inversion breakpoints favors proximity to pairing sensitive sites in *Drosophila melanogaster*. *Genetics* 204, 259–265. doi: 10.1534/genetics.116.190389
- Corbett-Detig, R. B., and Hartl, D. L. (2012). Population genomics of inversion polymorphisms in *Drosophila melanogaster*. *PLoS Genet.* 8:e1003056. doi: 10.1371/journal.pgen.1003056
- Crescente, J. M., Zavallo, D., Helguera, M., and Vanzetti, L. S. (2018). MITE Tracker: an accurate approach to identify miniature inverted-repeat transposable elements in large genomes. *BMC Bioinformatics* 19:348. doi: 10.1186/s12859-018-2376-y
- de Frutos, R. (1972). Contribution to the study of chromosomal polymorphism in the Spanish populations of *Drosophila subobscura*. *Genét. Ibér.* 24, 123–140.
- de Jong, G., and Bochdanovits, Z. (2003). Latitudinal clines in *Drosophila melanogaster*: body size, allozyme frequencies, inversion frequencies, and the insulin-signalling pathway. *J. Genet.* 82, 207–223. doi: 10.1007/BF02715819
- Delprat, A., Guillén, Y., and Ruiz, A. (2019). Computational sequence analysis of inversion breakpoint regions in the cactophilic *Drosophila mojavensis* lineage. *J. Hered.* 110, 102–117. doi: 10.1093/jhered/esy057
- Deng, W., Maust, B. S., Nickle, D. C., Learn, G. H., Liu, Y., Heath, L., et al. (2010). DIVEIN: a web server to analyze phylogenies, sequence divergence, diversity, and informative sites. *Biotechniques* 48, 405–408. doi: 10.2144/000113370
- Dionne, M. S., Pham, L. N., Shirasu-Hiza, M., and Schneider, D. S. (2006). Akt and foxo dysregulation contribute to infection-induced wasting in *Drosophila*. *Curr. Biol.* 16, 1977–1985. doi: 10.1016/j.cub.2006.08.052
- Dobzhansky, T. (1947). Genetics of natural populations. XIV. A response of certain gene arrangements in the third chromosome of *Drosophila pseudoobscura* to natural selection. *Genetics* 32, 142–160.
- Dolezal, T., Krejčová, G., Bajgar, A., Nedbalová, P., and Strasser, P. (2019). Molecular regulations of metabolism during immune response in insects. *Insect. Biochem. Mol. Biol.* 109, 31–42. doi: 10.1016/j.ibmb.2019.04.005
- Dolgova, O. (2013). *Genetic and Phenotypic Differentiation in three Chromosomal Arrangements of Drosophila subobscura*. Ph.D. thesis, Universitat Autònoma de Barcelona, Barcelona.
- Durmaz, E., Rajpurohit, S., Betancourt, N., Fabian, D. K., Kapun, M., Schmidt, P., et al. (2019). A clinal polymorphism in the insulin signaling transcription factor foxo contributes to life-history adaptation in *Drosophila*. *Evolution* 73, 1774–1792. doi: 10.1111/evo.13759
- Eilbeck, K., Lewis, S., Mungall, C., Yandell, M., Stein, L., Durbin, R., et al. (2005). The sequence ontology: a tool for the unification of genome annotations. *Genome Biol.* 6:R44. doi: 10.1186/gb-2005-6-5-r44
- Ellinghaus, D., Kurtz, S., and Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* 9:18. doi: 10.1186/1471-2105-9-18
- Fabian, D. K., Kapun, M., Nolte, V., Kofler, R., Schmidt, P. S., Schlotterer, C., et al. (2012). Genome-wide patterns of latitudinal differentiation among populations of *Drosophila melanogaster* from North America. *Mol. Ecol.* 21, 4748–4769. doi: 10.1111/j.1365-294X.2012.05731.x

- Faria, R., Johannesson, K., Butlin, R. K., and Westram, A. M. (2019). Evolving inversions. *Trends Ecol. Evol.* 34, 239–248. doi: 10.1016/j.tree.2018.12.005
- Farré, M., Robinson, T. J., and Ruiz-Herrera, A. (2015). An integrative breakage model of genome architecture, reshuffling and evolution. *Bioessays* 37, 479–488. doi: 10.1002/bies.201400174
- Finn, R. D., Attwood, T. K., Babbitt, P. C., Bateman, A., Bork, P., Bridge, A. J., et al. (2017). InterPro in 2017—beyond protein family and domain annotations. *Nucleic Acids Res.* 45, D190–D199. doi: 10.1093/nar/gkw1107
- Finn, R. D., Coghill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., et al. (2016). The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* 44, D279–D285. doi: 10.1093/nar/gkv1344
- Flatt, T., and Partridge, L. (2018). Horizons in the evolution of aging. *BMC Biol.* 16:93. doi: 10.1186/s12915-018-0562-z
- Fontdevila, A., Zapata, C., Alvarez, G., Sanchez, L., Méndez, J., and Enriquez, I. (1983). Genetic coadaptation in the chromosomal polymorphism of *Drosophila subobscura*. I. Seasonal changes of gametic disequilibrium in a natural population. *Genetics* 105, 935–955.
- Fragata, I., Lopes-Cunha, M., Bárbaro, M., Kellen, B., Lima, M., Santos, M. A., et al. (2014). How much can history constrain adaptive evolution? A real-time evolutionary approach of inversion polymorphisms in *Drosophila subobscura*. *J. Evol. Biol.* 27, 2727–2738. doi: 10.1111/jeb.12533
- Fuller, Z. L., Haynes, G. D., Richards, S., and Schaeffer, S. W. (2016). Genomics of natural populations: How differentially expressed genes shape the evolution of chromosomal inversions in *Drosophila pseudoobscura*. *Genetics* 204, 287–301. doi: 10.1534/genetics.116.191429
- Fuller, Z. L., Haynes, G. D., Richards, S., and Schaeffer, S. W. (2017). Genomics of natural populations: evolutionary forces that establish and maintain gene arrangements in *Drosophila pseudoobscura*. *Mol. Ecol.* 26, 6539–6562. doi: 10.1111/mec.14381
- Fuller, Z. L., Koury, S. A., Phadnis, N., and Schaeffer, S. W. (2019). How chromosomal rearrangements shape adaptation and speciation: case studies in *Drosophila pseudoobscura* and its sibling species *Drosophila persimilis*. *Mol. Ecol.* 28, 1283–1301. doi: 10.1111/mec.14923
- Gienapp, P., Teplitsky, C., Alho, J. S., Mills, J. A., and Merilä, J. (2008). Climate change and evolution: disentangling environmental and genetic responses. *Mol. Ecol.* 17, 167–178. doi: 10.1111/j.1365-294X.2007.03413.x
- Götz, W. (1965). Beitrag zur Kenntnis der Inversionen, Duplikationen und Strukturtypen von *Drosophila subobscura* Coll. *Z. Vererbungsl.* 96, 285–296. doi: 10.1007/BF00896828
- Götz, W. (1967). Untersuchungen über den chromosomalen Strukturpolymorphismus in kleinasiatischen und persischen Populationen von *Drosophila subobscura* Coll. *Mol. Gen. Genet.* 100, 1–38. doi: 10.1007/BF00425773
- Grant, C. E., Bailey, T. L., and Noble, W. S. (2011). FIMO: scanning for occurrences of a given motif. *Bioinformatics* 27, 1017–1018. doi: 10.1093/bioinformatics/btr064
- Greig, D. (2007). A screen for recessive speciation genes expressed in the gametes of F1 hybrid yeast. *PLoS Genet.* 3:e21. doi: 10.1371/journal.pgen.0030021
- Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* 59, 307–321. doi: 10.1093/sysbio/syq010
- Hanson, M. A., and Lemaître, B. (2020). New insights on *Drosophila* antimicrobial peptide function in host defense and beyond. *Curr. Opin. Immunol.* 62, 22–30. doi: 10.1016/j.coi.2019.11.008
- Hastings, P. J., Ira, G., and Lupski, J. R. (2009). A microhomology-mediated break-induced replication model for the origin of human copy number variation. *PLoS Genet.* 5:e1000327. doi: 10.1371/journal.pgen.1000327
- Hill, T., and Betancourt, A. J. (2018). Extensive exchange of transposable elements in the *Drosophila pseudoobscura* group. *Mob. DNA* 9:20.
- Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q., and Vinh, L. S. (2018). UFBoot2: improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* 35, 518–522. doi: 10.1093/molbev/msx281
- Hoffmann, A. A., and Rieseberg, L. H. (2008). Revisiting the impact of inversions in evolution: from population genetic markers to drivers of adaptive shifts and speciation? *Annu. Rev. Ecol. Syst.* 39, 21–42. doi: 10.1146/annurev.ecolsys.39.110707.173532
- Holt, C., and Yandell, M. (2011). MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* 12:491. doi: 10.1186/1471-2105-12-491
- Hughes, L. (2000). Biological consequences of global warming: Is the signal already apparent? *Trends Ecol. Evol.* 15, 56–61.
- Hughes, S. E., Miller, D. E., Miller, A. L., and Hawley, R. S. (2018). Female meiosis: synapsis, recombination, and segregation in *Drosophila melanogaster*. *Genetics* 208, 875–908. doi: 10.1534/genetics.117.300081
- Huson, D. H., and Bryant, I. D. (2006). Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* 23, 254–267. doi: 10.1093/molbev/msj030
- Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., et al. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30, 1236–1240. doi: 10.1093/bioinformatics/btu031
- Joyce, E. F., Paul, A., Chen, K. E., Tanneti, N., and McKim, K. S. (2012). Multiple barriers to nonhomologous DNA end joining during meiosis in *Drosophila*. *Genetics* 191, 739–746. doi: 10.1534/genetics.112.140996
- Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., and Jermini, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* 14, 587–589. doi: 10.1038/nmeth.4285
- Kapun, M., Fabian, D. K., Goudet, J., and Flatt, T. (2016). Genomic evidence for adaptive inversion clines in *Drosophila melanogaster*. *Mol. Biol. Evol.* 33, 1317–1336. doi: 10.1093/molbev/msw016
- Kapun, M., and Flatt, T. (2018). The adaptive significance of chromosomal inversion polymorphisms in *Drosophila melanogaster*. *Mol. Ecol.* 28, 1263–1282. doi: 10.1111/mec.14871
- Karageorgiou, C., Gámez-Visairas, V., Tarrío, R., and Rodríguez-Trelles, F. (2019). Long-read based assembly and synteny analysis of a reference *Drosophila subobscura* genome reveals signatures of structural evolution driven by inversions recombination-suppression effects. *BMC Genomics* 20:223.
- Katoh, K., Rozewicki, J., and Yamada, K. D. (2019). MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief. Bioinformatics* 20, 1160–1166. doi: 10.1093/bib/bbx108
- Kaushal, S., and Freudenreich, C. H. (2019). The role of fork stalling and DNA structures in causing chromosome fragility. *Genes Chromosomes Cancer* 58, 270–283. doi: 10.1002/gcc.22721
- Kehrer-Sawatzki, H., Sandig, C. A., Goidts, V., and Hameister, H. (2005). Breakpoint analysis of the pericentric inversion between chimpanzee chromosome 10 and the homologous chromosome 12 in humans. *Cytogenet. Genome Res.* 108, 91–97. doi: 10.1159/000080806
- Kirkpatrick, M. (2010). How and why chromosome inversions evolve. *PLoS Biol.* 8:e1000501. doi: 10.1371/journal.pbio.1000501
- Kirkpatrick, M., and Barton, N. (2006). Chromosome inversions, local adaptation and speciation. *Genetics* 173, 419–434. doi: 10.1534/genetics.105.047985
- Kolb, J., Chuzhanova, N. A., Högel, J., Vasquez, K. M., Cooper, D. N., Bacolla, A., et al. (2009). Cruciform-forming inverted repeats appear to have mediated many of the microinversions that distinguish the human and chimpanzee genomes. *Chromosome Res.* 17, 469–483.
- Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., and Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* 27, 722–736. doi: 10.1101/gr.215087.116
- Koske, T., and Maynard-Smith, J. (1954). Genetics and cytology of *Drosophila subobscura*. X. The fifth linkage group. *J. Genet.* 52, 521–541. doi: 10.1007/BF02985076
- Kramara, J., Osia, B., and Malkova, A. (2018). Break-induced replication: the where, the why, and the how. *Trends Genet.* 34, 518–531. doi: 10.1016/j.tig.2018.04.002
- Krimbas, C. B. (1992). “The inversion polymorphism of *Drosophila subobscura*,” in *Drosophila Inversion Polymorphism*, eds C. B. Krimbas and J. R. Powell (Boca Raton, FL: CRC Press), 127–220.
- Kulakovskiy, I. V., and Makeev, V. J. (2009). Discovery of DNA motifs recognized by transcription factors through integration of different experimental sources. *Biophysics* 54, 667–674. doi: 10.1134/S0006350909060013
- Kunze-Mühl, E., and Müller, E. (1958). Weitere Untersuchungen über die chromosomale Struktur und die natürlichen Strukturtypen von *Drosophila subobscura* Coll. *Chromosoma* 9, 559–570. doi: 10.1007/BF02568093

- Kunze-Mühl, E., and Sperlich, D. (1955). Inversionen und chromosomale Strukturtypen bei *Drosophila subobscura* Coll. Z. Vererbungsl. 87, 65–84. doi: 10.1007/BF00308333
- Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., et al. (2004). Versatile and open software for comparing large genomes. *Genome Biol.* 5:R12. doi: 10.1186/gb-2004-5-2-r12
- Lagesen, K., Hallin, P., Rødland, E. A., Staerfeldt, H. H., Rognes, T., and Ussery, D. W. (2007). RNAMmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* 35, 3100–3108. doi: 10.1093/nar/gkm160
- Lazzaro, A., Hilfiker, D., and Zeyer, J. (2015). Structures of microbial communities in alpine soils: seasonal and elevational effects. *Front. Microbiol.* 6:1330. doi: 10.3389/fmicb.2015.01330
- Lazzaro, B. P., and Clark, A. G. (2001). Evidence for recurrent paralogous gene conversion and exceptional allelic divergence in the Attacin genes of *Drosophila melanogaster*. *Genetics* 159, 659–671.
- Lazzaro, B. P., and Clark, A. G. (2003). Molecular population genetics of inducible antibacterial peptide genes in *Drosophila melanogaster*. *Mol. Biol. Evol.* 20, 914–923. doi: 10.1093/molbev/msg109
- Lee, J. A., Carvalho, C. M., and Lupski, J. R. (2007). A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell* 131, 1235–1247. doi: 10.1016/j.cell.2007.11.037
- Leigh Brown, A. J., and Ish-Horowitz, D. (1981). Evolution of the 87A and 87C heat-shock loci in *Drosophila*. *Nature* 290, 677–682.
- Lobachev, K. S., Rattray, A., and Narayanan, V. (2007). Hairpin- and cruciform-mediated chromosome breakage: causes and consequences in eukaryotic cells. *Front. Biosci.* 12, 4208–4220. doi: 10.2741/2381
- Lopes, M., Foiani, M., and Sogo, J. M. (2006). Multiple mechanisms control chromosome integrity after replication fork uncoupling and restart at irreparable UV lesions. *Mol. Cell* 21, 15–27. doi: 10.1016/j.molcel.2005.11.015
- Loukas, M., Krimbas, C. B., and Vergini, Y. (1979). The genetics of *Drosophila subobscura* populations. IX. Studies on linkage disequilibrium in four natural populations. *Genetics* 93, 497–523.
- Lowe, T. M., and Chan, P. P. (2016). tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. *Nucleic Acids Res.* 44, W54–W57. doi: 10.1093/nar/gkw413
- Lowry, D. B., Popovic, D., Brennan, D. J., and Holeski, L. M. (2019). Mechanisms of a locally adaptive shift in allocation among growth, reproduction, and herbivore resistance in *Mimulus guttatus*. *Evolution* 73, 1168–1181. doi: 10.1111/evo.13699
- Lu, S., Wang, G., Bacolla, A., Zhao, J., Spitzer, S., and Vasquez, K. M. (2015). Short inverted repeats are hotspots for genetic instability: relevance to cancer genomes. *Cell Rep.* 10, 1674–1680. doi: 10.1016/j.celrep.2015.02.039
- Mani, R. S., and Chinnaiyan, A. M. (2010). Triggers for genomic rearrangements: insights into genomic, cellular and environmental influences. *Nat. Rev. Genet.* 11, 819–829. doi: 10.1038/nrg2883
- Matzkin, L. M., Merritt, T. J., Zhu, C. T., and Eanes, W. F. (2005). The structure and population genetics of the breakpoints associated with the cosmopolitan chromosomal inversion In(3R)Payne in *Drosophila melanogaster*. *Genetics* 170, 1143–1152. doi: 10.1534/genetics.104.038810
- Maynard-Smith, J., and Maynard-Smith, S. (1954). Genetics and cytology of *Drosophila subobscura*. VIII. Heterozygosity, viability and rate of development. *J. Genet.* 52, 152–164. doi: 10.1007/BF02981496
- McBroome, J., Liang, D., and Corbett-Detig, R. (2020). Fine-scale position effects shape the distribution of inversion breakpoints in *Drosophila melanogaster*. *Genome Biol. Evol.* doi: 10.1093/gbe/evaa103 [Epub ahead of print].
- McKinney, J. A., Wang, G., Mukherjee, A., Christensen, L., Subramanian, S. H. S., Zhao, J., et al. (2020). Distinct DNA repair pathways cause genomic instability at alternative DNA structures. *Nat. Commun.* 11:236.
- Menozi, P., and Krimbas, C. B. (1992). The inversion polymorphism of *D. subobscura* revisited: synthetic maps of gene arrangement frequencies and their interpretation. *J. Evol. Biol.* 5, 625–641. doi: 10.1046/j.1420-9101.1992.5040625.x
- Messer, P. W., Ellner, S. P., and Hairston, N. G. Jr. (2016). Can population genetics adapt to rapid evolution? *Trends Genet.* 32, 408–418. doi: 10.1016/j.tig.2016.04.005
- Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274. doi: 10.1093/molbev/msu300
- Orengo, D. J., Puerma, E., Cereijo, U., and Aguadé, M. (2019). The molecular genealogy of sequential overlapping inversions implies both homologous chromosomes of a heterokaryotype in an inversion origin. *Sci. Rep.* 9:17009.
- Orengo, D. J., Puerma, E., Papaceit, M., Segarra, C., and Aguadé, M. A. (2015). Molecular perspective on a complex polymorphic inversion system with cytological evidence of multiply reused breakpoints. *Heredity* 114, 610–618. doi: 10.1038/hdy.2015
- Paaby, A. B., Bergland, A. O., Behrman, E. L., and Schmidt, P. S. (2014). A highly pleiotropic amino acid polymorphism in the *Drosophila* insulin receptor contributes to life-history adaptation. *Evolution* 68, 3395–3409. doi: 10.1111/evo.12546
- Papaceit, M., Segarra, C., and Aguadé, M. (2012). Structure and population genetics of the breakpoints of a polymorphic inversion in *Drosophila subobscura*. *Evolution* 67, 66–79.
- Parmesan, C. (2006). Ecological and evolutionary responses to recent climate change. *Annu. Rev. Ecol. Evol. Syst.* 37, 637–669. doi: 10.1146/annurev.ecolsys.37.091305.110100
- Pegueroles, C., Aquadro, C. F., Mestres, F., and Pascual, M. (2013). Gene flow and gene flux shape evolutionary patterns of variation in *Drosophila subobscura*. *Heredity* 110, 520–529. doi: 10.1038/hdy.2012.118
- Pegueroles, C., Araúz, P. A., Pascual, M., and Mestres, F. (2010a). A recombination survey using microsatellites: the O chromosome of *Drosophila subobscura*. *Genetica* 138, 795–804.
- Pegueroles, C., Ferrés-Coy, A., Martí-Solano, M., Aquadro, C. F., Pascual, M., and Mestres, F. (2016). Inversions and adaptation to the plant toxin ouabain shape DNA sequence variation within and between chromosomal inversions of *Drosophila subobscura*. *Sci. Rep.* 6:23754. doi: 10.1038/srep23754
- Pegueroles, C., Ordoñez, V., Mestres, F., and Pascual, M. (2010b). Recombination and selection in the maintenance of the adaptive value of inversions. *J. Evol. Biol.* 23, 2709–2717. doi: 10.1111/j.1420-9101.2010.02136.x
- Powell, J. R. (1997). *Progress and Prospects in Evolutionary Biology: The Drosophila Model*. Oxford: Oxford University Press.
- Prazeres da Costa, O., González, J., and Ruiz, A. (2009). Cloning and sequencing of the breakpoint regions of inversion 5g fixed in *Drosophila buzzatii*. *Chromosoma* 118, 349–360. doi: 10.1007/s00412-008-0201-5
- Prevosti, A., Ribo, G., Serra, L., Aguadé, M., Balaña, J., Monclus, M., et al. (1988). Colonization of America by *Drosophila subobscura*: experiment in natural populations that supports the adaptive role of chromosomal-inversion polymorphism. *Proc. Natl. Acad. Sci. U.S.A.* 85, 5597–5600. doi: 10.1073/pnas.85.15.5597
- Puerma, E., Orengo, D. J., and Aguadé, M. (2016a). Multiple and diverse structural changes affect the breakpoint regions of polymorphic inversions across the *Drosophila* genus. *Sci. Rep.* 6:36248. doi: 10.1038/srep36248
- Puerma, E., Orengo, D. J., and Aguadé, M. (2016b). The origin of chromosomal inversions as a source of segmental duplications in the *Sophophora* subgenus of *Drosophila*. *Sci. Rep.* 6:30715. doi: 10.1038/srep30715
- Puerma, E., Orengo, D. J., and Aguadé, M. (2017). Inversion evolutionary rates might limit the experimental identification of inversion breakpoints in non-model species. *Sci. Rep.* 7:17281.
- Puerma, E., Orengo, D. J., Cruz, F., Gómez-Garrido, J., Librado, P., Salguero, D., et al. (2018). The high-quality genome sequence of the oceanic island endemic species *Drosophila guanche* reveals signals of adaptive evolution in genes related to flight and genome stability. *Genome Biol. Evol.* 10, 1956–1969. doi: 10.1093/gbe/evy135
- Puerma, E., Orengo, D. J., Salguero, D., Papaceit, M., Segarra, C., and Aguadé, M. (2014). Characterization of the breakpoints of a polymorphic inversion complex detects strict and broad breakpoint reuse at the molecular level. *Mol. Biol. Evol.* 31, 2331–2341. doi: 10.1093/molbev/msu177
- Puig-Giribets, M., Guerreiro, M. P. G., Santos, M., Ayala, F. J., Tarrío, R., and Rodríguez-Trelles, F. (2019). Chromosomal inversions promote genomic islands of concerted evolution of Hsp70 genes in the *Drosophila subobscura* species subgroup. *Mol. Ecol.* 28, 1316–1332. doi: 10.1111/mec.14511
- Ranz, J. M., Maurin, D., Chan, Y. S., von Grothuss, M., Hillier, L. W., Roote, J., et al. (2007). Principles of genome evolution in the *Drosophila melanogaster* species group. *PLoS Biol.* 5:e152. doi: 10.1371/journal.pbio.0050152
- Reese, M. G. (2001). Application of a time-delay neural network to promoter annotation in the *Drosophila melanogaster* genome. *Comput. Chem.* 26, 51–56.

