DOI: 10.1534/genetics.107.086108

# Chromosomal Rearrangement Inferred From Comparisons of 12 Drosophila Genomes

**Arjun Bhutkar,*,†,1 Stephen W. Schaeffer,‡ Susan M. Russo,* Mu Xu,‡ Temple F. Smith† and William M. Gelbart***

**Department of Molecular and Cellular Biology, Harvard University, Cambridge, Massachusetts 02138, †Bioinformatics Program, Boston University, Boston, Massachusetts 02215 and ‡Department of Biology and Institute of Molecular Evolutionary Genetics, Pennsylvania State University, University Park, Pennsylvania 16802-5301*

Manuscript received December 16, 2007
Accepted for publication May 22, 2008

## ABSTRACT

The availability of 12 complete genomes of various species of genus Drosophila provides a unique opportunity to analyze genome-scale chromosomal rearrangements among a group of closely related species. This article reports on the comparison of gene order between these 12 species and on the fixed rearrangement events that disrupt gene order. Three major themes are addressed: the conservation of syntenic blocks across species, the disruption of syntenic blocks (via chromosomal inversion events) and its relationship to the phylogenetic distribution of these species, and the rate of rearrangement events over evolutionary time. Comparison of syntenic blocks across this large genomic data set confirms that genetic elements are largely (95%) localized to the same Muller element across genus Drosophila species and paracentric inversions serve as the dominant mechanism for shuffling the order of genes along a chromosome. Gene-order scrambling between species is in accordance with the estimated evolutionary distances between them and we find it to approximate a linear process over time (linear to exponential with alternate divergence time estimates). We find the distribution of synteny segment sizes to be biased by a large number of small segments with comparatively fewer large segments. Our results provide estimated chromosomal evolution rates across this set of species on the basis of whole-genome synteny analysis, which are found to be higher than those previously reported. Identification of conserved syntenic blocks across these genomes suggests a large number of conserved blocks with varying levels of embryonic expression correlation in *Drosophila melanogaster.* On the other hand, an analysis of the disruption of syntenic blocks between species allowed the identification of fixed inversion breakpoints and estimates of breakpoint reuse and lineage-specific breakpoint event segregation.

DROSOPHILA research has a rich history in the study of genome rearrangements that now culminates with the analysis of complete genomic sequences. Comparative genomics in Drosophila began when linkage maps of morphological traits were used to establish the homologies of six chromosomal arms in closely related species (Donald 1936; Sturtevant and Tan 1937; Muller 1940; Sturtevant and Novitski 1941). These early studies established the idea that genes are syntenic or conserved on the same chromosome arm among species. One difficulty encountered with these early comparative genomic analyses was that chromosomal arm nomenclatures varied among species (Crew and Lamy 1935; Sturtevant and Tan 1937). Muller (1940) overcame this problem when he developed a standard nomenclature that assigned a letter to each of the chromosomal arms or elements on the basis of the *Drosophila melanogaster* genome (chromosomal arm equals Muller element: X = A, 2L = B, 2R = C, 3L = D, 3R = E, 4 = F). Sturtevant and Novitski (1941) showed that the conservation of chromosomal elements extended to species across the entire genus of Drosophila.

The conservation of the gene content within Muller elements across the genus Drosophila has been well supported as more species have been examined and as molecular genetic markers have been used to develop more detailed genetic and physical maps (Spassky and Dobzhansky 1950; Loukas *et al.* 1979; Steinemann *et al.* 1984; Whiting *et al.* 1989; Segarra *et al.* 1995, 1996; Ranz *et al.* 1997, 2001, 2003). As the density of genetic and physical markers on the maps of Drosophila species has increased, it has become clear that gene order within Muller elements is not conserved among species (Ranz *et al.* 2001).

Chromosomal inversions that result from two breakage and rejoining events in DNA are the agents of gene-order change. Examination of the polytene chromosomes of Drosophila salivary glands provided the first glimpse into the structural mutations that alter the genome (Painter 1934). In a landmark study, Dobzhansky and

[1] *Corresponding author:* Gelbart Lab, Department of Molecular and Cellular Biology, Harvard University, 16 Divinity Ave., Room 4059, Cambridge, MA 02138. E-mail: arjunb@morgan.harvard.edu

Sturtevant (1938) discovered that natural populations of *D. pseudoobscura* harbored a wealth of gene arrangement polymorphism. The gene arrangement diversity in *D. pseudoobscura* was organized into an unrooted network or phylogeny of gene arrangements that were linked together in an evolutionary sequence of single paracentric inversion events.

There is a continuum of gene arrangement diversity within species of Drosophila. Some species are fixed for a single gene arrangement on each chromosomal arm, while others have gene arrangement polymorphisms on all major Muller elements (Sperlich and Pfriem 1986). Comparisons of polytene chromosomes among closely related species have been used to reconstruct the history of rearrangements (Dobzhansky 1944; Lemeunier and Ashburner 1976), but the polytene maps of more distantly related species are not readily comparable because of the accumulation of large numbers of rearrangements. Comparison of gene order in the newly available 12 Drosophila genome sequences (Drosophila 12 Genomes Consortium 2007) provides a unique opportunity to understand the evolutionary forces that have acted on chromosomal inversions during the history of the Drosophila genus.

The level of detail available in the annotation of genomic sequence requires us to carefully define terms that are relevant to the study of gene-order variation among species. A gene is syntenic when it is found on the same chromosome (Muller element) in two or more species (Ehrlich *et al.* 1997). We denote this feature with the term "arm-level synteny." A syntenic block is a genomic region containing a set of two or more syntenic genes that are in the same order and orientation in two or more species. A single gene would not be considered a syntenic block by this definition. However, the evolutionary process of inversions could lead to a one-gene block if an inversion happened up- and downstream of a gene. Chromosomal inversion events, which are thought to disrupt gene order, can be considered at the macrolevel or microlevel depending on the number of genes involved in an inversion. Our definition of a syntenic block uses relaxed criteria where localized scrambling of gene order due to micro-inversions is permitted within larger blocks resulting from macro-inversion events. We use a cutoff parameter (see materials and methods) to permit localized scrambling involving few genes (micro-inversions) within larger inversion units consisting of a large number of genes (macro-inversions). Macro-inversions will break up a chromosome into many syntenic blocks where localized gene order and transcriptional orientation involving a few genes may vary between species because of micro-inversion events. A multispecies conserved block is one where there is conservation of synteny across species for that set of genes (and species-specific localized scrambling of gene order due to micro-inversions is permitted). Finally, a rearrangement breakpoint is defined as a genomic region between pairs of syntenic and/or relaxed syntenic blocks. The breakpoint would consist of the nucleotide sequence from the end of one block to the beginning of the next block. When we infer that a breakpoint is reused we mean that two or more breakage events occurred within the nucleotide interval between blocks, but the events are not necessarily coincident within the breakpoint (Gonzalez *et al.* 2007).

Comparison of the complete sequences in *D. melanogaster* and *D. pseudoobscura* provided the first comprehensive view of genomic rearrangements in the genus Drosophila (Richards *et al.* 2005). The process of rearrangement was examined by inferring syntenic blocks. Rearrangement breakpoints were found to harbor repetitive sequences that might have facilitated the rearrangement process. Inversions were observed to accumulate at a rate of 10 per 1 million years since the two species diverged from a common ancestor. Richards *et al.* (2005), however, were unable to determine which breakpoints were used on the two lineages, which limited inferences about the role repetitive sequences played in the rearrangement process. The *Anopheles gambiae* sequence was available at the time (Holt *et al.* 2002), but extensive rearrangement between the mosquito and fly lineages reduced confidence in the lineage inference.