- Regan, J. C., Froy, H., Walling, C. A., Moatt, J. P., and Nussey, D. H. (2020). Dietary restriction and insulin-like signalling pathways as adaptive plasticity: a synthesis and re-evaluation. *Funct. Ecol.* 34, 107–128. doi: 10.1111/1365-2435.13418
- Rego, C., Balanyà, J., Fragata, I., Matos, M., Rezende, E. L., and Santos, M. (2010). Clinal patterns of chromosomal inversion polymorphisms in *Drosophila subobscura* are partly associated with thermal preferences and heat stress resistance. *Evolution* 64, 385–397. doi: 10.1111/j.1558-5646.2009.00835.x
- Rezende, E. L., Balanyà, J., Rodríguez-Trelles, F., Rego, C., Fragata, I., Matos, M., et al. (2010). Climate change and chromosomal inversions in *Drosophila subobscura*. *Clim. Res.* 43, 103–114. doi: 10.1007/s10709-018-0035-x
- Richards, S., Liu, Y., Bettencourt, B. R., Hradecky, P., Letovsky, S., and Nielsen, R. (2005). Comparative genome sequencing of *Drosophila pseudoobscura*: chromosomal, gene, and cis-element evolution. *Genome Res.* 15, 1–18. doi: 10.1101/gr.3059305
- Rivera, J., Keränen, S. V. E., Gallo, S. M., and Halfon, M. S. (2019). REDfly: the transcriptional regulatory element database for *Drosophila*. *Nucleic Acids Res.* 47, D828–D834. doi: 10.1093/nar/gky957
- Rodríguez-Trelles, F. (2003). Seasonal cycles of allozyme-by-chromosomal-inversion gametic disequilibrium in *Drosophila subobscura*. *Evolution* 57, 839–848. doi: 10.1111/j.0014-3820.2003.tb00295.x
- Rodríguez-Trelles, F., Alvarez, G., and Zapata, C. (1996). Time-series analysis of seasonal changes of the O inversion polymorphism of *Drosophila subobscura*. *Genetics* 142, 179–187.
- Rodríguez-Trelles, F., and Rodríguez, M. A. (1998). Rapid micro-evolution and loss of chromosomal diversity in *Drosophila* in response to climate warming. *Evol. Ecol.* 12, 829–838. doi: 10.1023/A:1006546616462
- Rodríguez-Trelles, F., and Rodríguez, M. A. (2007). Comment on ‘Global genetic change tracks global climate warming in *Drosophila subobscura*’. *Science* 315:1497. doi: 10.1126/science.1136298
- Rodríguez-Trelles, F., and Rodríguez, M. A. (2010). Measuring evolutionary responses to global warming: cautionary lessons from *Drosophila*. *Insect Conserv. Divers.* 3, 44–50. doi: 10.1111/j.1752-4598.2009.00071.x
- Rodríguez-Trelles, F., Tarrío, R., and Santos, M. (2013). Genome-wide evolutionary response to a heat wave in *Drosophila*. *Biol. Lett.* 9:20130228. doi: 10.1098/rsbl.2013.0228
- Romero-Olivares, A. L., Allison, S. D., and Treseder, K. K. (2017). Soil microbes and their response to experimental warming over time: a meta-analysis of field studies. *Soil Biol. Biochem.* 107, 32–40. doi: 10.1016/j.soilbio.2016.12.026
- Said, I., Byrne, A., Serrano, V., Cardeno, C., Vollmers, C., and Corbett-Detig, R. B. (2018). Linked genetic variation and not genome structure causes widespread differential expression associated with chromosomal inversions. *Proc. Natl. Acad. Sci. U.S.A.* 115, 5492–5497. doi: 10.1073/pnas.1721275115
- Sakofsky, C. J., Ayyar, S., Deem, A. K., Chung, W. H., Ira, G., and Malkova, A. (2015). Translesion polymerases drive microhomology-mediated break-induced replication leading to complex chromosomal rearrangements. *Mol. Cell* 60, 860–872. doi: 10.1016/j.molcel.2015.10.041
- Santos, J., Serra, L., Solé, E., and Pascual, M. (2010). FISH mapping of microsatellite loci from *Drosophila subobscura* and its comparison to related species. *Chromosome Res.* 18, 213–226.
- Santos, M., Céspedes, W., Balanyà, J., Trotta, V., Calboli, F. C., Fontdevila, A., et al. (2005). Temperature-related genetic changes in laboratory populations of *Drosophila subobscura*: evidence against simple climatic-based explanations for latitudinal clines. *Am. Nat.* 165, 258–273. doi: 10.1086/427093
- Savage, J. R. (1976). Classification and relationships of induced chromosomal structural changes. *J. Med. Genet.* 13, 103–122. doi: 10.1136/jmg.13.2.103
- Schwenke, R. A., Lazzaro, B. P., and Wolfner, M. F. (2016). Reproduction-immunity trade-offs in insects. *Annu. Rev. Entomol.* 61, 239–256.
- Scully, R., Panday, A., Elango, R., and Willis, N. A. (2019). DNA double-strand break repair-pathway choice in somatic mammalian cells. *Nat. Rev. Mol. Cell Biol.* 20, 698–714.
- Sebastián, A., and Contreras-Moreira, B. (2014). FootprintDB: a database of transcription factors with annotated cis elements and binding interfaces. *Bioinformatics* 30, 258–265. doi: 10.1093/bioinformatics/btt663
- Senger, K., Armstrong, G. W., Rowell, W. J., Kwan, J. M., Markstein, M., and Levine, M. (2004). Immunity regulatory DNAs share common organizational features in *Drosophila*. *Mol. Cell* 13, 19–32.
- Seppy, M., Manni, M., and Zdobnov, E. M. (2019). BUSCO: assessing genome assembly and annotation completeness. *Methods Mol. Biol.* 1962, 227–245. doi: 10.1007/978-1-4939-9173-0_14
- Shigyo, N., Umeki, K., and Hirao, T. (2019). Seasonal dynamics of soil fungal and bacterial communities in cool-temperate montane forests. *Front. Microbiol.* 10:1944. doi: 10.3389/fmicb.2019.01944
- Shtraizent, N., DeRossi, C., Nayar, S., Sachidanandam, R., Katz, L. S., Prince, A., et al. (2017). MPI depletion enhances O-GlcNAcylation of p53 and suppresses the Warburg effect. *eLife* 6:e22477. doi: 10.7554/eLife.22477
- Slade, J. D., and Staveley, B. E. (2016). Enhanced survival of *Drosophila* Akt1 hypomorphs during amino-acid starvation requires foxo. *Genome* 59, 87–93.
- Smit, A. F. A., Hubley, R., and Green, P. (2013/2015). RepeatMasker Open-4.0. Available online at: <http://www.repeatmasker.org> (accessed November 15, 2019).
- Smith, C., Llorente, B., and Symington, L. (2007). Template switching during break-induced replication. *Nature* 447, 102–105. doi: 10.1038/nature05723
- Soderlund, C., Bomhoff, M., and Nelson, W. M. (2011). SyMAP v3.4: a turnkey synteny system with application to plant genomes. *Nucleic Acids Res.* 39:e68. doi: 10.1093/nar/gkr123
- Sogo, J. M., Lopes, M., and Foiani, M. (2002). Fork reversal and ssDNA accumulation at stalled replication forks owing to checkpoint defects. *Science* 297, 599–602. doi: 10.1126/science.1074023
- Solé, E., Balanyà, J., Sperlich, D., and Serra, L. (2002). Long-term changes in the chromosomal inversion polymorphism of *Drosophila subobscura*. I. Mediterranean populations from southwestern Europe. *Evolution* 56, 830–835. doi: 10.1111/j.0014-3820.2002.tb01393.x
- Sperlich, D., Feuerbach-Mravlag, H., Lange, P., Michaelidis, A., and Pentzos-Daponte, A. (1977). Genetic load and viability distribution in central and marginal populations of *Drosophila subobscura*. *Genetics* 86, 835–848.
- Sun, X., Qun, Y., and Xia, X. (2013). An improved implementation of effective number of codons (Nc). *Mol. Biol. Evol.* 30, 191–196. doi: 10.1093/molbev/mss201
- Sunder, S., and Wilson, E. T. (2019). Frequency of DNA end joining in trans is not determined by the predamage spatial proximity of double-strand breaks in yeast. *Proc. Natl. Acad. Sci. U.S.A.* 116, 9481–9490. doi: 10.1073/pnas.1818595116
- Tajima, F. (1993). Simple methods for testing the molecular evolutionary clock hypothesis. *Genetics* 135, 599–607.
- Teleman, A. A. (2010). Molecular mechanisms of metabolic regulation by insulin in *Drosophila*. *Biochem. J.* 425, 13–26.
- The Gene Ontology Consortium. (2017). Expansion of the gene ontology knowledgebase and resources. *Nucleic Acids Res.* 45, D331–D338. doi: 10.1093/nar/gkwl108
- Tremblay-Belzile, S., Lepage, É., Zampini, É., and Brisson, N. (2015). Short-range inversions: rethinking organelle genome stability: template switching events during DNA replication destabilize organelle genomes. *Bioessays* 37, 1086–1094. doi: 10.1002/bies.201500064
- Voineagu, I., Narayanan, V., Lobachev, K. S., and Mirkin, S. M. (2008). Replication stalling at unstable inverted repeats: interplay between DNA hairpins and fork stabilizing proteins. *Proc. Natl. Acad. Sci. U.S.A.* 105, 9936–9941. doi: 10.1073/pnas.0804510105
- Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., et al. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 9:e112963. doi: 10.1371/journal.pone.0112963
- Wang, G., Christensen, L. A., and Vasquez, K. M. (2006). Z-DNA-forming sequences generate large-scale deletions in mammalian cells. *Proc. Natl. Acad. Sci. U.S.A.* 103, 2677–2682. doi: 10.1073/pnas.0511084103
- Wang, G., and Vasquez, K. M. (2006). Non-B DNA structure-induced genetic instability. *Mutat. Res.* 598, 103–119.
- Wasserman, M. (1968). Recombination-induced chromosomal heterosis. *Genetics* 58, 125–139.
- Weirauch, M. T., Yang, A., Albu, M., Cote, A. G., Montenegro-Montero, A., Drewe, P., et al. (2014). Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* 158, 1431–1443. doi: 10.1016/j.cell.2014.08.009

- Wellenreuther, M., and Bernatchez, L. (2018). Eco-evolutionary genomics of chromosomal inversions. *Trends Ecol. Evol.* 33, 427–440. doi: 10.1016/j.tree.2018.04.002
- Wesley, C. S., and Eanes, W. F. (1994). Isolation and analysis of the breakpoint sequences of chromosome inversion In(3L)Payne in *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. U.S.A.* 91, 3132–3136. doi: 10.1073/pnas.91.8.3132
- Zhang, F., Khajavi, M., Connolly, A. M., Towne, C. F., Batish, S. D., and Lupski, J. R. (2009). The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. *Nat. Genet.* 41, 849–853. doi: 10.1038/ng.399
- Zhao, J., Bacolla, A., Wang, G., and Vasquez, K. M. (2010). Non-B DNA structure-induced genetic instability and evolution. *Cell. Mol. Life Sci.* 67, 43–62.
- Zollinger, E. (1950). Ein strukturell homozygoter Stamm von *Drosophila subobscura* aus einer Wild-population. *Arch. Klaus-Stiftg* 25, 33–35.
- Zouros, E., Krimbas, C. B., Tsakas, S., and Loukas, M. (1974). Genic versus chromosomal variation in natural populations of *D. subobscura*. *Genetics* 78, 1223–1244.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Karageorgiou, Tarrío and Rodríguez-Trelles. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

4. Discussion

The two Result Sections, 3.1 and 3.2, aim to address questions regarding the evolution of structural variation and the effect of polymorphic inversions and that of their breakpoints in natural *D. subobscura* populations. The two published articles include a “Discussion” section where our findings are discussed in depth. Thus, in this section I intend to provide a general overview and extend the discussion on those points that were briefly mentioned or not covered previously in detail in the published articles.

4.1 *De novo* genome assembly and short-read limitations

Next-generation sequencing has enabled the sequencing of whole genomes at a relatively low cost (Wetterstrand, 2021). Up-to-date long-read sequencing and *de novo* genome assembly continue to advance rapidly reaching a level of accuracy similar to that of short-read sequencing. Further, long reads not only facilitate *de novo* genome assemblies, but can improve detection of structural variants (SV). While short reads are well suited and widely used for the identification of single nucleotide polymorphisms (SNPs) and small insertion and deletions (indels), their application in structural variant detection can be troublesome and rather challenging. Reads shorter than 300 bases, such as those typically generated by Illumina next-generation sequencing platforms, are too short for intermediate-size structural variant detection (Logsdon *et al.*, 2020). Similarly, structural variants that generate duplications at their breakpoints or those that carry duplicated sequences at their ends in both the ancestral and rearranged state cannot be detected accurately using short-read sequencing and thus, SVs are commonly misidentified as alignment or sequencing artifacts (Cameron *et al.*, 2019; Mahmoud *et al.*, 2019). Further, SV can be detected via assembly comparison yet, until recently issues with genome completeness stood a major limitation to their discovery when employing this approach. Here to overcome the aforementioned limitations we have resorted to long-read sequencing and *de novo* genome assembly for the obtention of a reference genome for *D. subobscura* and the identification of the O₇ inversion and its breakpoints. This approach has allowed us to identify the molecular mechanisms for the O₇ formation and further investigate and propose alternative models of the inversion formation using basepair resolution. Overall, long-read sequencing over the past decade has revolutionized genome assembly and allowed us to overcome the technical obstacles that have hindered the genomic study of several species, enabling comprehensive studies of the entire genome and its structural variation (Jiang *et al.*, 2012; Nurk *et al.*, 2022). Currently, studies utilizing short-read sequencing possibly represent only a glimpse of the structural variation that lies within genomes. Hence, long-read sequencing and re-sequencing studies are now required to comprehensively examine structural variation and capture the full spectrum of diversity within species.

4.2 Genome browser a resource to handle data efficiently

The rapid generation of genomic data has progressively led to the development of novel tools for their manipulation and visualization. Currently there is a growing popularity and demand for genome browsers. Genome browsers are versatile web-based user interfaces that integrate genomic sequences and annotation data which are then displayed in a graphical format (Schattner, 2008). They can facilitate the visualization of different annotations such as genes, gene predictions, sliceforms, proteins, gene-expression, transposable elements, repetitive sequences, SNPs, indels and so on (Wang *et al.*, 2013). Over the years a large number of genome browsers have been developed for model species and non-model species (Stein, 2013; Buels *et al.*, 2016). To facilitate genomic research in *D. subobscura* we have built a genome browser for the reference genome and additionally integrated the genome of the O_{3+4+7} line. Our custom genome browser features all the results covered in the results section and in the two published papers *e.g.* gene annotations and predictions, transposable elements and their distribution, orthology relationships with other *Drosophila* species, fixed inversion breakpoints between *D. subobscura* and *D. guanche*, the proximal and distal breakpoints of inversion O_7 and their composition of repeats. Additionally, we have incorporated “elasticsearch” which enables the indexing of the featured annotations and allows browsing, searching and retrieving genomic sequences. Our genome browser was launched using the JBrowse application (Buels *et al.*, 2016). Moreover, a custom BLAST server for the two genomes has been built and integrated to the genome browser using SequenceServer (Priyam *et al.*, 2019). The custom BLAST server can vastly benefit researchers interested in working with *D. subobscura* by making it accessible to virtually anyone without the prior knowledge of bioinformatics or computational biology. Our genome browser offers an intuitive way to explore the genome of *D. subobscura* while facilitating comparative genomics analyses particularly for researchers that are not familiar with the species. In summary, the “*Drosophila subobscura* Genome Browser” portal (<http://dsubobscura.serveftp.com/>) serves as a compilation of the work conducted in the two published papers, provides a visual representation of the *D. subobscura* genome and its annotated features, enables BLAST searches and offers the possibility to explore the genome of *D. subobscura*.