One of the goals for the 12 Drosophila genomes project was to develop bioinformatics tools for the assembly, annotation, and analysis of groups of related taxa such as mammalian genomes. Information about synteny block boundaries is a valuable bioinformatics tool to aid in the ordering of scaffolds that emerge from a whole-genome shotgun project (Schaeffer *et al.* 2008, accompanying article, this issue), to infer the rearrangement history, and to determine rates of breakpoint reusage. Scaffolds can be mapped to Muller elements and ordered using syntenic block information. For instance, two scaffolds can be joined together if the genes at the end of the joined scaffolds are adjacent in other species. The large number of gene-order states allows the history of gene-order changes to be recovered even when breakpoints are reutilized. In addition, genes at breakpoint boundaries can be used to reconstruct ancestral states of common ancestors (Ma *et al.* 2006). Thus, the availability of the 12 genomes presents a unique opportunity to develop new tools for the assembly, annotation, and analysis of gene-order information.

We present here an analysis of gene-order data from the 12 Drosophila genomes. A computational approach (Synpipe) is presented for the annotation of orthologous genes among the 12 species of Drosophila that uses conserved synteny information to increase the confidence in homologous gene calls. Synpipe (Bhutkar *et al.* 2006) output helped to assign assembly scaffolds to one of six Muller elements and join contiguous scaffolds together. The inferred ordered scaffolds were largely consistent with the genetic and physical map data (Schaeffer *et al.* 2008). The assembled and ordered scaffolds provided a unique opportunity to study how gene order changes among species of Drosophila by the processes of inversion and transposition, using the study of adjacency information at conserved linkage breakpoints. This study found that 95% of genes exhibit arm-level synteny among the 12 Drosophila species, supporting the established Muller element hypothesis (Sturtevant and Tan 1937; Muller 1940; Sturtevant and Novitski 1941). The remaining genes have likely transposed to a different chromosome (Bhutkar *et al.* 2007b) in one or more species. Inferring chromosomal evolution rates from whole-genome synteny data suggests higher rates than previously reported (Ranz *et al.* 2001; Bartolome and Charlesworth 2006). The analysis of adjacency information reveals that the paracentric inversion rate is higher in Sophophoran species than in the subgenus Drosophila species and that each region with an observed breakpoint across the genus Drosophila gets used 1.5 times.

## MATERIALS AND METHODS

**Inference of syntenic blocks:** Multispecies *in silico* comparison of gene order and rearrangements was performed using Synpipe (Bhutkar *et al.* 2006), a graph-based chaining algorithm that utilizes syntenic block maximization criteria in the presence of genome assembly gaps. Starting with a reference peptide set for a given species and a contig or scaffold assembly for another species, Synpipe provides refined homology data, syntenic block computation, and reciprocal breakpoint annotation. Homology refinement is based on maximizing syntenic block size in the presence of paralogs, assembly gaps, or missing data. Synpipe accommodates for contig and scaffold gaps in the assembly by identifying homologous elements that might either fall in unsequenced assembly gaps or lie on the edges of sequenced segments or on small assembly fragments. The resulting Synpipe Drosophila data set has been used for breakpoint analysis, a comparative study of chromosomal rearrangements between species, multispecies alignment and orthology refinement, and for mapping and orienting scaffolds onto the Drosophila Muller elements or chromosome arms (Schaeffer *et al.* 2008).

One protein isoform, that with the 5′-most translation start site annotation, was chosen for each predicted gene in the reference set. Syntenic blocks were computed as segments of the assembly where orthologs shared the gene order of the reference species while allowing for gaps and localized scrambling due to micro-inversions. Missing orthologs that were predicted to fall in intercontig assembly gaps were assumed not to disrupt a syntenic block. One ortholog per gene was chosen in this analysis (Bhutkar *et al.* 2006) and gene duplication did not factor into this analysis. Localized gene-order scrambling was permitted within syntenic blocks and while merging smaller blocks with larger adjacent blocks. A threshold value of a number of genes (10 genes in this case) was used to allow for localized scrambling of gene order within a syntenic block. Missing genes did not disrupt syntenic blocks as long as two or more of the missing genes did not form small syntenic blocks elsewhere in the genome.

In this study, two reference peptide sets were used for independent analysis: the well-annotated Release 4.3 peptide set for *D. melanogaster* (Crosby *et al.* 2007) and the GLEAN-R (Drosophila 12 Genomes Consortium 2007) predicted gene set for *D. virilis*. The use of two sets provides for synteny computation from a subgenus Sophophora vantage point and a subgenus Drosophila vantage point. Additionally, outgroup species including *An. gambiae* (Holt *et al.* 2002), *Aedes aegypti* (Nene *et al.* 2007), *Apis mellifera* (Hgsc 2006), and *Tribolium castaneum* (Tribolium Genome Sequencing Consortium 2007) were also included in this analysis to glean information on ancestral syntenic relationships, where possible (see supplemental information).

Orthology information was also used to assign assembly scaffolds to the Muller elements within a given species, on the basis of majority hits from a specific Muller element of *D. melanogaster.* These assignments were also used to pinpoint locations of probable genome assembly misjoins, especially in cases where a scaffold had large syntenic blocks belonging to different Muller elements without supporting evidence from other species (Schaeffer *et al.* 2008). These scaffold assignments were used in conjunction with synteny information to infer scaffold order and orientation along chromosome arms in a candidate assembly. This information supplemented experimental analysis using known markers, especially for the order and orientation of small scaffolds without markers (Schaeffer *et al.* 2008).

**Analysis of syntenic block boundaries:** The breakpoints between identified syntenic blocks harbor information about the past inversion history between two species. We define a breakpoint as the nucleotide sequence between two conserved linkage blocks whose adjacent genes are not adjacent in the reference species. A single inversion will be inferred from a pair of breakpoints in a candidate assembly compared to the gene order in the reference set, while accounting for any missing genes. Furthermore, multibreak events or breakpoint reuse events can be inferred compared to the reference order, using a technique similar to a depth first search to search for syntenic block edges that can reconstruct the reference order.

We developed a method called linkage chain analysis to infer the numbers of fixed inversion events between pairs of species. If each gene along a chromosome is labeled from 1 to *n*, then the history of inversion events can be determined from the differences in the gene order among the different species. When a paracentric inversion flips the order of genes on a chromosomal element, the rearrangement leaves a signature of the mutation at the two nucleotide sites of double-strand breakage (Figure 1). The Synpipe analysis will define these breakpoints as the interval between two different syntenic blocks. Simulations assuming a random-breakage model (Nadeau and Taylor 1984) show that the majority of linkage chains will be composed of two breakpoints when the number of genes is large and the number of inversions is small (S. W. Schaeffer, unpublished data). Linkage chains with more than two breakpoints result from multiple inversion events that reuse some of the breakpoint nucleotide sequence intervals (Figure 2). The number of inversions $n_{\mathrm{inv}}$ inferred from a linkage chain of $n_{\mathrm{bp}}$ breakpoints will be $n_{\mathrm{inv}} = n_{\mathrm{bp}} - 1$. One can estimate a reusage statistic (Sankoff and Trinh 2005) $r$ from the $n_{\mathrm{inv}}$ and $n_{\mathrm{bp}}$ as $r = (2n_{\mathrm{inv}})/n_{\mathrm{bp}}$, where $r$ varies from a value of 1 and 2. An $r$ value of 1 indicates that each breakpoint within the chain was used a single time, while an $r$ value of 2 indicates that each breakpoint within the chain was used twice. The overall estimate of $r$ is obtained by summing over all linkage chains for a pairwise comparison of species. *The observation of reusage is not to say that the breakage events occur at the same nucleotide, but occur within the same interval defined by the two genes that flank the region where conserved linkage groups are*

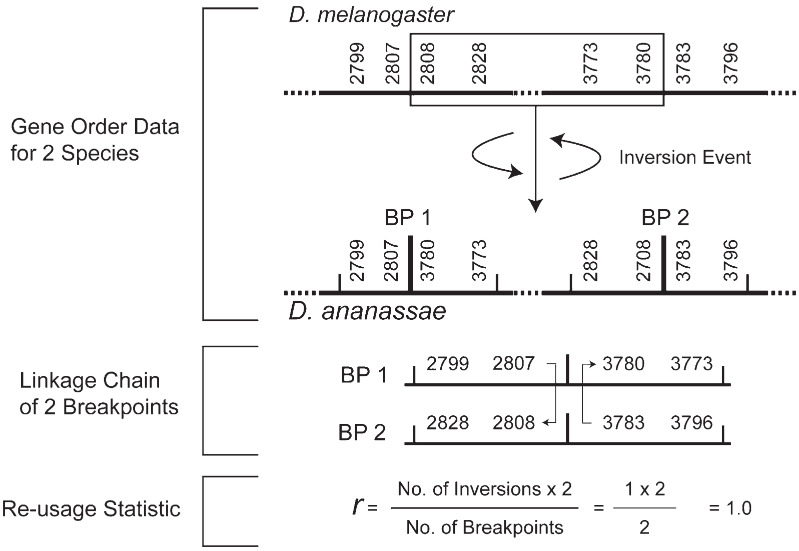

FIGURE 1.—Linkage chain analysis for reciprocal breakpoints generated by a single inversion event. Gene order is shown for two species where *D. melanogaster* has the standard gene order and *D. ananassae* has the rearranged gene order. The box indicates the segment that was inverted. The reusage statistic *r* (SANKOFF and TRINH 2005) of 1.0 estimated from this linkage chain indicates no reusage of these breakpoints.

*broken*. There is precedent for reutilization of breakpoints within *D. pseudoobscura* where the Tree Line and Santa Cruz arrangements share a common breakpoint within the resolution of polytene maps (DOBZHANSKY 1944). In addition, one of the breakpoints that converted the *D. pseudoobscura* Standard chromosome to the Arrowhead arrangement was reused since *D. melanogaster* and *D. pseudoobscura* shared a common ancestor (RICHARDS *et al.* 2005).

The basic procedure for inferring linkage chains between two species requires that genes be sorted according to the gene order of a reference species and then linkage chains are inferred from the altered gene order of target species. Let us say that we are inferring the linkage chain for the *D. melanogaster* and *D. ananassae* species pair; then we can sort the genes according to the *D. melanogaster* gene order and analyze the gene order changes for *D. ananassae*. Linkage chain analysis uses the Synpipe defined syntenic blocks for a chromosome and the associated breakpoints defined at the boundaries of each pair of syntenic blocks. By definition, if there are $n_{\mathrm{sb}}$ syntenic blocks on a chromosome, then there will be $n_{\mathrm{bp}} = n_{\mathrm{sb}} - 1$ breakpoints. Linkage chain analysis proceeds by following the sequence of gene identification numbers at the boundary of the breakpoint. If the gene identification numbers are increasing toward the breakpoint, then the breakpoint adjacent to the $n + 1$th gene is located. If the gene identification numbers are decreasing toward the breakpoint, then the breakpoint adjacent to the $n - 1$th gene is located. In the example, the gene identification numbers adjacent to the first breakpoint in *D. ananassae* (BP 1 in Figure 1) are increasing so the breakpoint adjacent to gene 2808 is located. BP 1 and BP 2 form a complete chain because the boundary genes at the two breakpoints comprise the set of adjacent neighbors in *D. melanogaster.* Both the gene order and the orientation of the genes are used in linkage chain analysis because some syntenic blocks consist of a single gene where expected adjacent genes at the boundary cannot be de-

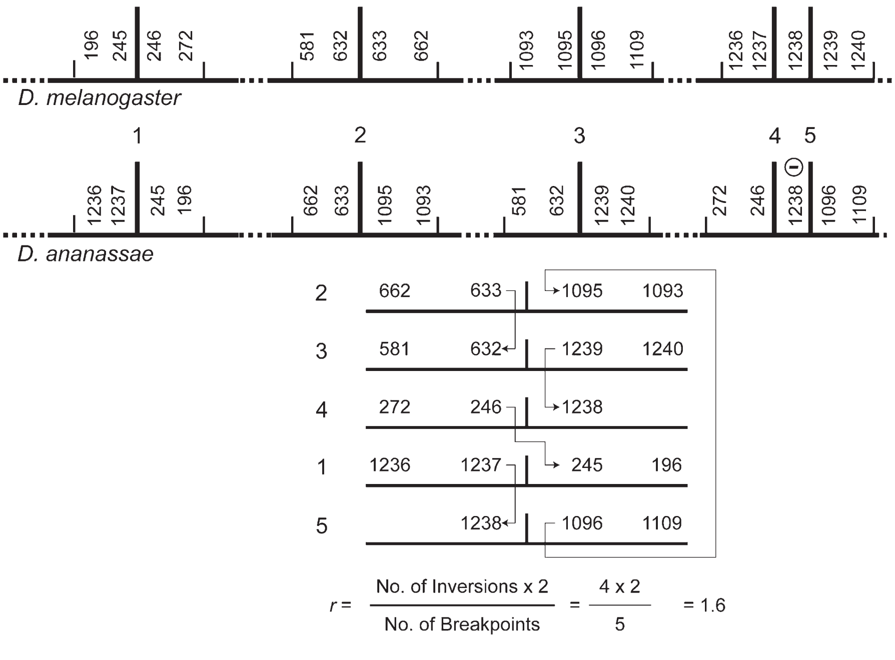

FIGURE 2.—Linkage chain with five breakpoints that was generated by four inversion events. Gene order is shown for two species where *D. melanogaster* has the standard gene order and *D. ananassae* has the rearranged gene order. The reusage statistic (SANKOFF and TRINH 2005) *r* of 1.6 estimated from this linkage chain indicates that 60% or three breakpoints were used more than once.

termined. The orientation information provides directional information about the adjacent neighbor of a single-gene block.

Linkage chain analysis was performed on gene-order data for the five major Muller elements (A–E) for comparisons of the seven most divergent Drosophila species, *D. melanogaster, D. ananassae, D. pseudoobscura, D. willistoni, D. virilis, D. mojavensis*, and *D. grimshawi*. Gene order has changed extensively among these species compared to members of *D. melanogaster* subgroup where <30 rearrangement events have accumulated among these species and have been considered elsewhere (Ranz *et al.* 2007; York *et al.* 2007). *D. persimilis* was also excluded from the analysis because the genomic sequence differs from the *D. pseudoobscura* sequence by four inversion differences (Moore and Taylor 1986) and the inversion events have been analyzed elsewhere (Machado *et al.* 2007; Noor *et al.* 2007).

The gene-order data from the seven Drosophila species were adjusted before linkage chain analysis by taking the intersection of the syntenic genes. If $G_1$–$G_7$ represent the gene-order list in species 1–7, where $G_1$ is *D. virilis*, $G_2$ is *D. mojavensis*, $G_3$ is *D. grimshawi*, $G_4$ is *D. willistoni*, $G_5$ is *D. pseudoobscura*, $G_6$ is *D. ananassae*, and $G_7$ is *D. melanogaster*, then the set of genes used for the linkage chain analysis $G$ is the intersection of genes in the different species or

$$G = \cap_{i=1}^{7} G_i.$$

Genes were removed from the joint set of genes across the seven species if (1) a gene was absent due to deletion events in at least one species, (2) a gene resulted from a duplication not in the common ancestor to all lineages, (3) a gene was not annotated because it was in an assembly gap, or (4) a gene was transposed to another chromosome in at least one species. Genes were also removed if they were transposed within a chromosome in at least one species. Transposed genes were identified as genes whose two adjacent neighbors were adjacent to each other. For instance, if the neighbors of gene 2088, genes 2087 and 2089, were adjacent at another chromosomal location, then gene 2088 was removed because we assumed that its rearrangement resulted from a transposition event rather than an inversion event. Finally, embedded or overlapping genes were removed from the joint data set because this organization of genes tended to be highly conserved, yet inclusion of these genes tended to artificially alter the order of the gene indexes. This can occur due to possible absence of upstream exons of the surrounding gene (as a result of assembly gaps). This can also occur because multiple embedded genes are numbered from the first exon of each gene model along the sequence of the reference species. If a segment with embedded genes is reversed in a target species, then one can get artificial rearrangements when the embedded genes of the target species are indexed in the reverse order.