4.3 Insights and future perspectives into inversion polymorphisms in *Drosophila subobscura*

Almost a hundred years after the early studies of Sturtevant and Dobzhansky, a number of major questions about inversion polymorphisms remain unresolved. The past few decades much effort has been put into understanding how inversions can impact evolutionary change. Currently, evolutionary geneticists around the world are investigating polymorphic inversions using different species, ranging from plants to mammals. In the chapters above we have attempted to generate the first high-quality genome assembly for *D. subobscura* and provide a detailed analysis and annotation of its genome in an effort to further establish the *subobscura* subgroup as a model for comparative genomics, and *D. subobscura* as a model organism for the study of polymorphic inversions. The resources developed can facilitate the characterization of all the polymorphic inversions in the species, help us unravel the mechanisms of formation of different polymorphic inversions and identify target loci and genes that might be under selections within the different chromosomal arrangements. Moreover, the acquisition of a high-quality genome can vastly improve comparative genomics, since fragmented genomes are not best suited for comparative analyses, while it can assist in identifying inversion breakpoints with great precision. The *D. subobscura* genome and developed resources can be the stepping stone to better understand inversion polymorphism and its maintenance in the species. Future work could help disentangle the evolutionary mechanisms that maintain inversion polymorphisms in *D. subobscura* and further explore and interpret its long-known systematic spatiotemporal patterns.

Up-to-date, major progress has been noted in illustrating that several polymorphic inversions in different species are shaped by selection. Yet, the types of selection that maintain and spread polymorphic inversions in populations warrants further investigation (Kirkpatrick & Kern, 2012). Here we propose that population genomics analyses using *D. subobscura* could help identify genes that serve as targets of selection and the distinct signatures of different types of section acting on polymorphic inversions. Further, the abundance of polymorphic inversions in the species and its wide geographic range could help explore how dominance or epistatic effects of loci captured within inversions may influence adaptation to diverse environments.

5. Conclusions

1. We present the first high-quality, long-read sequencing, *D. subobscura* reference genome.
2. We demonstrated that the sequenced genome exhibits a relatively compact size, compared to other assembled genomes of the *obscura* group.
3. *D. subobscura* exhibits the highest rate of accumulation of paracentric inversions of its subgroup.
4. All identified inversions originated by chromosomal breakage, supporting that the prevailing mechanism of inversion formation in the *Sophophora* subgenus of *Drosophila* is NHEJ.
5. Inversion fixation rates appear to be 10 times higher in continental *D. subobscura* than in the two island species, *D. guanche* and *D. madeirensis*.
6. Genome structure evolution in *D. subobscura* is driven indirectly, through the inversions' recombination-suppression effects in maintaining sets of adaptive alleles together in the face of gene flow.
7. We have assembled and annotated a high-quality genome for an $O_{\underline{3+4}+7}$ isogenic line.
8. We have isolated and functionally characterized the O_7 breakpoints.
9. Our findings have general implications for current theories on the molecular mechanisms of formation of inversions.
10. Inversion O_7 spatiotemporal clines are driven by selective factors other than temperature alone.
11. Inversion O_7 may have altered *D. subobscura* immunometabolism, by disrupting the concerted evolution of two *AttA2* immunity genes, and reattached them to putative dFOXO metabolic enhancers.
12. Considering the length of O_7 , it is likely that its evolution is shaped by additional direct or/and indirect effects on genes other than those near its breakpoints.

6. Publications from this Thesis

The results presented in “Chapter 1: Long-read based assembly and synteny analysis of a reference *Drosophila subobscura* genome reveals signatures of structural evolution driven by inversions recombination-suppression effects ” have been published in *BMC Genomics* **20**, 223 (2019).

The results presented in “Chapter 2: The Cyclically Seasonal *Drosophila subobscura* Inversion O₇ Originated From Fragile Genomic Sites and Relocated Immunity and Metabolic Genes ” have been published in *Front. Genet.* **11**, 565836 (2020).

Appendices

A. Supplementary material of “Long-read based assembly and synteny analysis of a reference *Drosophila subobscura* genome reveals signatures of structural evolution driven by inversions recombination-suppression effects”

i. Supplementary Figures

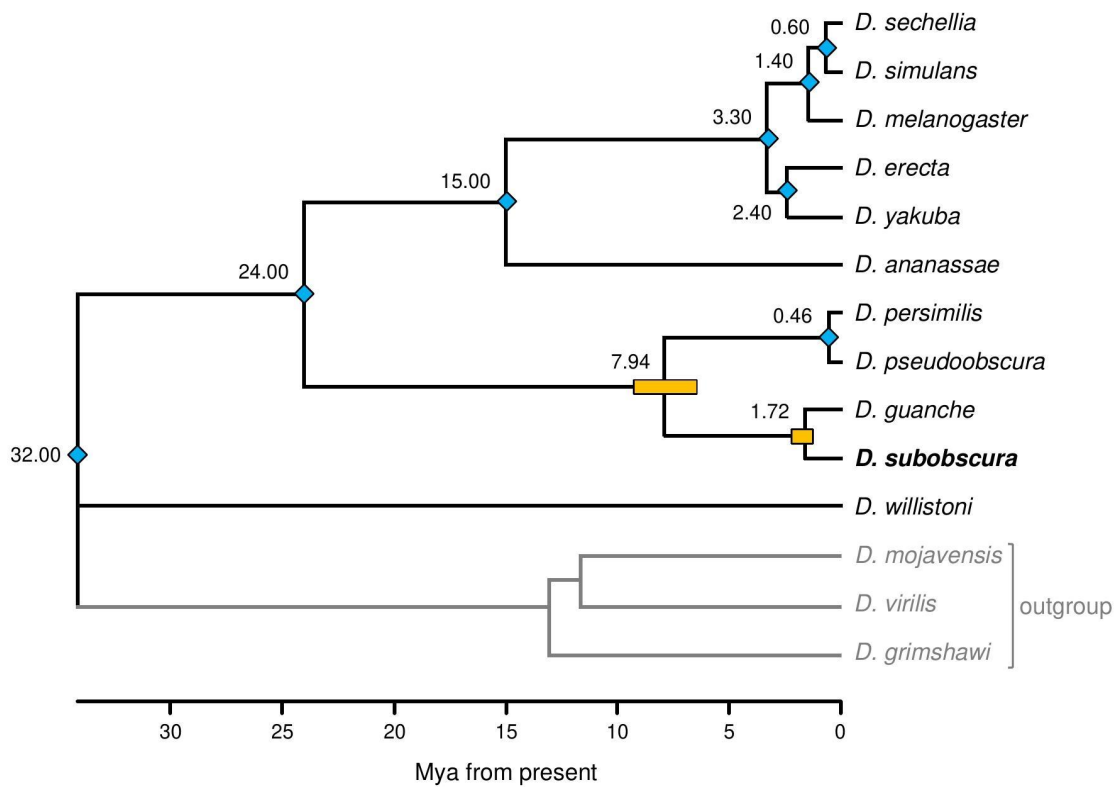


Figure S1. RelTime timetree of 14 *Drosophila* species obtained using the maximum-likelihood tree-topology that results after GTR + G + I best-fit modeling of a 50 concatenated nuclear low-codon bias orthologous gene alignment dataset. Blue diamonds indicate Obbard *et al.* mutation-based calibrated nodes, and orange boxes 95% confidence intervals for target divergences.

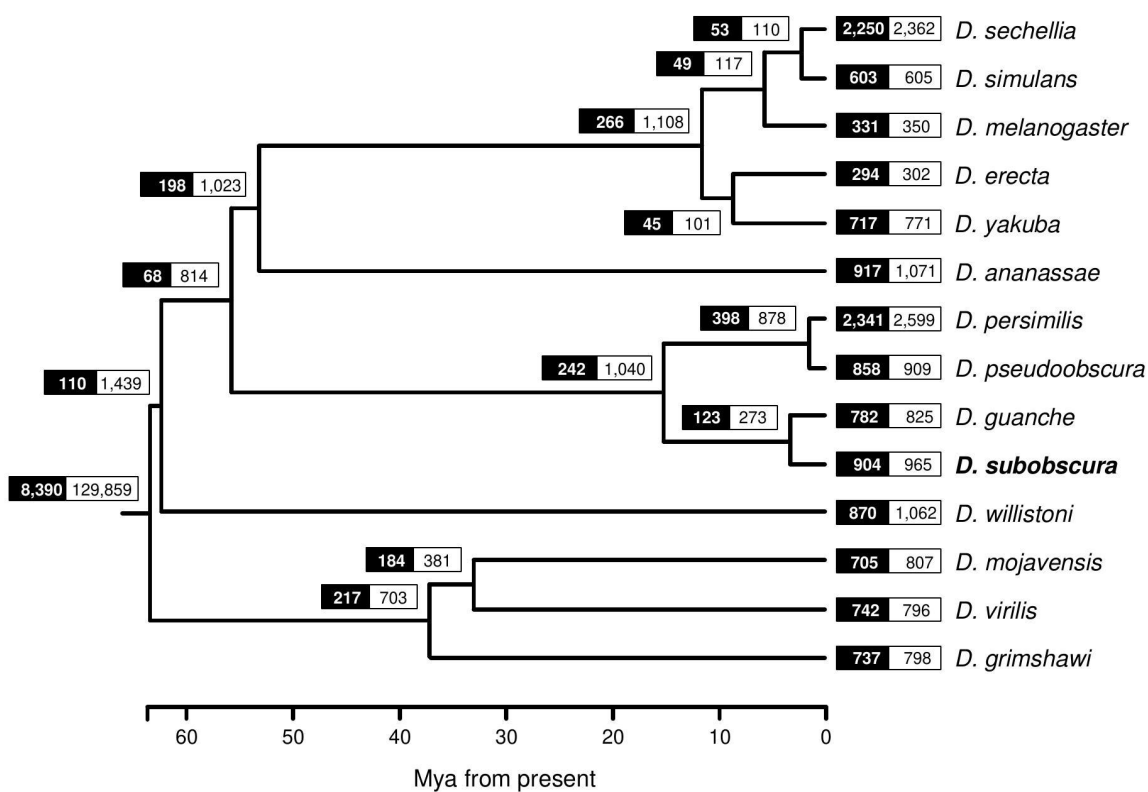


Figure S2. OrthoMCL analysis of gene families in *D. subobscura*. Numbers of orthoMCL clusters and of genes within those clusters on each node are given in black and white rectangles, respectively.

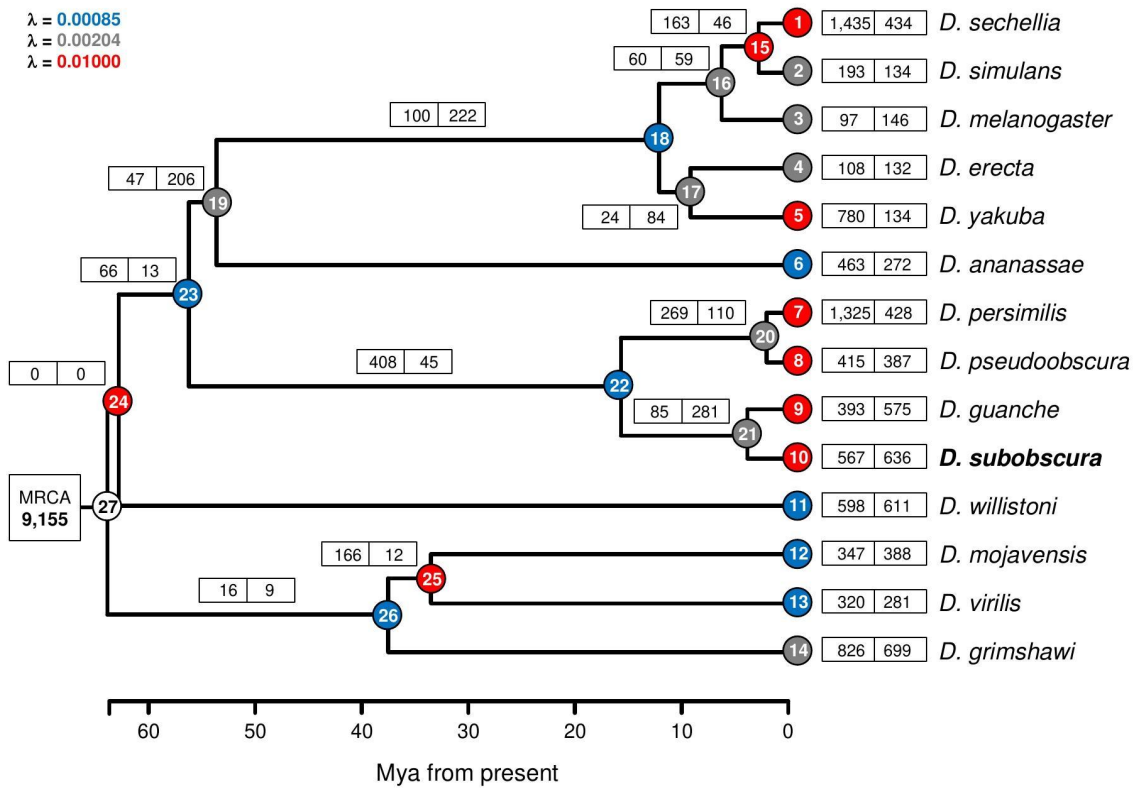


Figure S3. CAFE analysis of the evolution of gene family size in *D. subobscura*. Shown on each branch are its corresponding numbers of expanded (left) and contracted (right) gene families. Circled numbers on nodes are identifiers for internal branches of the phylogeny leading to those nodes. The colors of the circles indicate estimated rates of gene gain and loss according to the legend on the upper left (blue: slow, grey: medium, red: fast).

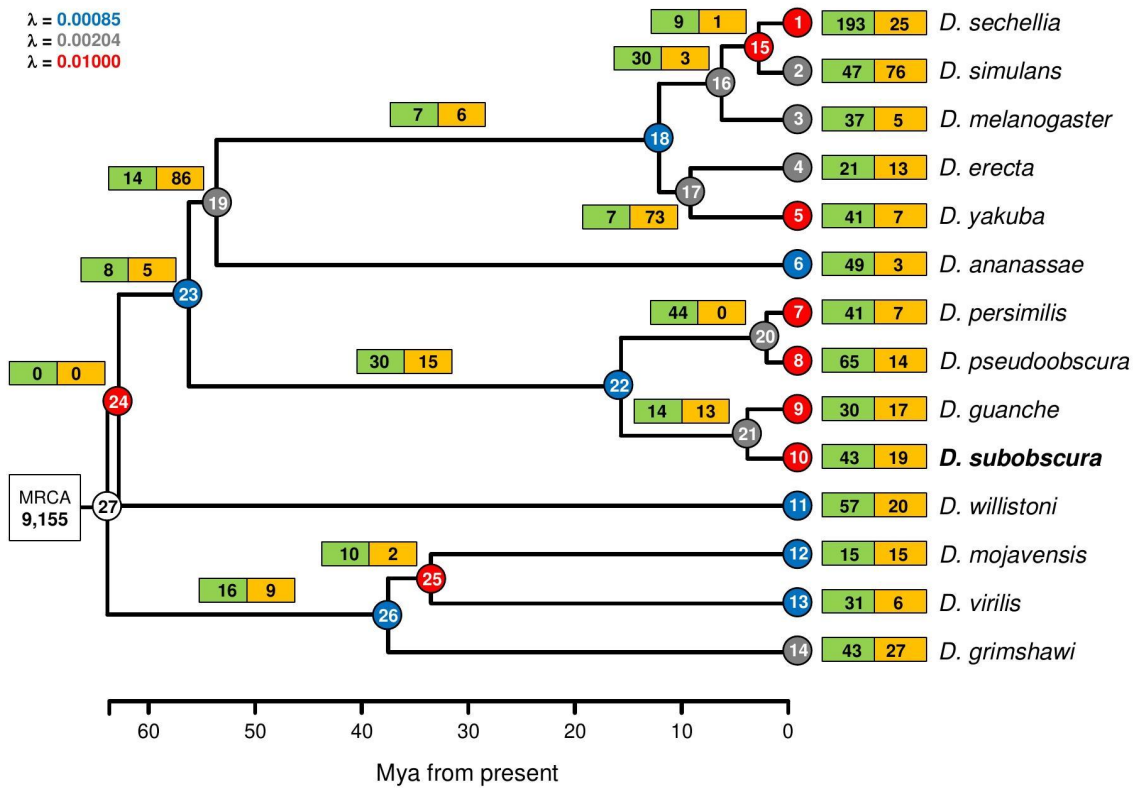


Figure S4. CAFE analysis of the evolution of gene family size in *D. subobscura*. Shown on each branch are its corresponding numbers of significantly expanded (green) and contracted (orange) gene families. Circled numbers on nodes are identifiers for internal branches of the phylogeny leading to those nodes. The colors of the circles indicate estimated rates of gene gain and loss according to the legend on the upper left (blue: slow, grey: medium, red: fast).

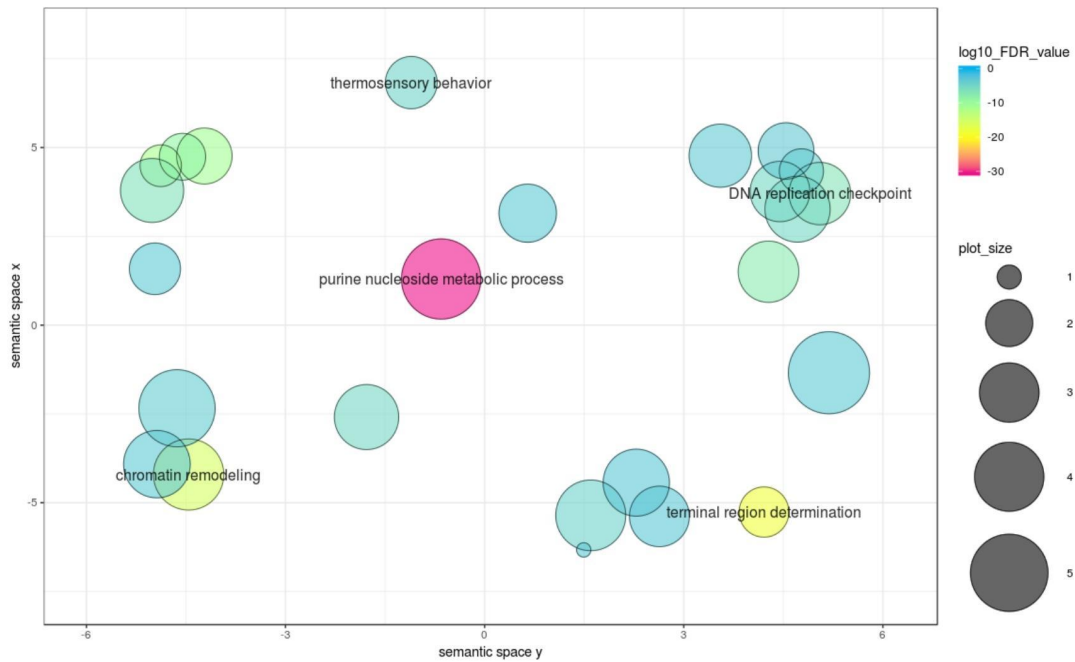


Figure S5. REVIGO summary scatterplot for 27 over-represented Biological Process GO terms in CAFE-expanded gene families. Shown GO term names denote cluster representatives centered on their corresponding GO term. Distances between GO terms are in units of semantic similarity. Circle color indicates FDR values, and circle size generality of the GO term (the lower, the greater the uniqueness of the term).