The algorithm to perform the linkage chain analysis is simpler when the list of gene identifiers is a sequential list of integers, where the number of genes in the list is equal to the maximum index value. The gene-order lists in set $G$ were converted to a new set of gene identifiers in set $H$, such that $H_{\mathrm{i}}$ has the new sequential list of gene identifiers for $G_{\mathrm{i}}$. If $H$ is sorted according to the *D. melanogaster* identifiers in $H_7$, then the gene-order lists in $H_1$–$H_6$ will reflect the rearrangement information when *D. melanogaster* is the reference gene order. Analysis of $H_1$–$H_6$ along with the associated sign information is then used in the linkage chain analysis between *D. melanogaster* and the six other species to estimate the number of breakpoints, the chain length distribution, and the numbers of inversions. This analysis is done using each species for the reference gene order. Pairwise inversion distances estimated in this manner from linkage chain analysis are reciprocal; that is, the number of breakpoints, the chain distribution, and the number of inversions are the same no matter which species is used for the reference gene order in the comparison.

**Analysis of syntenic block conservation among Drosophila species:** In addition to determining syntenic blocks between pairs of species, blocks conserved across larger sets of species were also computed. Eliminating *D. persimilis*, *D. sechellia*, and *D. simulans* (for assembly quality and due to representation from other evolutionarily close species), conserved blocks across nine species were computed. On the basis of each species' gene orthology with *D. melanogaster*, a dynamic programming methodology was employed to determine sets of conserved order. To allow low levels of scrambling within syntenic blocks, genes in a synteny block were sorted according to the *D. melanogaster* order. Gene deletions in one or more species, however, were considered as block boundaries. Using a dynamic programming approach provides a fast method to determine segments of conserved gene order between two species, the results of which are used to compute conserved segments with a third species. This procedure is continued until all nine species have been processed. These conserved blocks were further analyzed using *D. melanogaster* gene expression patterns during embryogenesis (Tomancak *et al.* 2002) (http://www.fruitfly.org). Time-course data from three replicates (Tomancak *et al.* 2002) over a 12-hr window were averaged for each 1-hr window. Correlation coefficients were derived for these time-course data between pairs of genes within a block. Additionally, we studied tissue expression specificity for various genes within a block across different embryonic developmental stages (Tomancak *et al.* 2002). Functional clustering of genes was also explored in conserved blocks, using gene ontology (Gene Ontology Consortium 2001) (www.geneontology.org) data for Drosophila genes (http://www.flybase.org). We also analyzed syntenic blocks for transposable-element insertion sites, for *P-transposable* elements (Bellen *et al.* 2004; Spradling *et al.* 1995) and natural transposons (Kaminker *et al.* 2002), in *D. melanogaster* (http://www.flybase.org, www.fruitfly.org). It is known that *P* elements show nonrandom insertion patterns (O'Hare and Rubin 1983), and we wished to test whether large syntenic blocks might be less likely to be interrupted by transposable-element insertions.

**Derivation of a rearrangement phylogeny:** Two approaches were used to derive phylogenetic relationships and rearrangement estimates for these genomes. The first approach uses pairwise numbers of inversions among all species from the linkage chain analysis to infer the phylogenetic relationships of the Drosophila species. The neighbor-joining method implemented in MEGA 3.0 (Kumar *et al.* 2004) was used to infer the tree topology and branch lengths (Saitou and Nei 1987). The pairwise numbers of breakpoints were also used in the phylogenetic analysis to infer the branch length-specific breakpoint numbers to determine whether most breakpoint reusage occurred in terminal or internal branches of the phylogeny.

Another approach based on the neighboring gene pair (NGP) method (Bhutkar *et al.* 2007a) was used to infer a rearrangement phylogeny for the genus Drosophila and to determine specific chromosomal disruptions between known genes on various branches of the phylogeny. The description of the NGP method (Bhutkar *et al.* 2007a) used a subset of the available Drosophila species as a case study. We have expanded the set of species and we report on inversion estimates for this expanded set and on some associated challenges. In addition to accounting for the macro-inversions responsible for making large-scale changes to gene order, NGP takes into account

**TABLE 1**

**Synteny statistics (with respect to the *D. melanogaster* gene order)**

| Species | Estimated time since most recent common ancestor with *D. melanogaster* A and B (MY) | No. of synteny blocks | No. of genes in synteny blocks | Maximum synteny block size (no. genes) | Average synteny block size (no. genes) | No. of singleton genes on same Muller element |
|---|---|---|---|---|---|---|
| *D. sechellia* | 5.4, 2.3 | 42 | 13,378 | 1,834 | 318.52 | 0 |
| *D. simulans* | 5.4, 2.3 | 139 | 11,851 | 1,075 | 85.26 | 9 |
| *D. yakuba* | 12.6, 6.1 | 114 | 13,175 | 763 | 115.57 | 1 |
| *D. erecta* | 12.6, 6.1 | 63 | 13,403 | 972 | 212.75 | 5 |
| *D. ananassae* | 44.2, 20 | 695 | 12,660 | 138 | 18.22 | 100 |
| *D. pseudoobscura* | 54.9, 24.3 | 908 | 11,932 | 109 | 13.14 | 154 |
| *D. persimilis* | 54.9, 24.3 | 962 | 11,993 | 109 | 12.47 | 151 |
| *D. willistoni* | 62.2, 36.3 | 1,430 | 11,670 | 88 | 8.16 | 383 |
| *D. virilis* | 62.9, 39.2 | 1,297 | 11,707 | 81 | 9.03 | 305 |
| *D. mojavensis* | 62.9, 39.2 | 1,312 | 11,509 | 73 | 8.77 | 328 |
| *D. grimshawi* | 62.9, 39.2 | 1,337 | 11,217 | 78 | 8.39 | 351 |

Synteny statistics (with respect to the *D. melanogaster* gene order) utilizing assembly scaffolds that were anchored to chromosome arms in various species using experimental and computational data are shown (Schaeffer *et al.* 2008). Synteny blocks were not artificially broken up by scaffold breaks and were inferred to be continuous when adjacent scaffolds permitted it. The number of synteny blocks and their composition (number of genes and maximum and average size) correspond to the phylogenetic distribution of these species. The number of singleton genes (isolated from their *D. melanogaster* neighbors) found on corresponding Muller elements in various species increases with evolutionary distance from *D. melanogaster*—presumed to largely be a result of paracentric inversions. Approximate divergence estimates are from earlier studies: A, Tamura *et al.* (2004); and B, Russo *et al.* (1995). We also used a time-independent calibration method for inversion rate estimates. See materials and methods for a discussion regarding selection of approximate divergence times.

micro-inversions that cause localized order and orientation changes in one or more genes. This method opens a window into fine-scale rearrangements where the syntenic blocks could be maintained between genomes, despite low-level scrambling of gene order and orientation involving a small number of genes. In fact, even single-gene inversions are tracked by this method. The NGP algorithm uses a two-stage tree walkthrough algorithm and allows for genome-scale data sets to be evaluated in a matter of minutes, which is highly suited to our data set of multiple species with >13,000 homologous genes in each species. The basic motivation behind this technique is the assumption of parsimony where the likelihood of the same inversion event taking place independently in two disjoint lineages is low. Consequently, pairs of neighboring genes where their adjacency and mutual orientation are conserved in distant species are assumed to have existed as an adjacent pair at their common ancestor. In other words, the probability of an inversion creating an identical NGP in a different species is assumed to be low. This method also serves as a technique for phylogenetic reconstruction based on maximization of shared pairs unique to a group or cluster ("exclusively shared NGPs"). We used this method to compute rearrangement counts (involving inferred macro- and micro-inversions) along each evolutionary path in the phylogeny.

**Approximate divergence time estimates:** We used two primary sources for approximate divergence time estimates (Russo *et al.* 1995; Tamura *et al.* 2004) in addition to a time-independent calibration method. The first set of estimates based on Russo *et al.* is derived from a study of the coding region of the *Adh* gene across species. The split between the Drosophila and the Sophophora subgenera is estimated to have taken place ~40 MYA. Divergence estimates for other species from *D. melanogaster,* based on this study, are listed in Table 1. Russo *et al.* do not provide a direct estimate for *D. ananassae*. They estimate the divergence between the *D. montium* subgroup and *D. melanogaster* to be ~12.7 MY. Although there are some reports to the contrary (Yang *et al.* 2004), a majority of the other studies estimate that *D. ananassae* branched off prior to the *D. montium* subgroup and after the *D. obscura* group (Pelandakis *et al.* 1991; Pelandakis and Solignac 1993; Clark *et al.* 1998; Harr *et al.* 2000; Goto and Kimura 2001; Tamura *et al.* 2004). We have chosen an approximate estimate of 20 MY for the divergence between *D. melanogaster* and *D. ananassae* on the basis of these observations (Table 1).

The second set of divergence time estimates is based on Tamura *et al.* (2004). This study used a large set of genes (176) for sequence comparisons to derive a genomic mutation clock. Approximate divergence estimates from this study are listed in Table 1. Both sets of divergence estimates (Russo *et al.* 1995; Tamura *et al.* 2004) show some accordance with various biogeographic data (Beverley and Wilson 1984; Powell and Desalle 1995). We have used both sets of estimated divergence times to compute evolutionary rearrangement rates, wherever possible.

We also used a time-independent method to calibrate the inversion rate. The inversion rates on tree branches were calibrated relative to the number of second-position changes per second-position site (second-position tree, S. Kumar, Arizona State University, unpublished results).

**Computation of chromosomal evolutionary rates:** We used the number of inferred synteny blocks for various species compared to *D. melanogaster* and the estimated divergence times between species (Russo *et al.* 1995; Tamura *et al.* 2004) to compute estimates of chromosomal evolutionary rates. These rearrangement rates were calculated in units of disruptions per megabase per million years (Ranz *et al.* 2001). We also used an alternate method based on the results of the NGP algorithm to compute estimates for rearrangement rates. The NGP approach gives us chromosomal breakpoint estimates along each line of descent from the root of the genus

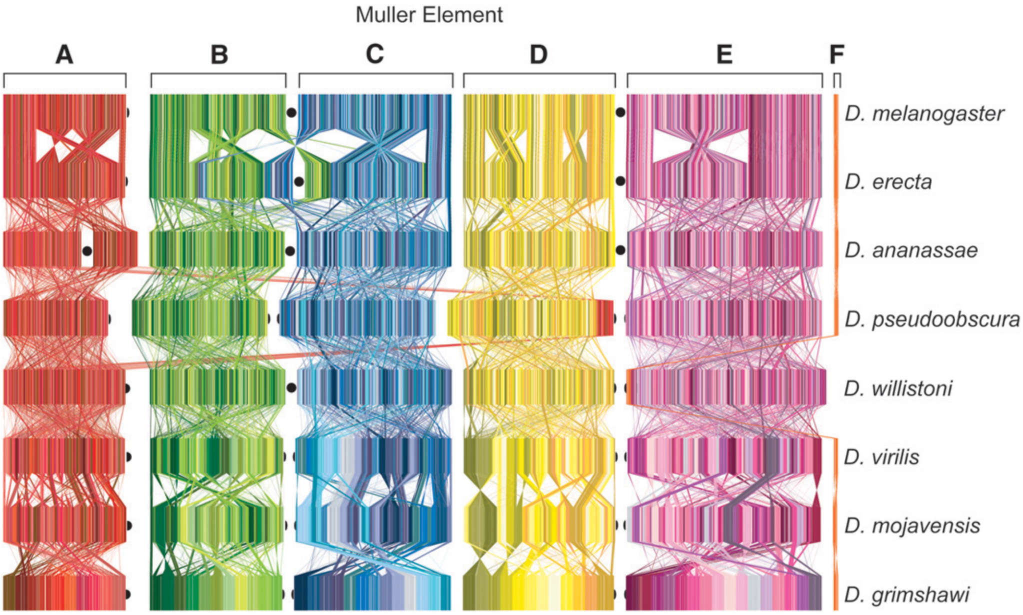

Figure 3.—Plot of genome rearrangement for the six Muller elements among eight species of Drosophila. Each vertical line represents a single gene and the lines that connect the genes between the species help to determine the movement of the genes within and between chromosomal arms among the different species. The centromeres are indicated with solid half and whole circles. Each Muller element is shown in shades of a single color: A, red; B, green; C, blue; D, yellow; E, magenta; F, orange. Blocks of genes in *D. grimshawi* are arbitrarily colored within each Muller element. The blocks of genes in *D. grimshawi* do not correspond to syntenic blocks, but are presented as a visual heuristic to help observe shuffling of genes among the eight species. Some rearrangements might not be visible due to the compressed scale (on Muller F, for example).

Drosophila tree. Using the time estimated for the divergence of the subgenus Sophophora from the subgenus Drosophila, rearrangement rates are calculated in units similar to those of the other method.

**Analysis of breakpoint sequences:** We examined breakpoint sequences for the presence of repetitive sequences that might be responsible for rearrangement hot spots. The methods used by Richards *et al.* (2005) were used to determine if breakpoint sequences in the seven most divergent species had repetitive elements. Breakpoint regions defined by linkage chain analysis were assembled in species-specific BLAST databases and each breakpoint sequence was compared to all other breakpoint sequences with BLASTN (Altschul *et al.* 1997), using an *E*-value of 1 × $10^5$. For each breakpoint, we estimated the fraction of breakpoints that were matched in the BLAST search to derive the breakpoint match distribution for the breakpoints. Breakpoints that had a higher degree of interbreakpoint matching provide a measure of repetitive sequences within the sequence that was involved in the rearrangement.

## RESULTS

**Overview of the rearrangement process:** Figure 3 is a graphical representation of how gene order has been shuffled among eight of the most divergent species of Drosophila. Each gene within each chromosome is represented as a single line and blocks of 150–200 genes within *D. grimshawi* are colored to help show the rearrangement of genes. The six chromosomal arms are indicated with a different hue. The small fraction of single positionally relocated genes are not shown. Several features are apparent. First, the majority of rearrangements take place within a chromosomal arm, although there are several major exceptions. There has been a pericentric inversion between Muller B and C in *D. erecta* that is shared with *D. yakuba* (Lemeunier and Ashburner 1976; Ranz *et al.* 2007). Genes from Muller A in *D. pseudoobscura* have moved from the left arm of the X to the right arm where Muller D genes reside (Segarra *et al.* 1995). Muller F of *D. willistoni* has fused to Muller E (Papaceit and Juan 1998). Second, gene order has undergone far more rearrangement in the Sophophoran than in the Drosophila subgenus. The Drosophila species tend to maintain homogeneous blocks of genes seen as the maintenance of color shade blocks and there is less evidence for rearrangement (fewer crossed lines between species), while the Sophophora have undergone extensive rearrangement (extensive crossed lines between species).

**Inference of syntenic blocks:** Cross-species synteny analysis was performed using Synpipe (Bhutkar *et al.* 2006). The initial analysis was based on the annotated protein set of the reference species, *D. melanogaster.* The set of orthologs assigned for the 13,733 release 4.3 *D. melanogaster* euchromatic genes (Crosby *et al.* 2007) was identified in each species, using Synpipe's synteny maximization criterion (see supplemental materials for complete orthology placements). Clear 1:1 high-confidence orthologs without collisions (*i.e.*, overlapping placements with paralogs or duplications) were first identified. This set was extended to include additional genes on the basis of results of processing to untangle collisions, where possible, to maximize the size of syntenic blocks. For example, 12,874 *D. melanogaster* genes had high-confidence noncollision Synpipe orthology placements in *D. sechellia,* 11,653 in *D. pseudoobscura,* and 10,971 in *D. grimshawi* (supplemental Table S1). Processing collisions on the basis of synteny evidence increased these numbers to 13,683 orthologous placements in *D. sechellia,* 12,905 in *D. pseudoobs-*