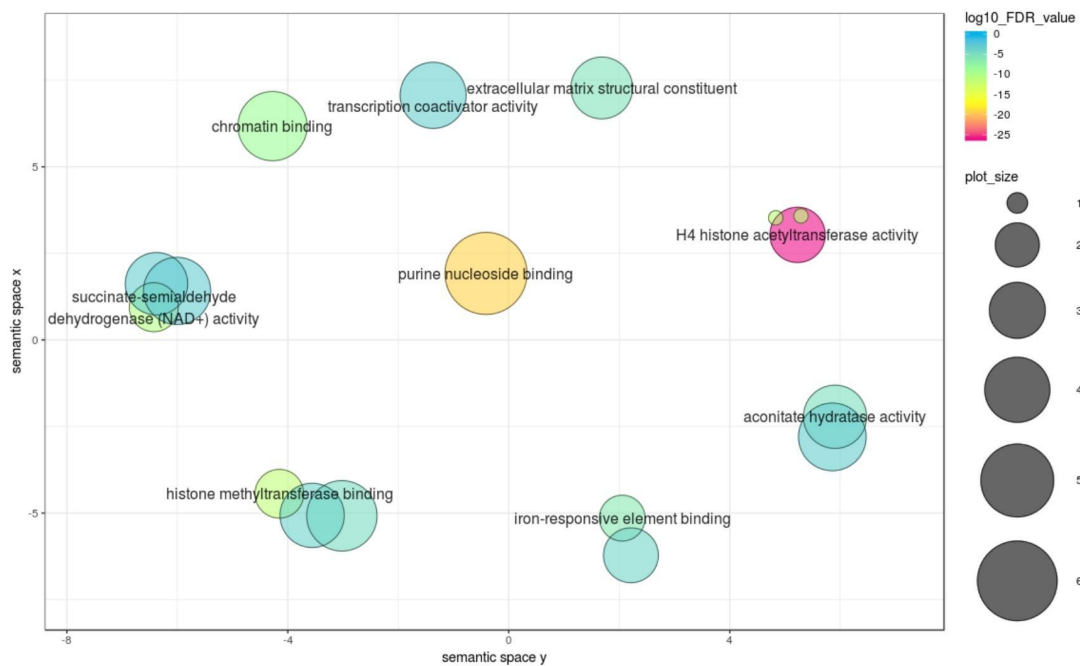


Figure S6. REVIGO summary scatterplot for 17 over-represented Molecular Function GO terms in CAFE-expanded gene families. Shown GO term names denote cluster representatives centered on their corresponding GO term. Distances between GO terms are in units of semantic similarity. Circle color indicates FDR values, and circle size generality of the GO term (the lower, the greater the uniqueness of the term).

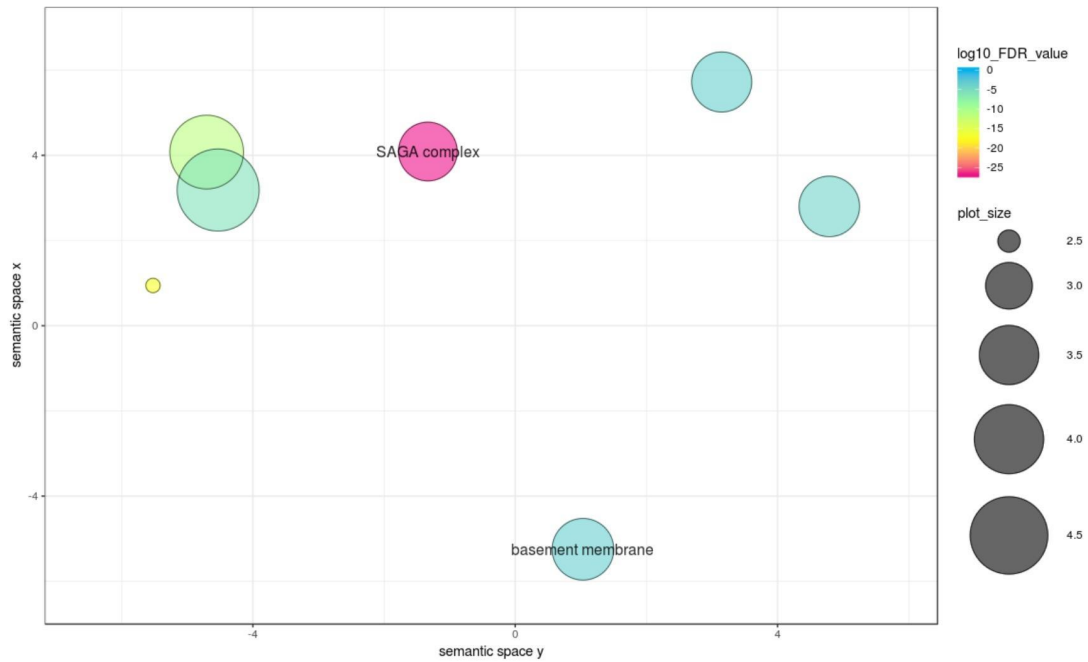


Figure S7. REVIGO summary scatterplot for 9 over-represented Cellular Component GO terms in CAFE-expanded gene families. Shown GO term names denote cluster representatives centered on their corresponding GO term. Distances between GO terms are in units of semantic similarity. Circle color indicates FDR values, and circle size generality of the GO term (the lower, the greater the uniqueness of the term).

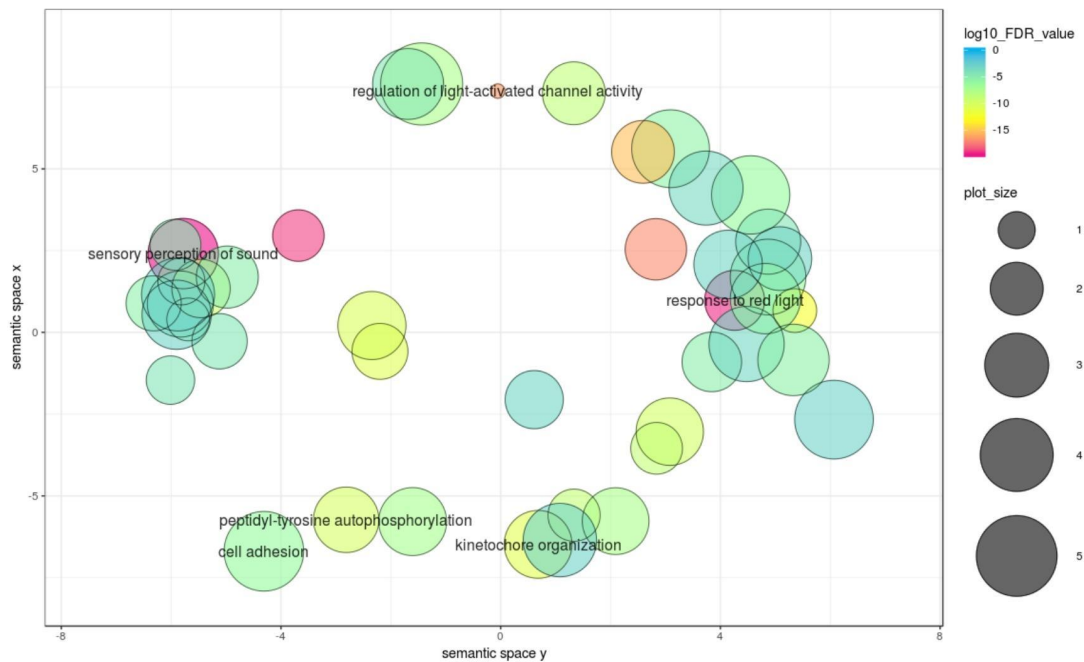


Figure S8. REVIGO summary scatterplot for 51 over-represented Biological Process GO terms in CAFE-contracted gene families. Shown GO term names denote cluster representatives centered on their corresponding GO term. Distances between GO terms are in units of semantic similarity. Circle color indicates FDR values, and circle size generality of the GO term (the lower, the greater the uniqueness of the term).

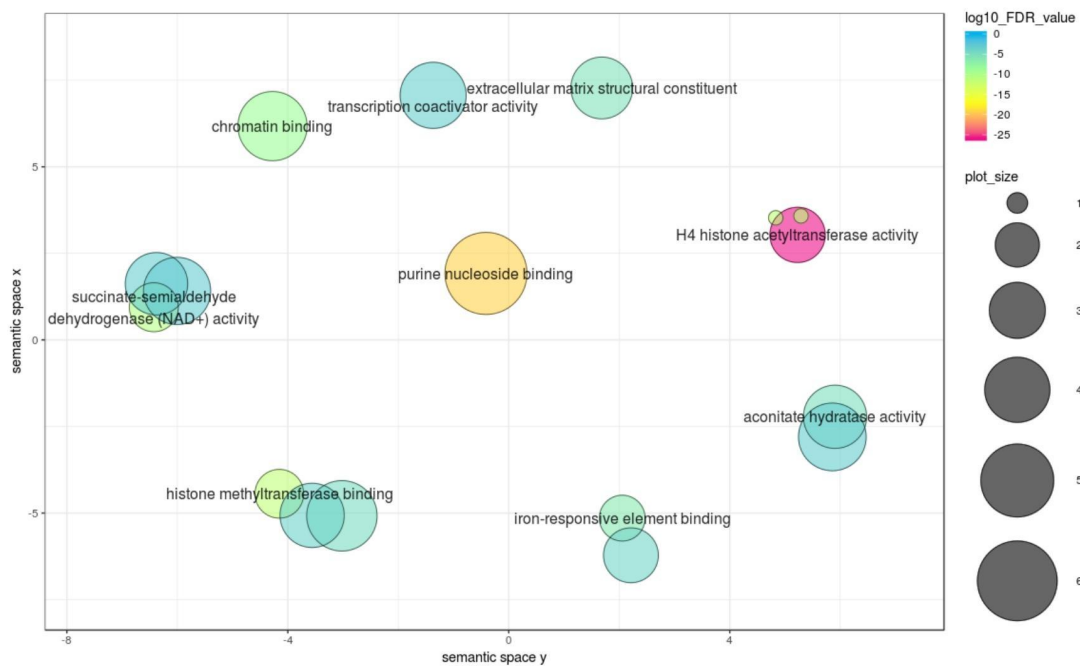


Figure S9. REVIGO summary scatterplot for 12 over-represented Molecular Function GO terms in CAFE-contracted gene families. Shown GO term names denote cluster representatives centered on their corresponding GO term. Distances between GO terms are in units of semantic similarity. Circle color indicates FDR values, and circle size generality of the GO term (the lower, the greater the uniqueness of the term).

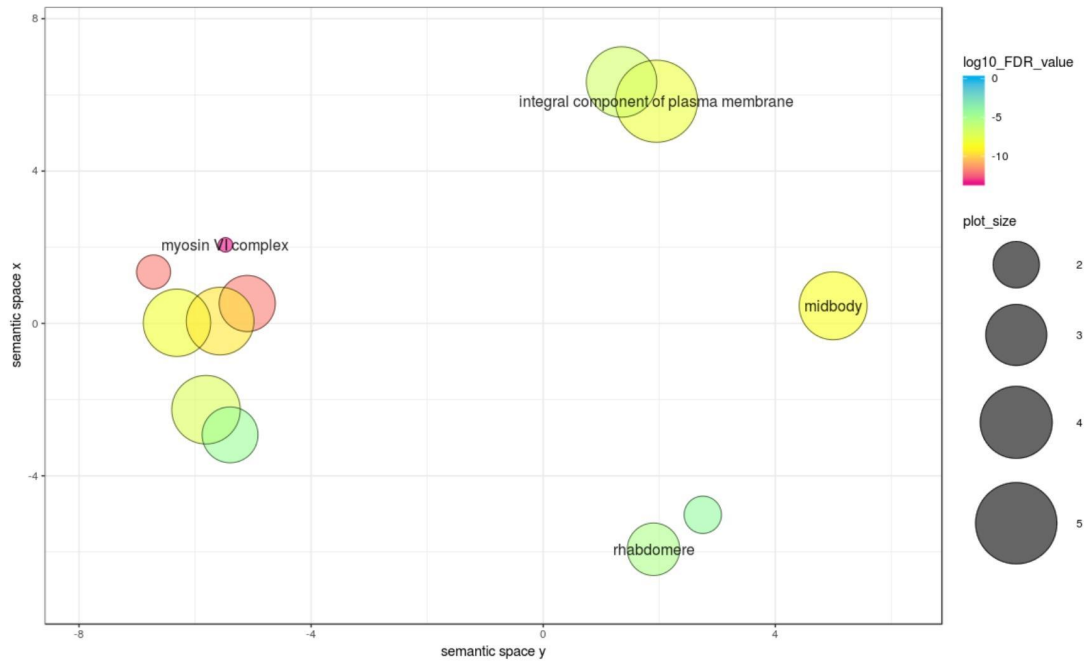


Figure S10. REVIGO summary scatterplot for 8 over-represented Cellular Component GO terms in CAFE-contracted gene families. Shown GO term names denote cluster representatives centered on their corresponding GO term. Distances between GO terms are in units of semantic similarity. Circle color indicates FDR values, and circle size generality of the GO term (the lower, the greater the uniqueness of the term).

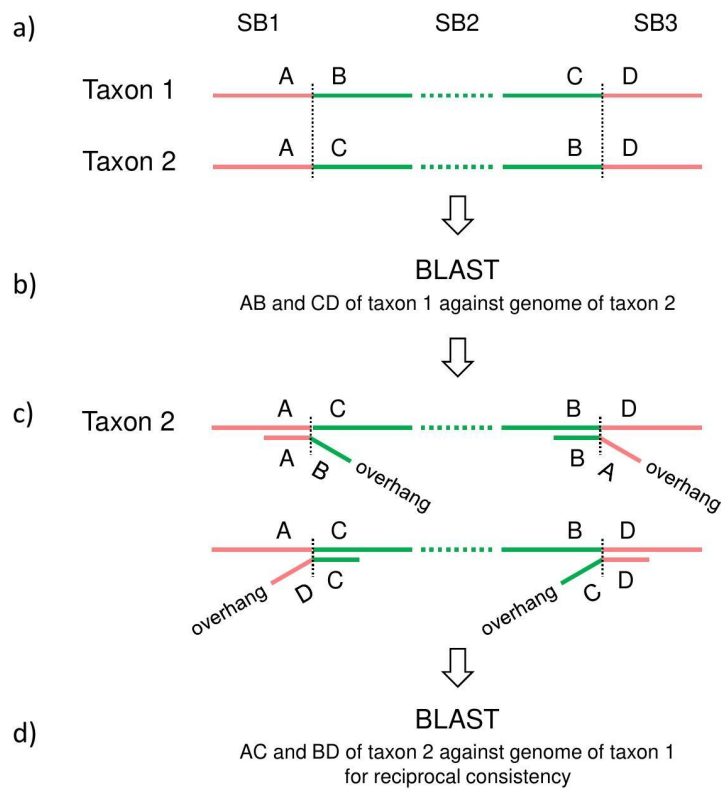


Figure S11. Schematic of the strategy used for inversion breakpoint detection. From top to bottom: shown are (a) two noninverted (SB1 and SB3; pink) and one inverted (SB2; green) hypothetical SyMAP synteny blocks between two taxa (1 and 2). The regions flanking the points of broken synteny (vertical dotted lines) are labelled A-D correspondingly; (b) BLASTing regions AB and CD from taxon 1 against the genome of taxon 2 each produces two hits (c) at opposite ends of the inverted synteny block with associated overhangs; (d) steps b-c are repeated using taxon 2 for the BLAST queries to test for reciprocal consistency (see main text for more detail).

ii. Supplementary Tables

Table S1. Genome sequencing and assembly statistics (coverages based on a 150Mb genome size; lengths are in bp).

PacBio sequencing statistics	
Reads	1,252,701
Avg. read length	8,003
Longest read	52,567
Bases	10,025,366,103
Bases per SMRT cell	1,432,195,158
Genome coverage	66.8×
Canu assembly statistics	
Reads > 1kb	1,183,935
Bases in reads > 1kb	9,990,943,511
Corrected reads	939,323
Bases in corrected reads	7,881,319,324
Trimmed reads	1,060,943
Bases in trimmed reads	6,474,536,519
Avg. trimmed read length	6,103
Longest trimmed read	32,866
Genome coverage	43.2×
Contigs	327
Bases in contigs	135,636,516
Final assembly statistics	
Contigs	212
Bases in contigs	129,182,992
Scaffolds	186
Bases in scaffolds	129,236,726
L50	7
N50	5,954,457
Unknown scaffolds	33
Bases in unknown scaffolds	823,561
Chromosome-assigned scaffolds	153
Bases in chromosome-assigned scaffolds	128,413,165
Assigned-only scaffolds	90
Bases in assigned-only scaffolds	3,551,346
rDNA scaffolds	19
Bases in rDNA scaffolds	862,921
Non rDNA scaffolds	71
Bases in non rDNA scaffolds	2,688,425
Pseudochromosome-scaffolds	63
Bases in pseudochromosome-scaffolds	124,861,819

Table S2. Genetic markers used for validation, and physical anchoring, ordering and orientation of scaffolds. In total, 683 markers were considered, of which 621 were used. 62 markers were not used because they showed inconsistencies as to their localization with respect to markers from other studies and our own data. The MS Excel file contains eight spreadsheets, including one for this title, one for each of the five major pseudochromosomes (*i.e.*, A, J, U, E and O), one with a summary, and one with the references for the marker data. For each pseudochromosome, markers are listed in column “A”, including used cytological (numbered, black), used linkage (numbered, blue), and nonused (nonnumbered, red) markers. For each marker, information relative to its name, cytological localization, authors, corresponding *D. pseudoobscura* “GA” gene model name, inconsistency where it applies, coordinates in the pseudochromosome, scaffold name, scaffold orientation and cytological span, and BLASTn statistics is provided in subsequent columns, from “B” to “Y”. Column “X” provides the number of used marker per scaffold. Alternating color in the background denotes different scaffolds. Cytological coordinates are always relative to the Kunze-Mühl and Müller [12] standard reference map. From the summary spreadsheet, most of the inconsistencies (72%) come from one (Laayouni *et al.*, 2007) out of the total 26 cited works. Excluding that study, the total percent of inconsistencies is only 2.85% (*i.e.*, 17 out of 638 markers).