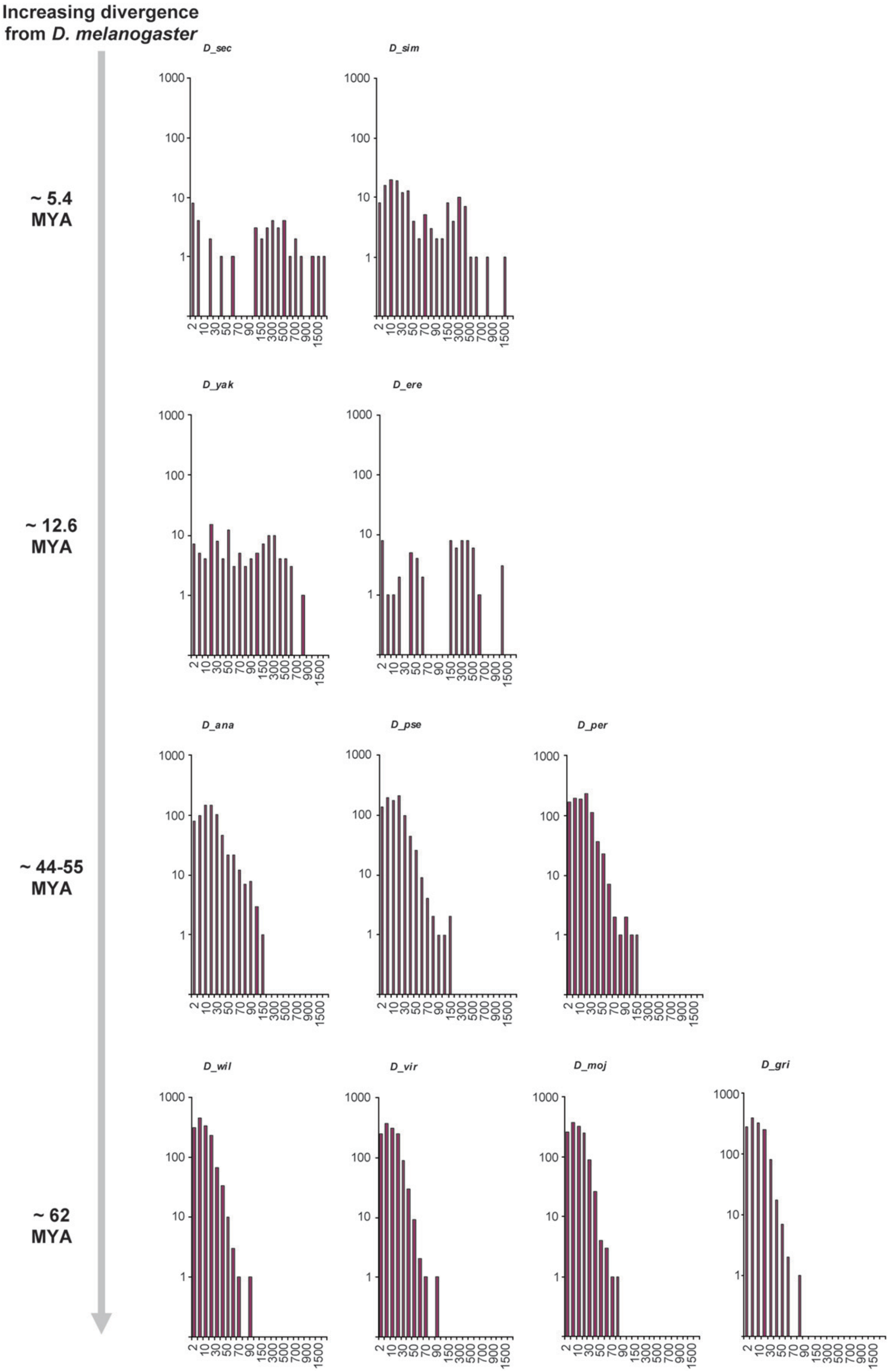

FIGURE 4.—Distribution of synteny blocks across various species based on the *D. melanogaster* euchromatic gene order. Species are shown in increasing evolutionary distance from *D. melanogaster*. The numbers to the left denote the approximate time range since a group of species shared most recent common ancestry with *D. melanogaster* (TAMURA *et al.* 2004). The graph for each species

*cura*, and 12,494 in *D. grimshawi* (supplemental Table S1).

On the basis of majority hits of these orthologs from a Muller element of the reference species (*D. melanogaster*) to a scaffold of a candidate genome assembly, we were able to assign assembly scaffolds to Muller elements in various species (Schaeffer *et al.* 2008) (see materials and methods). Other than a few small scaffolds, scaffold to Muller element assignment was largely unambiguous and was performed with a high level of confidence. As reported by other Drosophila comparative studies using evidence from a few genomes (Richards *et al.* 2005), a majority of the orthologous genes are found on the same Muller element, albeit in different syntenic blocks, across all species. We extended this analysis across the 12 Drosophila genomes and find this to be true across the Drosophila phylogeny. We find that ~95% of all the *D. melanogaster* genes found in at least one other species are located on the same Muller element in all those species (*i.e.*, arm-level synteny is maintained) (Figure 3).

We then studied synteny relationships across species using the *D. melanogaster* annotation as the reference set. Our results showed a distribution of syntenic blocks that was in agreement with the phylogeny. Using the genome scaffold assembly as is, without anchoring scaffolds to chromosome arms in various species, provided an initial set of numbers where synteny blocks were terminated at scaffold ends (supplemental Table S2). Between 55% (*D. ananassae*) and 77% (*D. erecta*, *D. simulans*) of the genome assembly was covered by synteny blocks (of size greater than one gene) with respect to *D. melanogaster* (supplemental Figure S1). To refine this synteny data set, only scaffolds anchored to chromosome arms using computational or experimental evidence (Schaeffer *et al.* 2008) were used and "scaffold joins" were inferred to determine continuous synteny blocks across genome scaffold breaks, where possible. This analysis reveals between 42 (*D. sechellia*) and 1430 (*D. willistoni*) syntenic blocks across various species on the basis of the *D. melanogaster* gene order (Table 1, Figure 4). The sizes of the largest syntenic blocks observed in various species fall in between a block of 1834 genes seen in *D. sechellia* (average of 319 genes/block) and another of 73 genes seen in *D. mojavensis* (average of ~9 genes/block). Additionally, compared to the *D. melanogaster* order, evolutionarily distant species show a higher number of same-arm singletons (single genes isolated from their *D. melanogaster* neighbors on the same Muller element) based on rearrangements in one lineage or the other. For example, the four species of the *melanogaster* subgroup have an average of <4 such cases compared to *D. grimshawi's* 351 and *D. willistoni's* 383 same-arm singletons. Between 50 and 75% of the genome assembly sequence mapped to various Muller elements (Schaeffer *et al.* 2008) in each species was inferred to be in syntenic segments (Table 1).

Muller element dot plots using synteny data show phylogenetic correspondence of syntenic blocks (Figure 5). In addition to the *D. melanogaster* protein set, we used the *D. virilis* GLEAN-R consensus annotation protein set (Drosophila 12 Genomes Consortium 2007) (rana.lbl.gov/drosophila) as an additional reference set to undertake synteny analysis. Synteny maps using both sets demonstrated phylogenetic synergy. In the case of the *D. virilis* reference set, *D. mojavensis* and *D. grimshawi* had fewer syntenic block disruptions (on the order of 500–650 blocks in these species, respectively) whereas the subgenus Sophophora species had on the order of 1500 blocks, all being evolutionarily equidistant from *D. virilis.* We also estimated chromosomal rearrangement rates across species, or the rate of syntenic disruptions, using these inferences of cross-species synteny (see discussion).

In addition to the genus Drosophila, syntenic blocks were inferred from a number of outgroup species including two mosquitoes *A. aegypti* and *An. gambiae*, the silkworm *Bombyx mori*, the honey bee *Apis mellifera*, and the red flour beetle *T. castaneum* (see supplemental information). Using either *D. melanogaster* proteins as the reference set or *An. gambiae* proteins as the reference set reveals very few large conserved blocks, all <15 genes in length, and provides insight into the ancestral colocation of genes on chromosome arms that might be the primary contributor to extant chromosomes (Bhutkar *et al.* 2007b). Additionally, such outgroup comparisons also support the inference of universally conserved syntenic blocks (see below).

**Inference of inversion events from analysis of syntenic block boundaries:** The inversion and breakpoint distances estimated from the linkage chain analysis of seven pairs of Drosophila species are shown in Table 2. These estimates are summed over all linkage chains for each pairwise comparison. The majority of linkage chains for each comparison of two species had a length of two breakpoints (Figure 6). The mean percentage of chains of length two was 68.3 and the range varied from a minimum of 51.9 to a maximum of

shows the size of derived synteny blocks (in number of genes) on the horizontal axis (nonlinear scale showing size buckets: 2, 5, 10–100 interval 10, 150, 200–1000 interval 100, 1500, 2000) and the number of such blocks (log scale) on the vertical axis. These distributions show greater fragmentation of the genome compared to *D. melanogaster* with increasing evolutionary divergence. Additionally, species equidistant from *D. melanogaster* might show different degrees of fragmentation (as seen in *D. yakuba vs. D. erecta*, where *D. yakuba* exhibits greater fragmentation). These distributions are based on genome assembly scaffolds that were anchored to chromosome arms, where synteny blocks were allowed to span across scaffold breaks wherever possible. Species: *D_sec*, *D. sechellia*; *D_sim*, *D. simulans*; *D_yak*, *D. yakuba*; *D_ere*, *D. erecta*; *D_ana*, *D. ananassae*; *D_pse*, *D. pseudoobscura*; *D_per*, *D. persimilis*; *D_wil*, *D. willistoni*; *D_vir*, *D. virilis*; *D_moj*, *D. mojavensis*; *D_gri*, *D. grimshawi.*

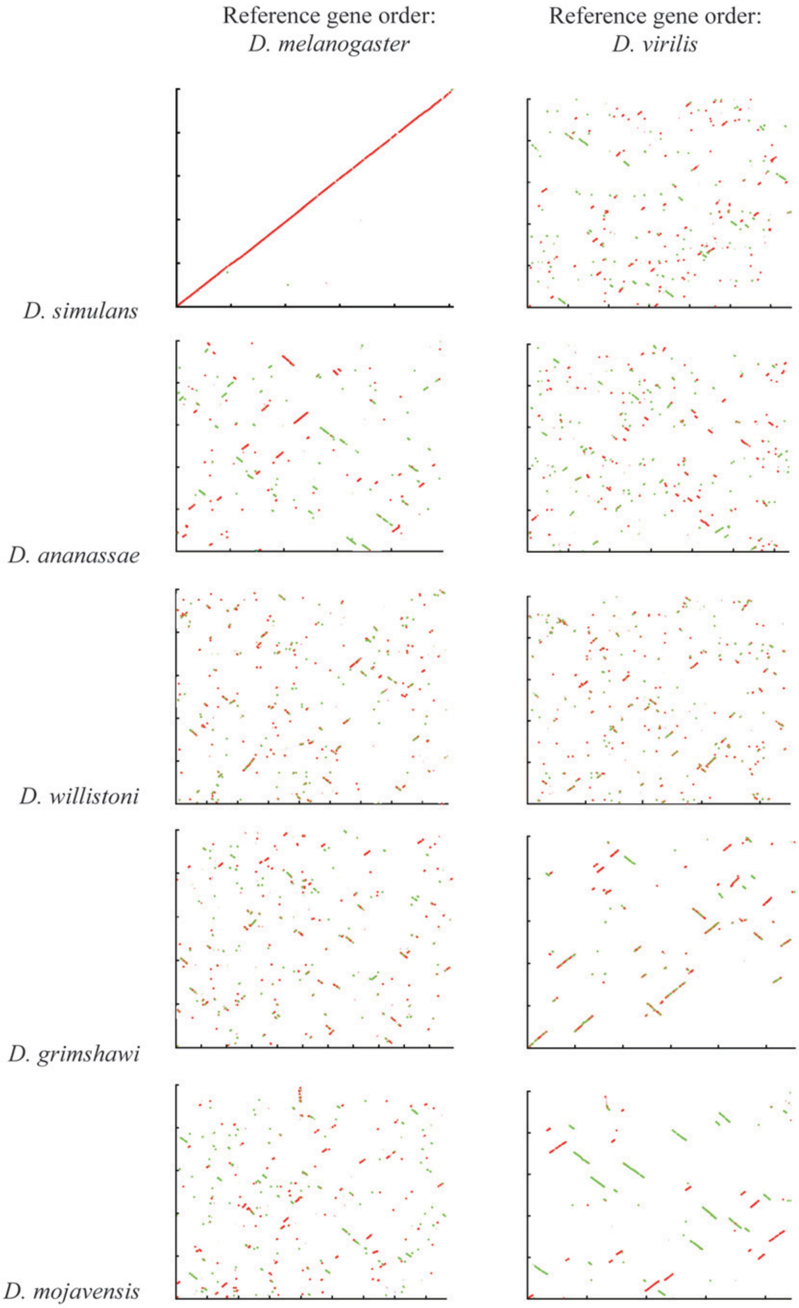

Figure 5.—Paracentric inversions correlated with phylogenetic distance: Muller element B dot plots based on *D. melanogaster* (left column) and *D. virilis* (right column) gene orders. Species are shown in increasing evolutionary distance from *D. melanogaster* from top to bottom (reverse for *D. virilis*). Evolutionarily distant species show greater scrambling of gene order due to independent paracentric inversions across various lineages. For example, *D. simulans* shows greater arm-level synteny with *D. melanogaster* than *D. virilis.* Similarly, *D. mojavensis* exhibits the opposite behavior with respect to the reference species. Compressed scale is used to depict the chromosome arm.

88.8 (Table 3). The mean estimate for the breakpoint reusage statistic $r$ (Sankoff and Trinh 2005) is 1.47 with minimum and maximum values of 1.05 and 1.80. The percentage of two-break chains is lower for comparisons of Drosophila species *vs.* Sophophora species (Table 3). Estimates of breakpoint reusage are greatest when species from the Drosophila and Sophophora subgenera are compared (Table 3).

We asked whether the observed breakpoint reusage ($r$) values were within the range of expected values when breakpoints are introduced by two Poisson events. We simulated this process by placing genes 1–$n$ on a chromosome in the +1 orientation and allowing $n_{\text{inv}}$ inversion events. Each inversion event was simulated by choosing two sites on the chromosome sampled from a uniform distribution and reversing the order of the intervening genes. The orientation of the inverted genes was reversed either +1 to −1 or −1 to +1, depending on the state of the gene at the time of the rearrangement. After the $n_{\text{inv}}$ events were complete, we used linkage chain analysis to estimate the number of breakpoints, the number of inversions, and the reusage statistic $r$ (Sankoff and Trinh 2005). We performed 2500 rearrangement simulations for each set of parameters (gene number, inversion number) and the probability of the observed $r$ value was determined from the rank-ordered set of simulated $r$ values. We used the sequential Bonferroni method to correct for multiple tests (Rice