[Link to supplementary table 2](#)

Table S3. *D. subobscura* mitogenome gene content and order (lengths in bp).

Annotation	Start	Stop	Length	Strand
tRNAI	0	46	47	+
tRNAQ	84	152	69	-
tRNAM	152	220	69	+
ND2	242	1236	995	+
tRNAW	1243	1308	66	+
tRNAC	1301	1363	63	-
tRNAY	1364	1429	66	-
CoI	1434	2942	1509	+
tRNAL2	2966	3031	66	+
CoII	3037	3706	672	+
tRNAK	3723	3792	70	+
tRNAD	3793	3859	67	+
ATP8	3886	4017	132	+
ATP6	4012	4676	665	+
CoIII	4693	5470	777	+
tRNAG	5490	5553	64	+
ND3	5554	5880	327	+
tRNAA	5906	5969	64	+
tRNAR	5971	6033	63	+

tRNAN	6034	6098	65	+
tRNAS1	6099	6166	68	+
tRNAE	6167	6232	66	+
tRNAF	6251	6316	66	-
ND5	6336	7955	1620	-
tRNAH	8052	8117	66	-
ND4	8138	9457	1320	-
ND4L	9454	9714	261	-
tRNAT	9750	9814	65	+
tRNAP	9815	9880	66	-
ND6	9895	10398	504	+
COB	10411	11517	1107	+
tRNAS2	11550	11615	66	+
ND1	11638	12561	924	-
tRNAL1	12581	12645	65	-
rRNAL	12604	13979	1376	-
tRNAV	13966	14037	72	-
rRNAS	14037	14820	784	-

Table S4. Repetitive content of the *D. subobscura* genome.

Class	No. of copies	Length (bp)	% of genome ¹
Retrotransposon			
SINE	168	14,783	0.01%
LINE	9,080	2,760,714	2.14%
LTR	7,036	2,317,534	1.79%
DNA TEs	25,953	4,853,394	3.76%
P	3,778	409,099	0.32%
CMC-EnSpm	2,475	264,762	0.20%
Tc1/Mariner	1,190	423,231	0.33%
hAT	3,369	684,490	0.53%
T2/Kolobok	981	93,683	0.07%
Helitrons	6,490	1,817,722	1.41%
Maverick	1,295	392,702	0.30%
Other	6,375	767,705	0.59%
Simple repeat	134,876	5,313,023	4.13%
Low complexity	15,271	723,001	0.56%
Satellites			
SGM	5,181	2,298,546	1.78%
Sat290	637	103,534	0.08%
Other	528	42,912	0.03%
Unclassified	341	86,020	0.07%
Other	2	367	0.0003%
Total	199,073	18,531,828	14.34%

1. Percents based on a 129,236,726 bp genome.

Table S5. Optimal CAFE model selection for the evolution of gene family size along the 14 *Drosophila* ultrametric tree in Figures S3-S4. Shown are the four assayed increasingly complex models, including the 1- λ and 3- λ models, and the 5- λ model without and with global assembling error term (ϵ); and their corresponding parameter estimates, including global (λ_G), slow (λ_S), medium (λ_M), fast (λ_F), *D. subobscura* (λ_{Ds}) and *D. guanche* (λ_{Dg}) lambdas, and global error, and maximum-likelihood scores (-lnL).

Model	λ_G	λ_S	λ_M	λ_F	λ_{Ds}	λ_{Dg}	ϵ	-lnL
1- λ	0.0027							103,172.40
3- λ		0.0009	0.0024	0.0218				88,126.00
5- λ		0.0009	0.0024	0.0216	0.0257	0.0191		88,104.66
5- $\lambda + \epsilon$		0.0007	0.0018	0.0124	0.0197	0.0112	0.0747	88,008.67

Table S6. Over represented GO Terms among CAFE significantly expanded gene families in *D. subobscura* inferred using one-sided Fisher exact test (FDR < 0.001) implemented in Blast2Go (BP: Biological Process; MF: Molecular Function; CC: Cellular Component).

GO ID	GO name	GO category	FDR
GO:0001883	Purine nucleoside binding	MF	4.55E-20
GO:0001745	Compound eye morphogenesis	BP	5.42E-04
GO:0042278	Purine nucleoside metabolic process	BP	3.58E-31
GO:0006338	Chromatin remodeling	BP	3.20E-16
GO:0005703	Polytene chromosome puff	CC	7.10E-18
GO:0003682	Chromatin binding	MF	1.58E-09
GO:0000124	SAGA complex	CC	1.50E-27
GO:0008134	Transcription factor binding	MF	1.59E-05
GO:0005730	Nucleolus	CC	1.46E-06
GO:0007362	Terminal region determination	BP	4.95E-19
GO:0010485	H4 histone acetyltransferase activity	MF	1.52E-26
GO:0005671	Ada2/Gcn5/Ada3 transcription activator complex	CC	2.63E-22
GO:0048515	Spermatid differentiation	BP	5.39E-06
GO:0016604	Nuclear body	CC	6.34E-12
GO:0007478	Leg disc morphogenesis	BP	4.46E-04
GO:0060828	Regulation of canonical Wnt signaling pathway	BP	9.00E-04
GO:1901605	Alpha-amino acid metabolic process	BP	7.60E-04
GO:0051568	Histone H3-K4 methylation	BP	2.64E-07
GO:0005201	Extracellular matrix structural constituent	MF	2.50E-06
GO:0008347	Glial cell migration	BP	7.84E-06
GO:0006352	DNA-templated transcription, initiation	BP	2.34E-04
GO:0003713	Transcription coactivator activity	MF	3.96E-04
GO:0032968	Positive regulation of transcription elongation from RNA polymerase II promoter	BP	4.78E-06
GO:0040040	Thermosensory behavior	BP	1.96E-05
GO:0048864	Stem cell development	BP	4.03E-04
GO:0043971	Histone H3-K18 acetylation	BP	4.44E-12

GO:1990226	Histone methyltransferase binding	MF	4.44E-12
GO:0043993	Histone acetyltransferase activity (H3-K18 specific)	MF	4.44E-12
GO:0044017	Histone acetyltransferase activity (H3-K27 specific)	MF	4.44E-12
GO:0043982	Histone H4-K8 acetylation	BP	1.60E-11
GO:0043983	Histone H4-K12 acetylation	BP	4.89E-11
GO:0043974	Histone H3-K27 acetylation	BP	1.41E-09
GO:0032922	Circadian regulation of gene expression	BP	2.42E-08
GO:0000076	DNA replication checkpoint	BP	1.35E-07
GO:0035023	Regulation of Rho protein signal transduction	BP	4.71E-05
GO:0007464	R3/R4 cell fate commitment	BP	8.61E-05
GO:0004777	Succinate-semialdehyde dehydrogenase (NAD+) activity	MF	3.56E-11
GO:0005604	Basement membrane	CC	3.48E-04
GO:0008266	Poly(U) RNA binding	MF	8.22E-05
GO:0005844	Polysome	CC	8.22E-05
GO:0005125	Cytokine activity	MF	1.57E-04
GO:0046426	Negative regulation of JAK-STAT cascade	BP	9.57E-04
GO:0030350	Iron-responsive element binding	MF	4.52E-07
GO:0003994	Aconitate hydratase activity	MF	7.55E-06
GO:0045252	Oxoglutarate dehydrogenase complex	CC	2.78E-04
GO:0016624	Oxidoreductase activity, acting on the aldehyde or oxo group of donors, disulfide as acceptor	MF	7.49E-04
GO:0006750	Glutathione biosynthetic process	BP	7.18E-04
GO:0016846	Carbon-sulfur lyase activity	MF	7.18E-04
GO:1900026	Positive regulation of substrate adhesion-dependent cell spreading	BP	5.93E-04
GO:0035386	Regulation of Roundabout signaling pathway	BP	5.93E-04
GO:0070899	Mitochondrial tRNA wobble uridine modification	BP	5.93E-04
GO:0004029	Aldehyde dehydrogenase (NAD) activity	MF	5.93E-04

Table S7. Over represented GO Terms among CAFE significantly contracted gene families in *D. subobscura* inferred using one-sided Fisher exact test (FDR < 0.001) implemented in Blast2Go (BP: Biological Process; MF: Molecular Function; CC: Cellular Component).

GO ID	GO name	GO category	FDR
GO:0007605	Sensory perception of sound	BP	2.61E-20
GO:0010114	Response to red light	BP	4.58E-20
GO:0060086	Circadian temperature homeostasis	BP	1.08E-19
GO:0004714	Transmembrane receptor protein tyrosine kinase activity	MF	8.37E-18
GO:0043153	Entrainment of circadian clock by photoperiod	BP	1.04E-17
GO:0035271	Ring gland development	BP	1.40E-17
GO:0016061	Regulation of light-activated channel activity	BP	2.62E-17
GO:0031489	Myosin V binding	MF	2.62E-17
GO:2001259	Positive regulation of cation channel activity	BP	4.75E-16
GO:0070855	Myosin VI head/neck binding	MF	1.15E-15
GO:0004715	Non-membrane spanning protein tyrosine kinase activity	MF	1.15E-15
GO:0031476	Myosin VI complex	CC	4.14E-14
GO:0031475	Myosin V complex	CC	1.64E-13
GO:0016060	Metarhodopsin inactivation	BP	5.09E-13
GO:0097431	Mitotic spindle pole	CC	1.13E-12
GO:0070865	Investment cone	CC	1.13E-12
GO:0051383	Kinetochore organization	BP	1.39E-11
GO:0016062	Adaptation of rhodopsin mediated signaling	BP	1.39E-11
GO:0072499	Photoreceptor cell axon guidance	BP	2.18E-11
GO:0010977	Negative regulation of neuron projection development	BP	4.77E-11
GO:0038083	Peptidyl-tyrosine autophosphorylation	BP	1.25E-10
GO:0031935	Regulation of chromatin silencing	BP	1.65E-10
GO:0005876	Spindle microtubule	CC	3.26E-10
GO:0001752	Compound eye photoreceptor fate commitment	BP	6.67E-10
GO:0003705	Transcription factor activity, RNA polymerase II distal enhancer sequence-specific binding	MF	9.93E-10

GO:0010705	Meiotic DNA double-strand break processing involved in reciprocal meiotic recombination	BP	1.46E-09
GO:0010780	Meiotic DNA double-strand break formation involved in reciprocal meiotic recombination	BP	1.46E-09
GO:0046716	Muscle cell cellular homeostasis	BP	2.30E-09
GO:0030496	Midbody	CC	2.47E-09
GO:0008514	Organic anion transmembrane transporter activity	MF	3.20E-09
GO:0005814	Centriole	CC	5.77E-09
GO:0005887	Integral component of plasma membrane	CC	9.94E-09
GO:0045316	Negative regulation of compound eye photoreceptor development	BP	1.14E-08
GO:0007099	Centriole replication	BP	3.37E-08
GO:0015711	Organic anion transport	BP	3.37E-08
GO:0042052	Rhabdomere development	BP	3.38E-08
GO:0043035	Chromatin insulator sequence binding	MF	3.50E-08
GO:0030048	Actin filament-based movement	BP	5.56E-08
GO:0070868	Heterochromatin organization involved in chromatin silencing	BP	7.02E-08
GO:0000792	Heterochromatin	CC	8.11E-08
GO:0031234	Extrinsic component of cytoplasmic side of plasma membrane	CC	2.32E-07
GO:0007169	Transmembrane receptor protein tyrosine kinase signaling pathway	BP	5.76E-07
GO:0045944	Positive regulation of transcription by RNA polymerase II	BP	6.38E-07
GO:0007155	Cell adhesion	BP	6.73E-07
GO:0061332	Malpighian tubule bud morphogenesis	BP	1.07E-06
GO:0042127	Regulation of cell proliferation	BP	2.21E-06
GO:0030178	Negative regulation of Wnt signaling pathway	BP	3.43E-06
GO:0045931	Positive regulation of mitotic cell cycle	BP	3.97E-06
GO:0007390	Germ-band shortening	BP	4.22E-06
GO:0005813	Centrosome	CC	4.37E-06
GO:0005326	Neurotransmitter transporter activity	MF	5.43E-06
GO:0035071	Salivary gland cell autophagic cell death	BP	6.35E-06
GO:0016028	Rhabdomere	CC	6.83E-06
GO:0035075	Response to ecdysone	BP	7.24E-06
GO:0046960	Sensitization	BP	1.75E-05
GO:0015695	Organic cation transport	BP	2.15E-05
GO:0007485	Imaginal disc-derived male genitalia development	BP	2.82E-05
GO:0000788	Nuclear nucleosome	CC	2.82E-05
GO:0035074	Pupation	BP	3.38E-05
GO:0007402	Ganglion mother cell fate determination	BP	3.38E-05
GO:0090303	Positive regulation of wound healing	BP	3.38E-05
GO:0035230	Cytoneme	CC	6.32E-05
GO:0007424	Open tracheal system development	BP	6.32E-05
GO:0001078	Proximal promoter DNA-binding transcription repressor activity, RNA polymerase II-specific	MF	9.79E-05
GO:0005509	Calcium ion binding	MF	1.48E-04
GO:0043065	Positive regulation of apoptotic process	BP	2.70E-04
GO:0007476	Imaginal disc-derived wing morphogenesis	BP	3.53E-04
GO:0045087	Innate immune response	BP	4.73E-04
GO:0000166	Nucleotide binding	MF	4.86E-04
GO:0035172	Hemocyte proliferation	BP	6.10E-04
GO:0000187	Activation of MAPK activity	BP	6.34E-04
GO:0000978	RNA pol II proximal promoter sequence-specific DNA binding	MF	6.46E-04
GO:0021579	Medulla oblongata morphogenesis	BP	7.59E-04
GO:1902843	Positive regulation of netrin-activated signaling pathway	BP	7.59E-04
GO:0007267	Cell-cell signaling	BP	8.31E-04
GO:0007517	Muscle organ development	BP	8.70E-04
GO:0006334	Nucleosome assembly	BP	8.80E-04

Table S8. Number of syntenic blocks between *D. subobscura* and increasingly distant relatives.

	A	J	U	E	O	Total
<i>D. subobscura</i> × <i>D. guanche</i>	12	3	4	6	6	31

× <i>D. pseudoobscura</i>	90	66	56	59	62	333
× <i>D. melanogaster</i>	125	100	87	115	113	540

Table S9. Average size of the syntenic block (in Mb) between *D. subobscura* and increasingly distant relatives.

	A	J	U	E	O	Total
<i>D. subobscura</i> × <i>D. guanche</i>	1.364	7.826	4.155	3.578	4.780	3.952
× <i>D. pseudoobscura</i>	0.224	0.333	0.439	0.333	0.462	0.345
× <i>D. melanogaster</i>	0.172	0.226	0.284	0.175	0.264	0.220

Table S10. Synteny analysis of inversion breakpoints. Provided is breakpoint information for 12 inversions, including six from pseudochromosome A (h1, h2, h3, h4, 5 and 6), one from J (ST), two from U (1 and 2), two from E (g1 and ST), and one from O (ms). The MS Excel file contains six spreadsheets, including one for this title, and one for each of the five major pseudochromosomes (*i.e.*, A, J, U, E and O). For each pseudochromosome, inversions are listed in column “A”. For each inversion, information about the three protein coding genes flanking each side of each breakpoint in three species, including *D. melanogaster*, *D. guanche* and *D. subobscura* is provided in subsequent columns, from “B” to “Q”. This information includes species names, names and pseudochromosome coordinates of the three coding gene markers on both sides of each distal and proximal breakpoint, and the size of the pseudochromosome segment spanned by the breakpoints in Mb. Also provided is, for each breakpoint, its cytological and estimated pseudochromosome coordinates, and its hypothetical originating mechanism with the length of the associated duplication where it applies. Cells color background indicate contiguity (brown) or altered (yellow) order of the markers relative to the outgroup (*D. melanogaster/D. pseudoobscura*). For example, in the case of hypothetical inversion 1 of the A chromosome (*i.e.*, h1) in *D. subobscura*, the three markers downstream the proximal breakpoint and upstream the distal breakpoint are in reverse order relative to *D. guanche*, which shows the markers ordered as in *D. melanogaster*. Reciprocal BLASTn searches with each breakpoint did not detect evidence of duplication, suggesting that the most likely originating mechanism of inversion A_{h1} (depicted in yellow) is simple, or nearly straight breaks.

[Link to supplementary Table 10](#)

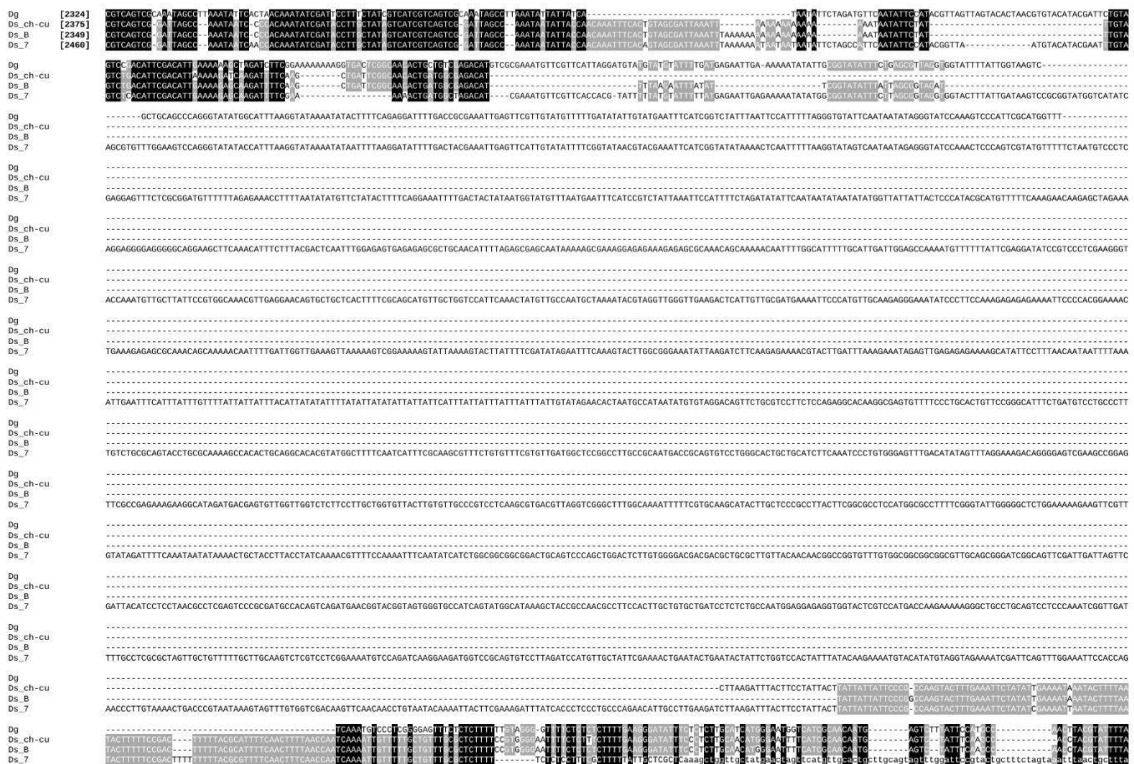
B. Supplementary material of “The cyclically seasonal *Drosophila subobscura* inversion O₇ originated from fragile genomic sites and relocated immunity and metabolic genes”

i. Supplementary Figures

A

Identification of segment A of the proximal breakpoint

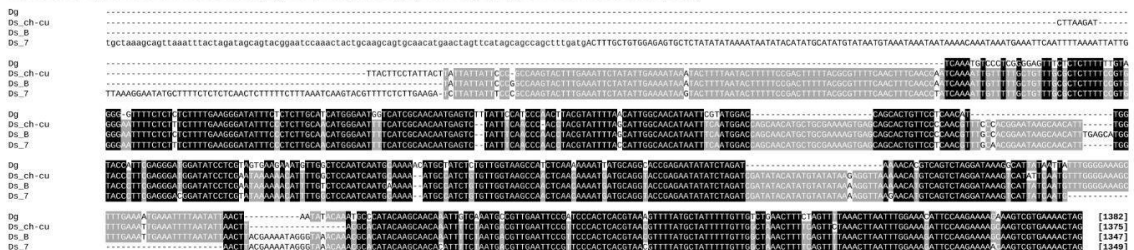
MSA of [+A|B] from the uninverted state (Dg, Ds_ch-cu, and Ds_B) with [+A|C] from the inverted state (Ds_7)



B

Identification of segment B of the proximal breakpoint

MSA of [+A|B] from the uninverted state (Dg, Ds_ch-cu, and Ds_B) with RC[-B|D] from the inverted state (Ds_7)



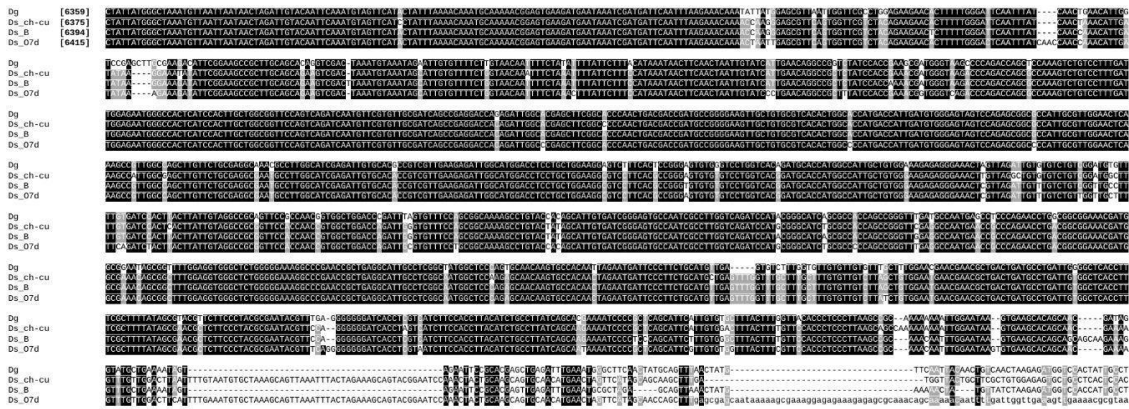
Supplementary Figure S1. Isolation of segments A and B for reconstruction of the proximal breakpoint of O₇. (A) MSA of the [+A|B] region from the uninverted state (Dg, Ds_ch-cu,

and Ds_B) with the [+A|-C] region from O₇. The regions of O₇ corresponding to segments A and C are denoted with capital and lower-case letters, respectively. **(B)** MSA of the [+A|+B] region from the uninverted state (Dg, Ds_ch-cu, and Ds_B) with the reverse complement (RC) of the [-B|+D] region from O₇. The regions of O₇ corresponding to segments D and B are denoted with lower-case and capital letters, respectively. Black and grey backgrounds denote invariant and 75% conserved MSA columns, respectively. Numbers in brackets are basepair distances to the nearest coding sequence.

A

Identification of segment C of the distal breakpoint

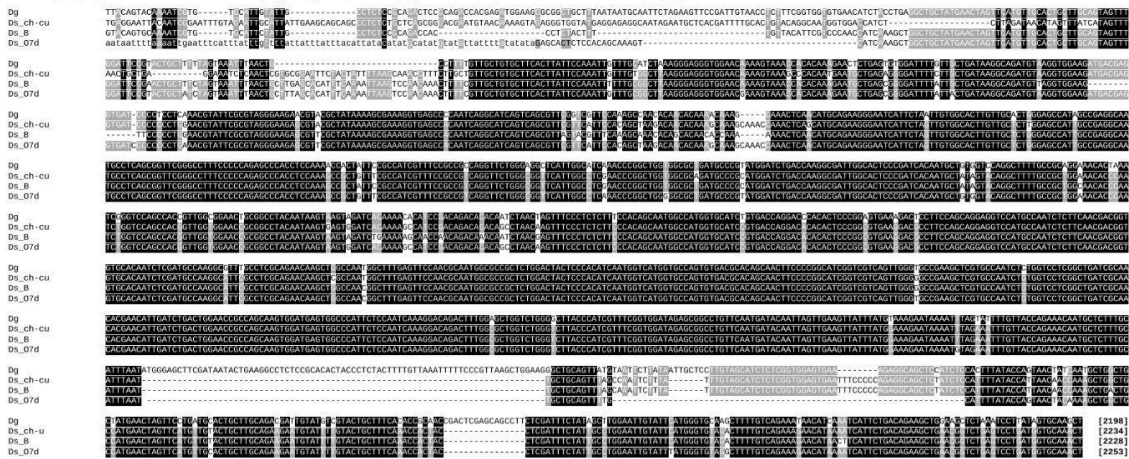
MSA of [+C|+D] from the uninverted state (Dg, Ds_ch-cu, and Ds_B) with RC[+A|-C] from the inverted state (Ds_7)



B

Identification of segment D of the distal breakpoint

MSA of [+C|+D] from the uninverted state (Dg, Ds_ch-cu, and Ds_B) with [-B|+D] from the inverted state (Ds_7)

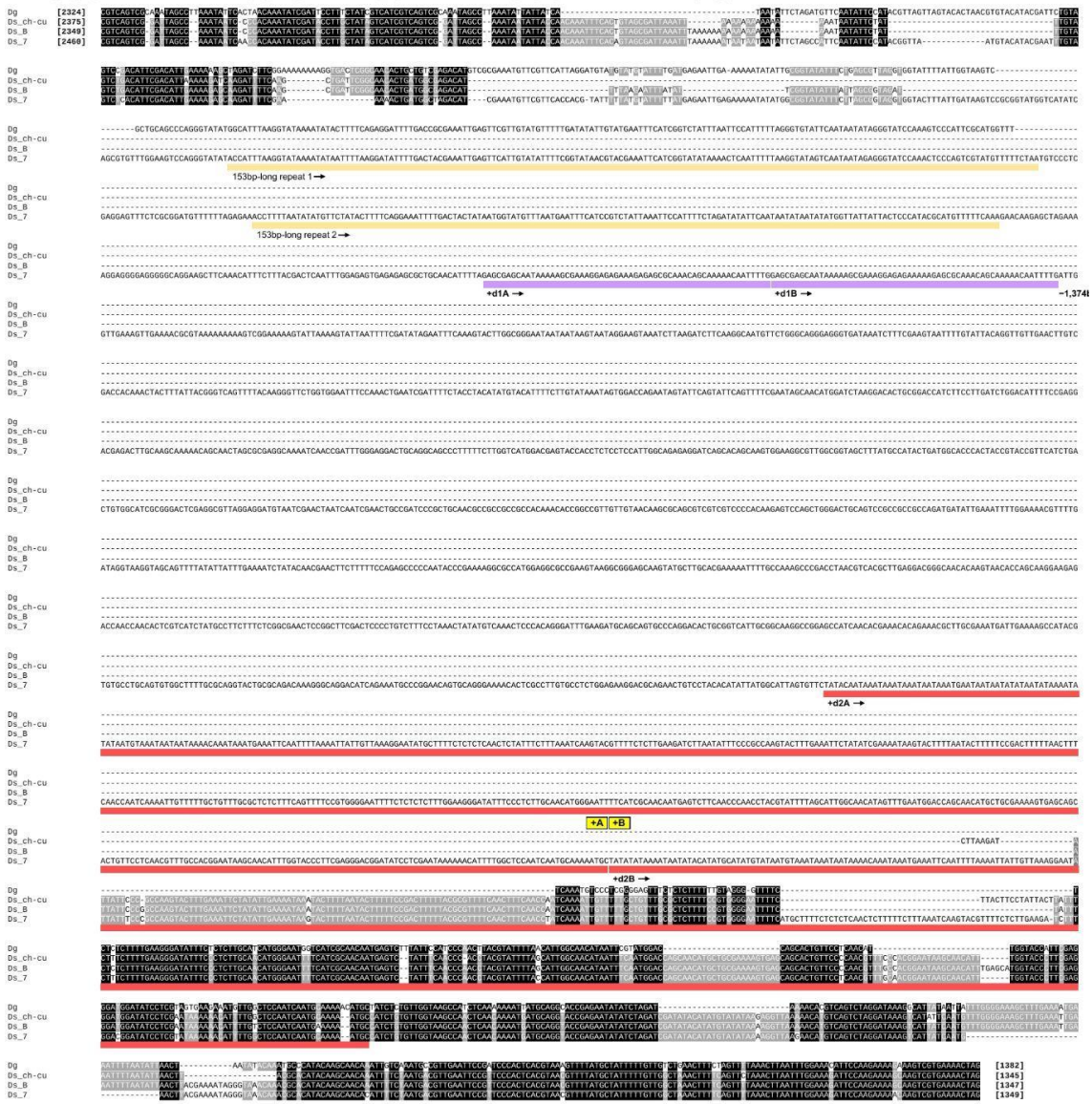


Supplementary Figure S2. Isolation of segments C and D for reconstruction of the distal breakpoint of O₇. **(A)** MSA of the [+C|+D] region from the uninverted state (Dg, Ds_ch-cu, and Ds_B) with the reverse complement (RC) of the [+A|-C] region from O₇. The regions of O₇ corresponding to segments C and A are denoted with capital and lower-case letters, respectively. **(B)** MSA of the [+C|+D] region from the uninverted state (Dg, Ds_ch-cu, and Ds_B) with the reverse complement (RC) of the [-B|+D] region from O₇. The regions of O₇ corresponding to segments B and D are

75% conserved MSA columns, respectively. Numbers in brackets are basepair distances to the nearest coding sequence.

Proximal breakpoint [+A]+B] with the microinversion reversed

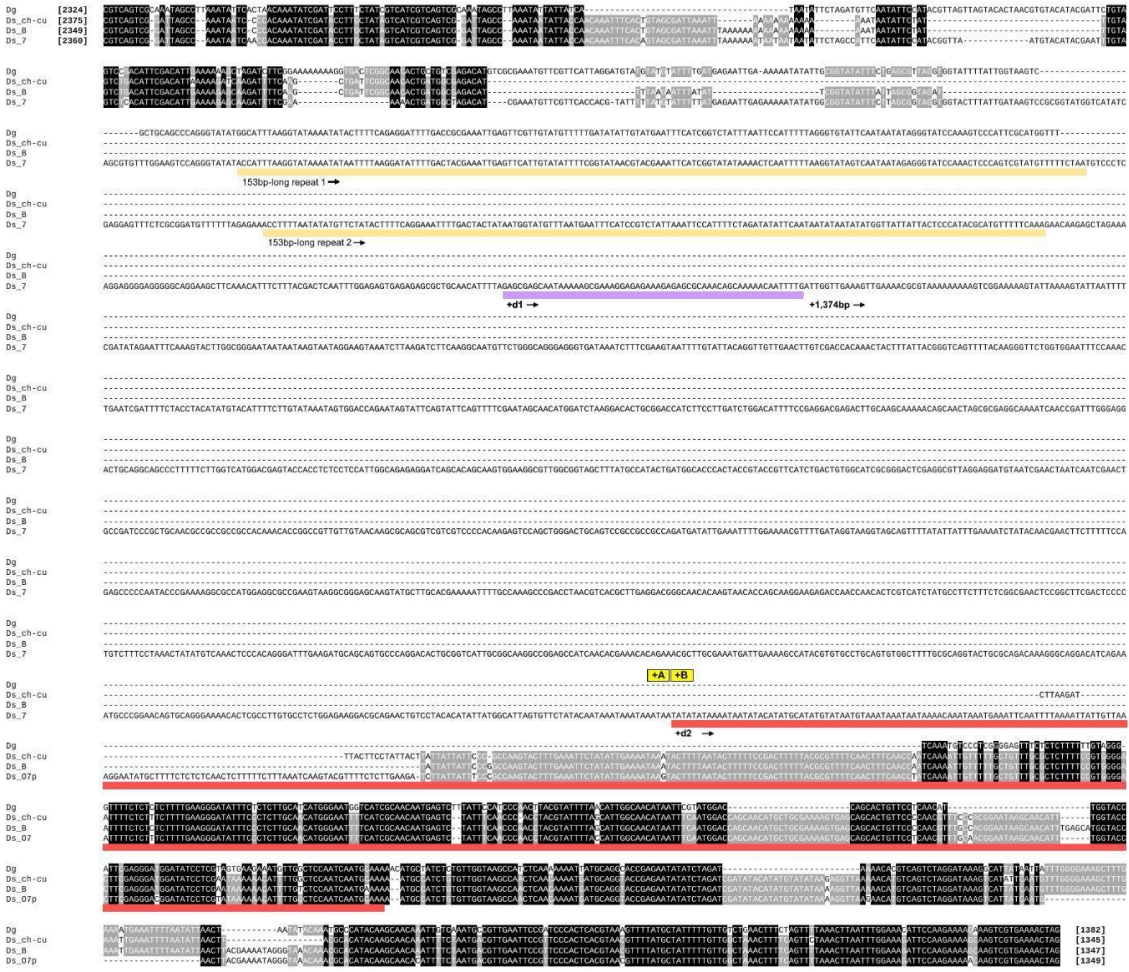
MSA of [+A]+B] from the uninverted state (Dg, Ds_ch-cu, and Ds_B) with reconstructed [+A]+B] from the inverted state (Ds_7)



Supplementary Figure S4. The proximal breakpoint after undoing the microinversion. Similar to Supplementary Figure S3, but with the microinversion reversed.

Proximal breakpoint [+A]+B as before the origination of O₇

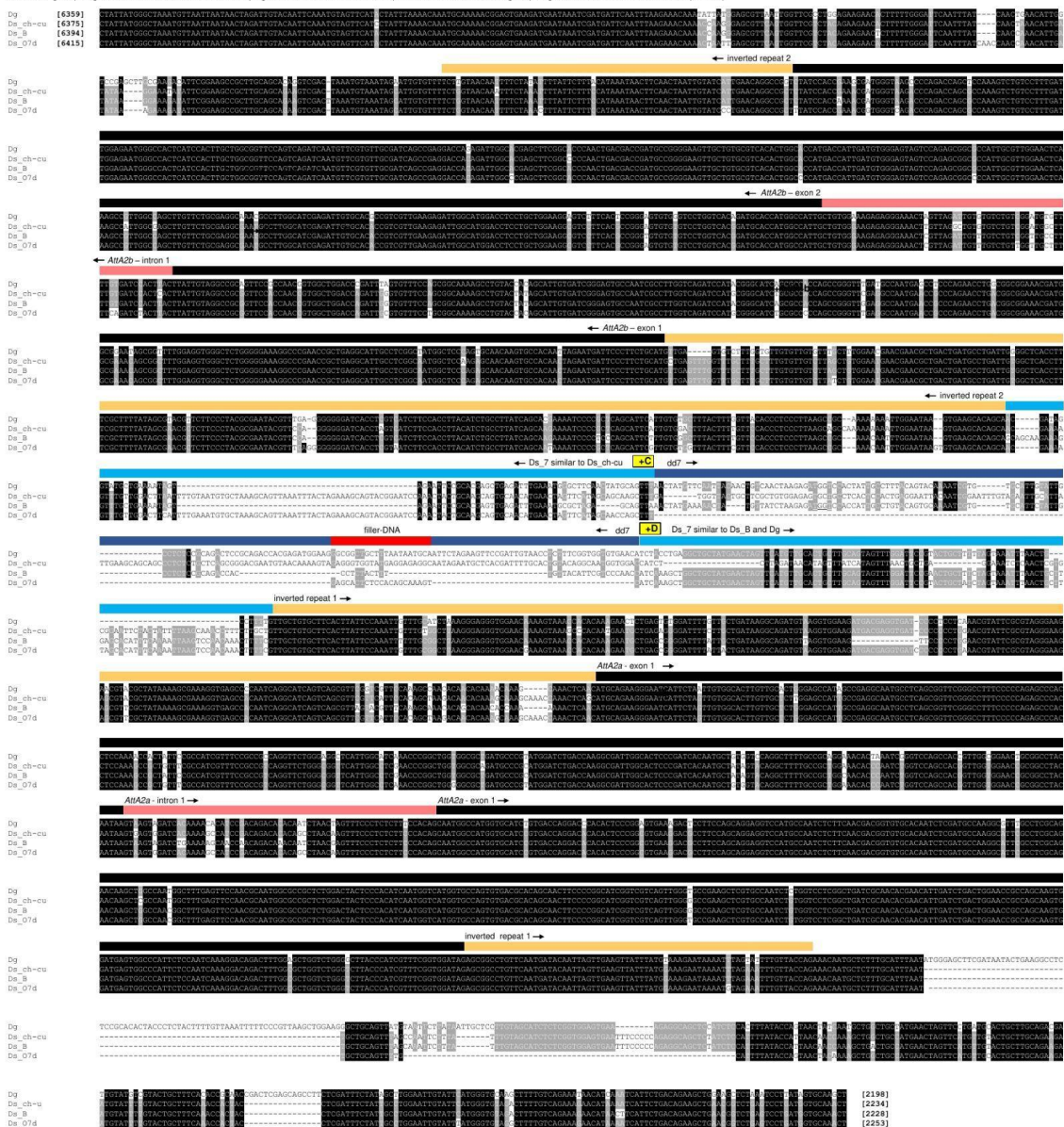
MSA of [+A]+B from the uninverted state (Dg, Ds, ch-cu, and Ds_B) with reconstructed [+A]+B from the inverted state (Ds_7) as it was before O₇ arose



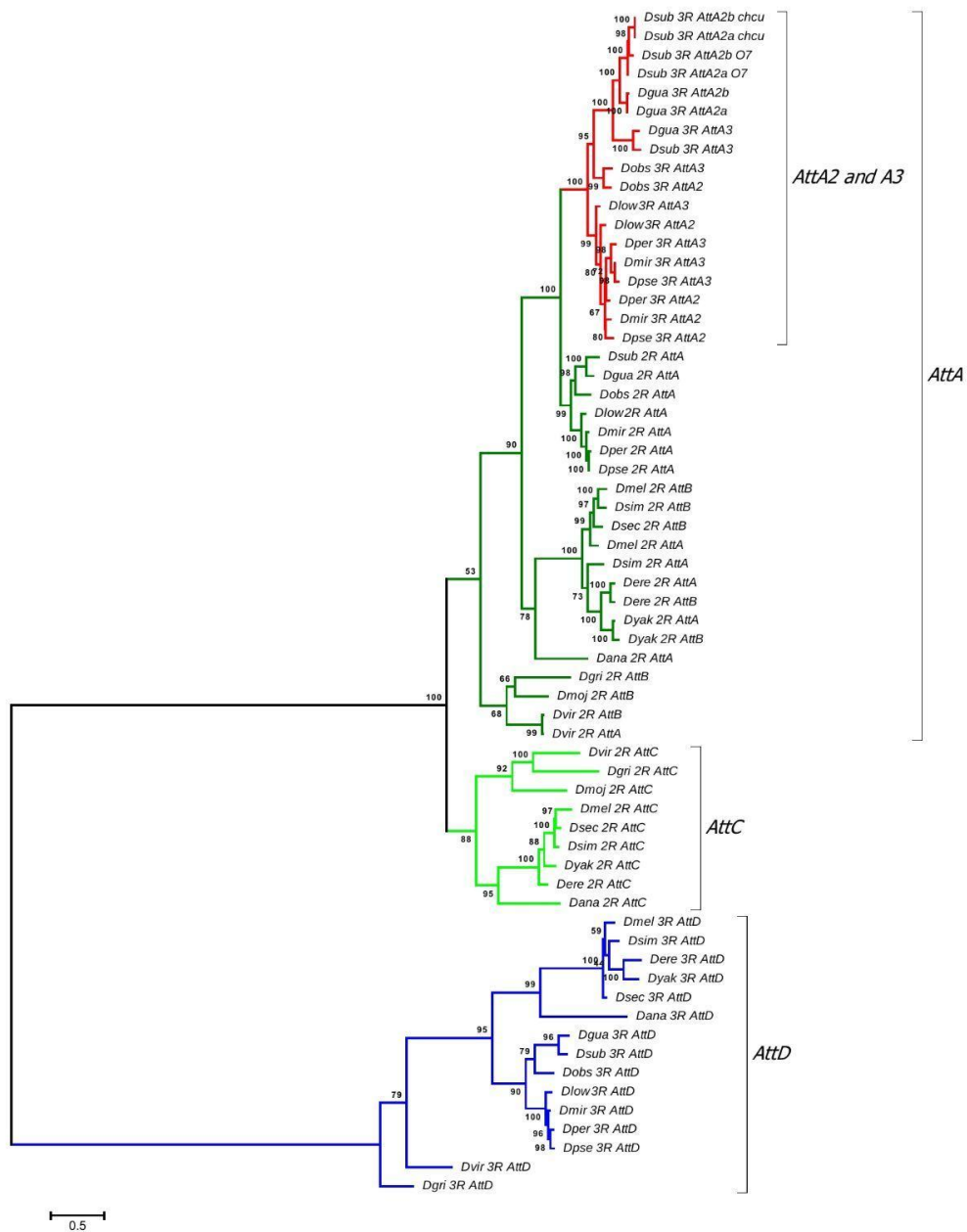
Supplementary Figure S5. Ancestral form of the proximal breakpoint before the occurrence of the DSBs. Similar to Supplementary Figure S4, but with one copy of each of d1 and d2 eliminated.

Distal breakpoint [+C|+D]

MSA of [+C|+D] from the uninverted state (Dg, Ds_ch-cu, and Ds_B) with reconstructed [+C|+D] from the inverted state (Ds_7)



Supplementary Figure S6. MSA of the distal breakpoint region of *O7*, including the [+C|+D] region from the uninverted state (Dg, Ds_ch-cu, and Ds_B) and the [+C|+D] region of *O7* reconstructed by cocatenation of the identified segments C and D. The breakpoint junction between segments C and D is located between the two corresponding yellow boxes above the aligned sequences. Colored boxes above the aligned sequences denote: sepia, the two IRs; black and pink within the IRs, two exons and one intron of each of *AttA2a* and *AttA2b*; light blue, the regions of the central spacer of *O7* between the two IRs that are similar to either Ds_ch-cu or Ds_B and Dg; dar blue within the central spacer, the CSR; red, putative repair-associated filler DNA (see also Figure 3). Black and grey backgrounds denote invariant and 75% conserved MSA columns, respectively. Numbers in brackets are basepair distances to the nearest coding sequence.



Supplementary Figure S7. Maximum likelihood tree of *Drosophila Attacin* genes. The tree includes 63 homologous nucleotide coding sequences (see Supplementary Table 1) with 250 codon sites. Numbers indicate bootstrap support values of IQ-Tree analysis (1000 replicates) with the MGK+F1X4+G4 model. The tree was rooted at the midpoint between the most divergent *Attacins*. The scale bar denotes the estimated number of nucleotide substitutions per site.

ii. Supplementary Tables

Supplementary Table 1. Synteny relationships for *Attacin* genes across *Drosophila*. The MS Excel file contains six spreadsheets, including one for this title, and one for each of five major *Drosophila Attacin* family members (*i.e.*, A-B, A2, A3, C and D). For each *Attacin*, columns “A” to “C” list the taxonomy of the sequences, including the subgenus within *Drosophila*, the group within the *Drosophila* subgenus, and the species. Columns “D” and “E” list the Muller element and the corresponding chromosome of the *Attacin* location. The remaining columns list the *Attacin* genes with the three upstream and downstream flanking genes. *Attacin* genes are highlighted in red, and syntenic orthologous flanking genes in yellow.

[Link to supplementary table 1](#)

References

1. Aguado, C., Gayà-Vidal, M., Villatoro, S., Oliva, M., Izquierdo, D., Giner-Delgado, C., Montalvo, V., García-González, J., Martínez-Fundichely, A., Capilla, L., Ruiz-Herrera, A., Estivill, X., Puig, M. & Cáceres, M. Validation and genotyping of multiple human polymorphic inversions mediated by inverted repeats reveals a high degree of recurrence. *PLoS Genet.* **10**, e1004208 (2014).
2. Anderson, W. W., Arnold, J., Baldwin, D. G., Beckenbach, A. T., Brown, C. J., Bryant, S. H., Coyne, J. A., Harshman, L. G., Heed, W. B. & Jeffery, D. E. Four decades of inversion polymorphism in *Drosophila pseudoobscura*. *Proc. Natl. Acad. Sci. USA* **88**, 10367–10371 (1991).
3. Andolfatto, P., Wall, J. D. & Kreitman, M. Unusual haplotype structure at the proximal breakpoint of In(2L)t in a natural population of *Drosophila melanogaster*. *Genetics* **153**, 1297–1311 (1999).
4. Anton, E., Blanco, J., Egozcue, J. & Vidal, F. Sperm studies in heterozygote inversion carriers: a review. *Cytogenet. Genome Res.* **111**, 297–304 (2005).
5. Ayala, F. J., Serra, L. & Prevosti, A. A grand experiment in evolution: the *Drosophila subobscura* colonization of the Americas. *Genome* **31**, 246–255 (1989).
6. Bächli, G. TaxoDros: The Database on Taxonomy of *Drosophilidae*. Available online at: <https://www.taxodros.uzh.ch> (2020).
7. Bachtrog, D. Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. *Nat. Rev. Genet.* **14**, 113–124 (2013).
8. Balanyá, J., Oller, J. M., Huey, R. B., Gilchrist, G. W. & Serra, L. Global genetic change tracks global climate warming in *Drosophila subobscura*. *Science* **313**, 1773–1775 (2006).
9. Balanyá, J., Serra, L., Gilchrist, G. W., Huey, R. B., Pascual, M., Mestres, F. & Solé, E. Evolutionary pace of chromosomal polymorphism in colonizing populations of *Drosophila subobscura*: an evolutionary time series. *Evolution* **57**, 1837–1845 (2003).
10. Barton, N. H. A general model for the evolution of recombination. *Genet. Res.* **65**, 123–144 (1995).
11. Begon, M. The relationships of *Drosophila obscura* fallén and *D. subobscura* collin to naturally-occurring fruits. *Oecologia* **20**, 255–277 (1975).

12. Besansky, N. J., Krzywinski, J., Lehmann, T., Simard, F., Kern, M., Mukabayire, O., Fontenille, D., Touré, Y. & Sagnon, N. Semipermeable species boundaries between *Anopheles gambiae* and *Anopheles arabiensis* : Evidence from multilocus DNA sequence variation. *Proc. Natl. Acad. Sci. USA* **100**, 10818–10823 (2003).
13. Bhutkar, A., Schaeffer, S. W., Russo, S. M., Xu, M., Smith, T. F. & Gelbart, W. M. Chromosomal rearrangement inferred from comparisons of 12 *Drosophila* genomes. *Genetics* **179**, 1657–1680 (2008).
14. Buels, R., Yao, E., Diesh, C. M., Hayes, R. D., Munoz-Torres, M., Helt, G., Goodstein, D. M., Elisk, C. G., Lewis, S. E., Stein, L. & Holmes, I. H. JBrowse: A dynamic web platform for genome visualization and analysis. *Genome Biol.* **17**, 66 (2016).
15. Burla, H. & Götz, W. Veränderlichkeit des chromosomalen polymorphismus bei *Drosophila subobscura*. *Genetica* **36**, 83–104 (1965).
16. Burla, H., Jungen, H. & Bächli, G. Population structure of *Drosophila subobscura*: Non-random microdispersion of inversion polymorphism on a mountain slope. *Genetica* **70**, 9–15 (1986).
17. Buzzati-Traverso, A. A., & Scossiroli, R. E. *The “Obscura Group” of the Genus Drosophila* (pp. 47–92), (1955).
18. Cáceres, M., Barbadilla, A. & Ruiz, A. Inversion length and breakpoint distribution in the *Drosophila buzzatii* species complex: is inversion length a selected trait? *Evolution* **51**, 1149–1155 (1997).
19. Cáceres, M., Barbadilla, A. & Ruiz, A. Recombination rate predicts inversion size in *Diptera*. *Genetics* **153**, 251–259 (1999).
20. Cáceres, M., Ranz, J. M., Barbadilla, A., Long, M. & Ruiz, A. Generation of a widespread *Drosophila* inversion by a transposable element. *Science* **285**, 415–418 (1999).
21. Cáceres, M., Sullivan, R. T. & Thomas, J. W. A recurrent inversion on the eutherian X chromosome. *Proc. Natl. Acad. Sci. USA* **104**, 18571–18576 (2007).
22. Cameron, D. L., Di Stefano, L. & Papenfuss, A. T. Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. *Nat. Commun.* **10**, 3240 (2019).
23. Carson, H. L. The selective elimination of inversion dicentric chromatids during meiosis in the eggs of *sciara impatiens*. *Genetics* **31**, 95–113 (1946).
24. Charlesworth, D. Plant contributions to our understanding of sex chromosome evolution. *New Phytol.* **208**, 52–65 (2015).

25. Cheng, C. & Kirkpatrick, M. Inversions are bigger on the X chromosome. *Mol. Ecol.* **28**, 1238–1245 (2019).
26. Cheng, C., Tan, J. C., Hahn, M. W. & Besansky, N. J. Systems genetic analysis of inversion polymorphisms in the malaria mosquito *Anopheles gambiae*. *Proc. Natl. Acad. Sci. USA* **115**, E7005-E7014(2018).
27. Coe, B. P., Witherspoon, K., Rosenfeld, J. A., van Bon, B. W. M., Vulto-van Silfhout, A. T., Bosco, P., Friend, K. L., Baker, C., Buono, S., Vissers, L. E. L. M., Schuurs-Hoeijmakers, J. H., Hoischen, A., Pfundt, R., Krumm, N., Carvill, G. L., Li, D., Amaral, D., Brown, N., Lockhart, P. J., Scheffer, I. E., Alberti, A., Shaw, M., Pettinato, R., Tervo, R., de Leeuw, N., Reijnders, M. R. F., Torchia, B. S., Peeters, H., Thompson, E., O’Roak, B. J., Fichera, M., Hehir-Kwa, J. Y., Shendure, J., Mefford, H. C., Haan, E., Géczy, J., de Vries, B. B. A., Romano, C. & Eichler, E. E. Refining analyses of copy number variation identifies specific genes associated with developmental delay. *Nat. Genet.* **46**, 1063–1071 (2014).
28. Collin, J. E. Note: *Drosophila subobscura* sp. n. male, female. *J Genet* **33**, 60 (1936).
29. Connallon, T. & Olito, C. Natural selection and the distribution of chromosomal inversion lengths. *Mol. Ecol.* **31**, 3627–3641 (2022).
30. Cooper, K. W. Concerning the origin of the polytene chromosomes of *Diptera*. *Proc. Natl. Acad. Sci. USA* **24**, 452–458 (1938).
31. Corbett-Detig, R. B. Selection on inversion breakpoints favors proximity to pairing sensitive sites in *Drosophila melanogaster*. *Genetics* **204**, 259–265 (2016).
32. Corbett-Detig, R. B. & Hartl, D. L. Population genomics of inversion polymorphisms in *Drosophila melanogaster*. *PLoS Genet.* **8**, e1003056 (2012).
33. Coughlan, J. M. & Willis, J. H. Dissecting the role of a large chromosomal inversion in life history divergence throughout the *Mimulus guttatus* species complex. *Mol. Ecol.* **28**, 1343–1357 (2019).
34. della Torre, A., Merzagora, L., Powell, J. R. & Coluzzi, M. Selective introgression of paracentric inversions between two sibling species of the *Anopheles gambiae* Complex. *Genetics* **146**, 239–244 (1997).
35. Delprat, A., Guillén, Y. & Ruiz, A. Computational sequence analysis of inversion breakpoint regions in the cactophilic *Drosophila mojavensis* Lineage. *J. Hered.* **110**, 102–117 (2019).
36. Dobzhansky, T. & Sturtevant, A. H. Inversions in the chromosomes of *Drosophila pseudoobscura*. *Genetics* **23**, 28–64 (1938).

37. Dobzhansky, T. Genetics of natural populations. XVI. Altitudinal and seasonal changes produced by natural selection in certain populations of *Drosophila pseudoobscura* and *Drosophila persimilis*. *Genetics* **33**, 158–176 (1948).
38. Dobzhansky, T. Rigid vs. flexible chromosomal polymorphisms in *Drosophila*. *Am. Nat.* **96**, 321–328 (1962).
39. Dobzhansky, T. Genetics of natural populations. XIV. A response of certain gene arrangements in the third chromosome of *Drosophila pseudoobscura* to natural selection. *Genetics* **32**, 142–160 (1947).
40. Dobzhansky, T. & Epling, C. The suppression of crossing over in inversion heterozygotes of *Drosophila pseudoobscura*. *Proc. Natl. Acad. Sci. USA* **34**, 137–141 (1948).
41. Dufresnes, C. & Crochet, P. A. Sex chromosomes as supergenes of speciation: why amphibians defy the rules? *Philos. Trans. R. Soc. B Biol. Sci.* **377**, 20210202 (2022).
42. Emmens, C. W. The morphology of the nucleus in the salivary glands of four species of *Drosophila*. *Zeitschrift für Zellforsch. und Mikroskopische Anat.* **26**, 1–20 (1937).
43. Feder, J. L., Berlocher, S. H., Roethele, J. B., Dambroski, H., Smith, J. J., Perry, W. L., Gavrilovic, V., Filchak, K. E., Rull, J. & Aluja, M. Allopatric genetic origins for sympatric host-plant shifts and race formation in *Rhagoletis*. *Proc. Natl. Acad. Sci. USA* **100**, 10314–10319 (2003).
44. Felsenstein, J. The evolutionary advantage of recombination. *Genetics* **78**, 737–756 (1974).
45. Fisher, R. A. XXI.—On the dominance ratio. *Proc. R. Soc. Edinburgh* **42**, 321–341 (1923).
46. Fontdevila, A., Zapata, C., Alvarez, G., Sanchez, L., Méndez, J. & Enriquez, I. Genetic coadaptation in the chromosomal polymorphism of *Drosophila subobscura*. I. seasonal changes of gametic disequilibrium in a natural population. *Genetics* **105**, 935–955 (1983).
47. Fuller, Z. L., Haynes, G. D., Richards, S. & Schaeffer, S. W. Genomics of natural populations: how differentially expressed genes shape the evolution of chromosomal inversions in *Drosophila pseudoobscura*. *Genetics* **204**, 287–301 (2016).
48. Fuller, Z. L., Haynes, G. D., Richards, S. & Schaeffer, S. W. Genomics of natural populations: Evolutionary forces that establish and maintain gene arrangements in *Drosophila pseudoobscura*. *Mol. Ecol.* **26**, 6539–6562 (2017).
49. Fuller, Z. L., Koury, S. A., Phadnis, N. & Schaeffer, S. W. How chromosomal rearrangements shape adaptation and speciation: Case studies in *Drosophila*

- pseudoobscura* and its sibling species *Drosophila persimilis*. *Mol. Ecol.* **28**, 1283–1301 (2019).
50. Fuller, Z. L., Leonard, C. J., Young, R. E., Schaeffer, S. W. & Phadnis, N. Ancestral polymorphisms explain the role of chromosomal inversions in speciation. *PLoS Genet.* **14**, e1007526 (2018).
 51. Goidts, V., Szamalek, J. M., de Jong, P. J., Cooper, D. N., Chuzhanova, N., Hameister, H. & Kehrer-Sawatzki, H. Independent intrachromosomal recombination events underlie the pericentric inversions of chimpanzee and gorilla chromosomes homologous to human chromosome 16. *Genome Res.* **15**, 1232–1242 (2005).
 52. Good, B. H., Walczak, A. M., Neher, R. A. & Desai, M. M. Genetic diversity in the interference selection limit. *PLoS Genet.* **10**, e1004222 (2014).
 53. Götz, K. G. Die optischen Übertragungseigenschaften der Komplexaugen von *Drosophila*. *Kybernetik* **2**, 215–221 (1965).
 54. Gutiérrez-Valencia, J., Hughes, P. W., Berdan, E. L. & Slotte, T. The genomic architecture and evolutionary fates of supergenes. *Genome Biol. Evol.* **13**, evab057 (2021).
 55. Hedrick, P. W. Genetic polymorphism in heterogeneous environments: The age of genomics. *Annu. Rev. Ecol. Evol. Syst.* **37**, 67–93 (2006).
 56. Hill, W. G. & Robertson, A. The effects of inbreeding at loci with heterozygote advantage. *Genetics* **60**, 615–628 (1968).
 57. Hoffmann, A. A. & Rieseberg, L. H. Revisiting the impact of inversions in evolution: from population genetic markers to drivers of adaptive shifts and speciation? *Annu. Rev. Ecol. Evol. Syst.* **39**, 21–42 (2008).
 58. Jay, P., Leroy, M., Le Poul, Y., Whibley, A., Arias, M., Chouteau, M. & Joron, M. Association mapping of colour variation in a butterfly provides evidence that a supergene locks together a cluster of adaptive loci. *Philos. Trans. R. Soc. B Biol. Sci.* **377**, 20210193 (2022).
 59. Jiang, T., Liu, S., Cao, S., Liu, Y., Cui, Z., Wang, Y. & Guo, H. Long-read sequencing settings for efficient structural variation detection based on comprehensive evaluation. *BMC Bioinformatics* **22**, 552 (2021).
 60. Kaiser, V. B. & Charlesworth, B. The effects of deleterious mutations on evolution in non-recombining genomes. *Trends Genet.* **25**, 9–12 (2009).
 61. Karageorgiou, C., Gámez-Visairas, V., Tarrío, R. & Rodríguez-Trelles, F. Long-read based assembly and synteny analysis of a reference *Drosophila subobscura* genome

reveals signatures of structural evolution driven by inversions
recombination-suppression effects. *BMC Genomics* **20**, 223 (2019).

62. Karageorgiou, C., Tarrío, R. & Rodríguez-Trelles, F. The cyclically seasonal *Drosophila subobscura* inversion O₇ originated from fragile genomic sites and relocated immunity and metabolic genes. *Front. Genet.* **11**, 565836 (2020).
63. Kay, T., Helleu, Q. & Keller, L. Iterative evolution of supergene-based social polymorphism in ants. *Philos. Trans. R. Soc. B Biol. Sci.* **377**, 20210196 (2022).
64. Kehrer-Sawatzki, H., Sandig, C. A., Goidts, V. & Hameister, H. Breakpoint analysis of the pericentric inversion between chimpanzee chromosome 10 and the homologous chromosome 12 in humans. *Cytogenet. Genome Res.* **108**, 91–97 (2005).
65. Kimura, M. & Ohta, T. Probability of fixation of a mutant gene in a finite population when selective advantage decreases with time. *Genetics* **65**, 525–534 (1970).
66. Kirkpatrick, M. How and why chromosome inversions evolve. *PLoS Biol.* **8**, e1000501 (2010).
67. Kirkpatrick, M. & Barrett, B. Chromosome inversions, adaptive cassettes and the evolution of species' ranges. *Mol. Ecol.* **24**, 2046–2055 (2015).
68. Kirkpatrick, M. & Barton, N. Chromosome inversions, local adaptation and speciation. *Genetics* **173**, 419–434 (2006).
69. Kirkpatrick, M. & Kern, A. Where's the money? Inversions, genes, and the hunt for genomic targets of selection. *Genetics* **190**, 1153–1155 (2012).
70. Korunes, K. L. & Noor, M. A. F. Gene conversion and linkage: effects on genome evolution and speciation. *Mol. Ecol.* **26**, 351–364 (2017).
71. Krimbas, C. B. & Loukas, M. Evolution of the *obscura* group *Drosophila* species. I. Salivary chromosomes and quantitative characters in *D. subobscura* and two closely related species. *Heredity (Edinb.)* **53**, 469–482 (1984).
72. Krimbas, C. B. *Drosophila subobscura: Biology, Genetics and Inversion Polymorphism*. (Verlag Dr. Kovac, 1993).
73. Krimbas, C. B. & Powell, J. R. *Drosophila Inversion Polymorphism*. (CRC press, 1992).
74. Kunze-Mühl, E. & Müller, E. Weitere Untersuchungen über die chromosomale Struktur und die natürlichen Strukturtypen von *Drosophila subobscura* coll. *Chromosoma* **9**, 559–570 (1957).
75. Lamichhaney, S., Fan, G., Widemo, F., Gunnarsson, U., Thalmann, D. S., Hoepfner, M. P., Kerje, S., Gustafson, U., Shi, C., Zhang, H., Chen, W., Liang, X., Huang, L., Wang,

- J., Liang, E., Wu, Q., Lee, S. M.-Y., Xu, X., Höglund, J., Liu, X. & Andersson, L. Structural genomic changes underlie alternative reproductive strategies in the ruff (*Philomachus pugnax*). *Nat. Genet.* **48**, 84–88 (2016).
76. Lavington, E. & Kern, A. D. The effect of common inversion polymorphisms *In(2L)t* and *In(3R)Mo* on patterns of transcriptional variation in *Drosophila melanogaster*. *G3 Genes|Genomes|Genetics* **7**, 3659–3668 (2017).
77. Lee, J. A., Carvalho, C. M. B. & Lupski, J. R. A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell* **131**, 1235–1247 (2007).
78. Levene, H. Genetic equilibrium when more than one ecological niche is available. *Am. Nat.* **87**, 331–333 (1953).
79. Lewontin, R., Kirk, D. & Crow, J. Selective mating, assortative mating, and inbreeding: Definitions and implications. *Eugen. Q.* **15**, 141–143 (1968).
80. Logsdon, G. A., Vollger, M. R. & Eichler, E. E. Long-read human genome sequencing and its applications. *Nat. Rev. Genet.* **21**, 597–614 (2020).
81. Lowry, D. B., Popovic, D., Brennan, D. J. & Holeski, L. M. Mechanisms of a locally adaptive shift in allocation among growth, reproduction, and herbivore resistance in *Mimulus guttatus*. *Evolution* **73**, 1168–1181 (2019).
82. Mahmoud, M., Gobet, N., Cruz-Dávalos, D. I., Mounier, N., Dessimoz, C. & Sedlazeck, F. J. Structural variant calling: the long and the short of it. *Genome Biol.* **20**, 246 (2019).
83. Mather, K. The genetical architecture of heterostyly in *Primula sinensis*. *Evolution* **4**, 340 (1950).
84. Matschiner, M., Barth, J. M. I., Tørresen, O. K., Star, B., Baalsrud, H. T., Briec, M. S. O., Pampoulie, C., Bradbury, I., Jakobsen, K. S. & Jentoft, S. Supergene origin and maintenance in Atlantic cod. *Nat. Ecol. Evol.* **6**, 469–481 (2022).
85. Matzkin, L. M., Merritt, T. J. S., Zhu, C.-T. & Eanes, W. F. The structure and population genetics of the breakpoints associated with the cosmopolitan chromosomal inversion *In(3R)Payne* in *Drosophila melanogaster*. *Genetics* **170**, 1143–1152 (2005).
86. Maynard Smith, J. *Evolutionary Genetics*. (Oxford University Press, 1998).
87. McBroome, J., Liang, D. & Corbett-Detig, R. Fine-scale position effects shape the distribution of inversion breakpoints in *Drosophila melanogaster*. *Genome Biol. Evol.* **12**, 1378–1391 (2020).

88. Menozzi, P. & Krimbas, C. B. The inversion polymorphism of *D. subobscura* revisited: Synthetic maps of gene arrangement frequencies and their interpretation. *J. Evol. Biol.* **5**, 625–641 (1992).
89. Messer, P. W. Measuring the rates of spontaneous mutation from deep and large-scale polymorphism data. *Genetics* **182**, 1219–1232 (2009).
90. Mestres, F., Balanyà, J., Arenas, C., Solé, E. & Serra, L. Colonization of America by *Drosophila subobscura* : Heterotic effect of chromosomal arrangements revealed by the persistence of lethal genes. *Proc. Natl. Acad. Sci. USA* **98**, 9167–9170 (2001).
91. Moltó, M. D., De Frutos, R. & Martínez-Sebastián, M. J. The banding pattern of polytene chromosomes of *Drosophila guanache* compared with that of *D. subobscura*. *Genetica* **75**, 55–70 (1987).
92. Murphy, W. J., Larkin, D. M., der Wind, A. E., Bourque, G., Tesler, G., Auvil, L., Beever, J. E., Chowdhary, B. P., Galibert, F., Gatzke, L., Hitte, C., Meyers, S. N., Milan, D., Ostrander, E. A., Pape, G., Parker, H. G., Raudsepp, T., Rogatcheva, M. B., Schook, L. B., Skow, L. C., Welge, M., Womack, J. E., O'Brien, S. J., Pevzner, P. A. & Lewin, H. A. Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science* **309**, 613–617 (2005).
93. Narayanan, V., Mieczkowski, P. A., Kim, H.-M., Petes, T. D. & Lobachev, K. S. The pattern of gene amplification is determined by the chromosomal location of hairpin-capped breaks. *Cell* **125**, 1283–1296 (2006).
94. Nei, M., Kojima, K.-I. & Schaffer, H. E. Frequency changes of new inversions in populations under mutation-selection equilibria. *Genetics* **57**, 741–750 (1967).
95. Noor, M. A. F., Grams, K. L., Bertucci, L. A. & Reiland, J. Chromosomal inversions and the reproductive isolation of species. *Proc. Natl. Acad. Sci. USA* **98**, 12084–12088 (2001).
96. Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S. J., Diekhans, M., Logsdon, G. A., Alonge, M., Antonarakis, S. E., Borchers, M., Bouffard, G. G., Brooks, S. Y., Caldas, G. V., Chen, N.-C., Cheng, H., Chin, C.-S., Chow, W., de Lima, L. G., Dishuck, P. C., Durbin, R., Dvorkina, T., Fiddes, I. T., Formenti, G., Fulton, R. S., Functamman, A., Garrison, E., Grady, P. G. S., Graves-Lindsay, T. A., Hall, I. M., Hansen, N. F., Hartley, G. A., Haukness, M., Howe, K., Hunkapiller, M. W., Jain, C., Jain, M., Jarvis, E. D., Kerpedjiev, P., Kirsche, M., Kolmogorov, M., Korlach, J., Kremitzki, M., Li, H., Maduro, V. V., Marschall, T., McCartney, A. M., McDaniel, J., Miller, D. E., Mullikin, J. C., Myers, E. W., Olson, N. D., Paten, B., Peluso, P., Pevzner, P. A., Porubsky, D., Potapova, T., Rogaev, E. I., Rosenfeld, J. A., Salzberg, S. L., Schneider, V. A., Sedlazeck, F. J., Shafin, K., Shew, C. J., Shumate, A., Sims, Y., Smit, A. F. A., Soto, D. C., Sović, I., Storer, J. M., Streets, A., Sullivan, B. A.,

- Thibaud-Nissen, F., Torrance, J., Wagner, J., Walenz, B. P., Wenger, A., Wood, J. M. D., Xiao, C., Yan, S. M., Young, A. C., Zarate, S., Surti, U., McCoy, R. C., Dennis, M. Y., Alexandrov, I. A., Gerton, J. L., O'Neill, R. J., Timp, W., Zook, J. M., Schatz, M. C., Eichler, E. E., Miga, K. H. & Phillippy, A. M. The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
97. Orengo, D. J., Puerma, E., Papaceit, M., Segarra, C. & Aguadé, M. A molecular perspective on a complex polymorphic inversion system with cytological evidence of multiply reused breakpoints. *Heredity (Edinb)*. **114**, 610–618 (2015).
98. Orengo, D.J. & Prevosti, A. Temporal changes in chromosomal polymorphism of *Drosophila subobscura* related to climatic changes. *Evolution* **50**, 1346 (1996).
99. Papaceit, M. & Prevosti, A. Differences in chromosome A arrangement between *Drosophila madeirensis* and *Drosophila subobscura*. *Experientia* **45**, 310–312 (1989).
100. Papaceit, M., Segarra, C. & Aguadé, M. Structure and population genetics of the breakpoints of a polymorphic inversion in *Drosophila subobscura*. *Evolution* **67**, 66–79 (2013).
101. Porubsky, D., Höps, W., Ashraf, H., Hsieh, P., Rodriguez-Martin, B., Yilmaz, F., Ebler, J., Hallast, P., Maria Maggiolini, F. A., Harvey, W. T., Henning, B., Audano, P. A., Gordon, D. S., Ebert, P., Hasenfeld, P., Benito, E., Zhu, Q., Lee, C., Antonacci, F., Steinrücken, M., Beck, C. R., Sanders, A. D., Marschall, T., Eichler, E. E. & Korbel, J. O. Recurrent inversion polymorphisms in humans associate with genetic instability and genomic disorders. *Cell* **185**, 1986–2005.e26 (2022).
102. Porubsky, D., Sanders, A. D., Höps, W., Hsieh, P., Sulovari, A., Li, R., Mercuri, L., Sorensen, M., Murali, S. C., Gordon, D., Cantsilieris, S., Pollen, A. A., Ventura, M., Antonacci, F., Marschall, T., Korbel, J. O. & Eichler, E. E. Recurrent inversion toggling and great ape genome evolution. *Nat. Genet.* **52**, 849–858 (2020).
103. Powell, J. R. *Progress and Prospects in Evolutionary Biology: The Drosophila Model*. (Oxford University Press, 1997).
104. Prevosti, A., Ribo, G., Serra, L., Aguade, M., Balaña, J., Monclus, M. & Mestres, F. Colonization of America by *Drosophila subobscura* : Experiment in natural populations that supports the adaptive role of chromosomal-inversion polymorphism. *Proc. Natl. Acad. Sci. USA* **85**, 5597–5600 (1988).
105. Priyam, A., Woodcroft, B. J., Rai, V., Moghul, I., Munagala, A., Ter, F., Chowdhary, H., Pieniak, I., Maynard, L. J., Gibbins, M. A., Moon, H., Davis-Richardson, A., Uludag, M., Watson-Haigh, N. S., Challis, R., Nakamura, H., Favreau, E., Gómez, E. A., Pluskal, T., Leonard, G., Rumpf, W. & Wurm, Y. Sequenceserver: A modern graphical user interface for custom BLAST Databases. *Mol. Biol. Evol.* **36**, 2922–2924 (2019).

106. Puerma, E., Orengo, D. J. & Aguadé, M. Inversion evolutionary rates might limit the experimental identification of inversion breakpoints in non-model species. *Sci. Rep.* **7**, 17281 (2017).
107. Puerma, E., Orengo, D. J. & Aguadé, M. Multiple and diverse structural changes affect the breakpoint regions of polymorphic inversions across the *Drosophila* genus. *Sci. Rep.* **6**, 36248 (2016).
108. Puerma, E., Orengo, D. J. & Aguadé, M. The origin of chromosomal inversions as a source of segmental duplications in the *Sophophora* subgenus of *Drosophila*. *Sci. Rep.* **6**, 30715 (2016).
109. Puerma, E., Orengo, D. J., Salguero, D., Papaceit, M., Segarra, C. & Aguadé, M. Characterization of the breakpoints of a polymorphic inversion complex detects strict and broad breakpoint reuse at the molecular level. *Mol. Biol. Evol.* **31**, 2331–2341 (2014).
110. Ranz, J. M., Maurin, D., Chan, Y. S., von Grotthuss, M., Hillier, L. W., Roote, J., Ashburner, M. & Bergman, C. M. Principles of genome evolution in the *Drosophila melanogaster* species group. *PLoS Biol.* **5**, e152 (2007).
111. Rego, C., Balanyà, J., Fragata, I., Matos, M., Rezende, E. L. & Santos, M. Clinal patterns of chromosomal inversion polymorphisms in *Drosophila subobscura* are partly associated with thermal preferences and heat stress resistance. *Evolution* **64**, 385–397 (2010).
112. Rezende, E., Balanyà, J., Rodríguez-Trelles, F., Rego, C., Fragata, I., Matos, M., Serra, L. & Santos, M. Climate change and chromosomal inversions in *Drosophila subobscura*. *Clim. Res.* **43**, 103–114 (2010).
113. Rieseberg, L. H. Chromosomal rearrangements and speciation. *Trends Ecol. Evol.* **16**, 351–358 (2001).
114. Rodríguez-Trelles, F., Alvarez, G. & Zapata, C. Time-series analysis of seasonal changes of the *O* inversion polymorphism of *Drosophila subobscura*. *Genetics* **142**, 179–187 (1996).
115. Rodríguez-Trelles, F. & Rodríguez, M. A. Measuring evolutionary responses to global warming: cautionary lessons from *Drosophila*. *Insect Conserv. Divers.* **3**, 44–50 (2010).
116. Rodríguez-Trelles, F. & Rodríguez, M. A. Rapid micro-evolution and loss of chromosomal diversity in *Drosophila* in response to climate warming. *Evol. Ecol.* **12**, 829–838 (1998).

117. Rodríguez-Trelles, F., Tarrío, R. & Santos, M. Genome-wide evolutionary response to a heat wave in *Drosophila*. *Biol. Lett.* **9**, 20130228 (2013).
118. Said, I., Byrne, A., Serrano, V., Cardeno, C., Vollmers, C. & Corbett-Detig, R. Linked genetic variation and not genome structure causes widespread differential expression associated with chromosomal inversions. *Proc. Natl. Acad. Sci. USA* **115**, 5492–5497 (2018).
119. Schattner, P. *Genomes, Browsers and Databases. Data-Mining Tools for Integrated Genomic Databases*. (Cambridge University Press, 2008).
120. Solé, E., Balanyá, J., Sperlich, D. & Serra, L. Long-term changes in the chromosomal inversion polymorphism of *Drosophila subobscura*. I. Mediterranean populations from southwestern Europe. *Evolution* **56**, 830–835 (2002).
121. Sperlich, D. & Feuerbach, H. Ist der chromosomale Struktur polymorphismus von *Drosophila Subobscura* stabil oder flexibel? *Z. Vererbungsl.* **98**, 16–24 (1966).
122. Stein, L. D. Using GBrowse 2.0 to visualize and share next-generation sequence data. *Brief. Bioinform.* **14**, 162–171 (2013).
123. Sturtevant, A. H. & Beadle, G. W. The relations of inversions in the X chromosome of *Drosophila melanogaster* to crossing over and disjunction. *Genetics* **21**, 554–604 (1936).
124. Sturtevant, A. H. A case of rearrangement of genes in *Drosophila*. *Proc. Natl. Acad. Sci. USA* **7**, 235–237 (1921).
125. Tadin-Strapps, M., Warburton, D., Baumeister, F. A. M., Fischer, S. G., Yonan, J., Gilliam, T. C. & Christiano, A. M. Cloning of the breakpoints of a *de novo* inversion of chromosome 8, Inv(8)(p11.2q23.1) in a patient with Ambras syndrome. *Cytogenet. Genome Res.* **107**, 68–76 (2004).
126. Takahashi, Y. & Kawata, M. A comprehensive test for negative frequency-dependent selection. *Popul. Ecol.* **55**, 499–509 (2013).
127. Thompson, M. J. & Jiggins, C. D. Supergenes and their role in evolution. *Heredity (Edinb)*. **113**, 1–8 (2014).
128. Wallace, B. *Genetic Load: Its Biological and Conceptual Aspects*. (Prentice Hall, 1970).
129. Wang, J., Kong, L., Gao, G. & Luo, J. A brief introduction to web-based genome browsers. *Brief. Bioinform.* **14**, 131–143 (2013).
130. Wasserman, M. Recombination-induced chromosomal heterosis. *Genetics* **58**, 125–139 (1968).

131. Wesley, C. S. & Eanes, W. F. Isolation and analysis of the breakpoint sequences of chromosome inversion In(3L)Payne in *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. USA* **91**, 3132–3136 (1994).
132. Wetterstrand KA. DNA sequencing costs: Data from the NHGRI Genome Sequencing Program (GSP). Available at: www.genome.gov/sequencingcostsdata
133. White, B. J., Hahn, M. W., Pombi, M., Cassone, B. J., Lobo, N. F., Simard, F. & Besansky, N. J. Localization of candidate regions maintaining a common polymorphic inversion (2La) in *Anopheles gambiae*. *PLoS Genet.* **3**, e217 (2007).
134. Wittmann, M. J., Bergland, A. O., Feldman, M. W., Schmidt, P. S. & Petrov, D. A. Seasonally fluctuating selection can maintain polymorphism at many loci via segregation lift. *Proc. Natl. Acad. Sci. USA* **114**, E9932-E9941 (2017).
135. Wright, D. & Schaeffer, S. W. The relevance of chromatin architecture to genome rearrangements in *Drosophila*. *Philos. Trans. R. Soc. B Biol. Sci.* **377**, 20210206 (2022).