**TABLE 2**

**Inversion and breakpoint distances based on comparisons of seven pairs of Drosophila species for the five major Muller elements**

| | *D_mel* | *D_ana* | *D_pse* | *D_wil* | *D_vir* | *D_moj* | *D_gri* |
|---|---|---|---|---|---|---|---|
| | | | Muller A | | | | |
| *D_mel* | | 156 | 194 | 393 | 308 | 316 | 341 |
| *D_ana* | 227 | | 205 | 409 | 319 | 325 | 359 |
| *D_pse* | 246 | 261 | | 390 | 288 | 299 | 325 |
| *D_wil* | 439 | 455 | 442 | | 426 | 427 | 451 |
| *D_vir* | 362 | 379 | 343 | 484 | | 42 | 99 |
| *D_moj* | 366 | 381 | 347 | 485 | 68 | | 111 |
| *D_gri* | 396 | 413 | 369 | 504 | 146 | 156 | |
| | | | Muller B | | | | |
| *D_mel* | | 94 | 111 | 238 | 179 | 179 | 187 |
| *D_ana* | 148 | | 121 | 244 | 182 | 184 | 196 |
| *D_pse* | 169 | 169 | | 233 | 170 | 170 | 188 |
| *D_wil* | 331 | 326 | 309 | | 240 | 245 | 261 |
| *D_vir* | 244 | 242 | 234 | 318 | | 24 | 57 |
| *D_moj* | 248 | 247 | 234 | 319 | 40 | | 59 |
| *D_gri* | 250 | 255 | 246 | 334 | 87 | 93 | |
| | | | Muller C | | | | |
| *D_mel* | | 88 | 159 | 255 | 203 | 204 | 217 |
| *D_ana* | 132 | | 162 | 265 | 216 | 219 | 230 |
| *D_pse* | 230 | 235 | | 284 | 241 | 245 | 256 |
| *D_wil* | 351 | 359 | 387 | | 275 | 279 | 288 |
| *D_vir* | 281 | 303 | 323 | 378 | | 37 | 48 |
| *D_moj* | 286 | 312 | 331 | 379 | 62 | | 53 |
| *D_gri* | 301 | 319 | 339 | 392 | 77 | 93 | |
| | | | Muller D | | | | |
| *D_mel* | | 67 | 122 | 266 | 180 | 182 | 204 |
| *D_ana* | 104 | | 128 | 274 | 187 | 189 | 209 |
| *D_pse* | 176 | 183 | | 280 | 188 | 190 | 210 |
| *D_wil* | 361 | 365 | 372 | | 304 | 302 | 322 |
| *D_vir* | 259 | 263 | 270 | 402 | | 10 | 38 |
| *D_moj* | 260 | 264 | 269 | 398 | 19 | | 38 |
| *D_gri* | 294 | 296 | 303 | 421 | 71 | 69 | |
| | | | Muller E | | | | |
| *D_mel* | | 119 | 157 | 274 | 234 | 235 | 253 |
| *D_ana* | 187 | | 161 | 278 | 236 | 236 | 255 |
| *D_pse* | 233 | 230 | | 272 | 231 | 233 | 247 |
| *D_wil* | 354 | 357 | 351 | | 300 | 305 | 318 |
| *D_vir* | 319 | 311 | 313 | 386 | | 42 | 64 |
| *D_moj* | 320 | 314 | 314 | 381 | 71 | | 68 |
| *D_gri* | 339 | 332 | 333 | 407 | 94 | 102 | |

Inversion distances are shown above the diagonal and the breakpoint distances are shown below the diagonal. Species: *D_mel*, *D. melanogaster*; *D_ana*, *D. ananassae*; *D_pse*, *D. pseudoobscura*; *D_wil*, *D. willistoni*; *D_vir*, *D. virilis*; *D_moj*, *D. mojavensis*; *D_gri*, *D. grimshawi*.

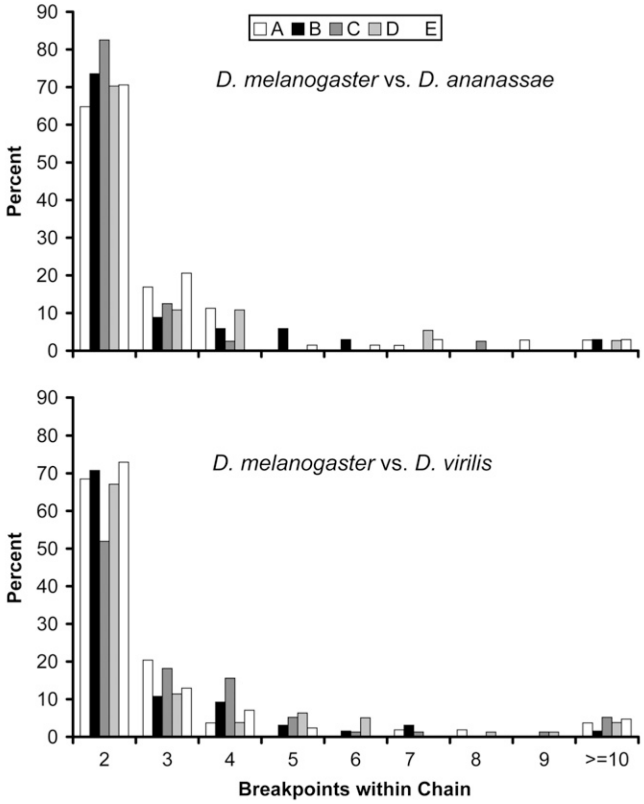

FIGURE 6.—Linkage chain distribution for two pairwise comparisons, *D. melanogaster vs. D. ananassae* and *D. melanogaster vs. D. virilis*. The numbers of chains are summed over Muller A–E for both comparisons.

1989). For all simulations, the observed reusage value was significantly greater than that expected from a model that assumes all sites are free to break given a uniform distribution (supplemental Table S3). In fact, the majority of the linkage chains were of length two, reflecting single inversion events. This suggests that double-strand breaks are not introduced as independent events in this simple model and that some noncoding sequences are hot or cold spots for rearrangement events.

We tested whether a hot- and cold-spot model for rearrangement could better explain the reusage values. Cytogenetic data suggest that particular sites on Drosophila chromosomes are prone to reusage on the basis of studies within and between species (Lemeunier and Ashburner 1976; Olvera *et al.* 1979). For these simulations, we assumed that only certain sites on the chromosome are able to rearrange, so-called hot spots. Alternatively, one can think of this model as avoiding breakage at some sites because of a functional reason such as maintaining suites of coordinately expressed genes (Stolc *et al.* 2004). These simulations used the same strategy as above except that initially a set of $n_{\mathrm{bp}}$ breakpoint sites was uniformly sampled from the total number of sites on the chromosome. For each inversion, two of the $n_{\mathrm{bp}}$ sites were chosen using a uniform distribution for the inversion event and the intervening genes were reversed. Once the $n_{\mathrm{inv}}$ inversions were performed, linkage chain analysis was used to estimate the number of breakpoints, the number of inversions, and the reusage statistic $r$ (Sankoff and Trinh 2005). We determined the probability of the observed $r$ value using the same approach as that for the simulations above.

**TABLE 3**

**Summary of linkage chain analysis**

| | Sophophora | Drosophila | Soph. *vs.* Dros. | Overall |
|---|---|---|---|---|
| Chain 2 BP % | 66.0 | 76.2 | 68.0 | 68.3 |
| Chain 2 BP min % | 51.9 | 66.0 | 51.9 | 51.9 |
| Chain 2 BP max % | 79.7 | 88.9 | 80.0 | 88.9 |
| BP reusage *r* | 1.47 | 1.23 | 1.53 | 1.47 |
| BP reusage min *r* | 1.27 | 1.05 | 1.39 | 1.05 |
| BP reusage max *r* | 1.80 | 1.42 | 1.79 | 1.80 |

Chain 2 BP, linkage chains with two breakpoints; BP reusage *r*, breakpoint reusage statistic *r* of Sankoff and Trinh (2005).

The breakpoint hot-spot model was rejected as an explanation for observed reusage values in 16 of 105 pairwise comparisons. Twelve of the 16 cases rejected the hot-spot model because breakpoint reusage was higher than expected under the model. All of these cases tended to involve comparisons of Muller A, tended to compare species between the Sophophora and Drosophila lineages, or tended to compare *D. willistoni* with other species. Four of the 16 cases rejected the hot-spot model because breakpoint reusage was less than expected. These deficiencies in breakpoint reusage were found in comparisons of closely related species on Muller elements C, D, and E.

Breakpoints could appear to be reutilized if the probability of breakage is directly related to the length of the breakpoint sequence. Because the specific site of a breakpoint could be anywhere between the boundary genes, larger breakpoint sequences could have a higher probability of being reutilized than smaller breakpoint intervals. We tested this possibility by comparing the lengths of breakpoint segments for breakpoints in linkage chains of two *vs.* linkage chains greater than two. Reutilized breakpoints would be in linkage chains greater than two. For this analysis, we considered pairwise linkage chain analyses that had larger numbers of chains (*D. melanogaster vs. D. willistoni*, *D. ananassae vs. D. willistoni*, *D. pseudoobscura vs. D. willistoni*, *D. willistoni vs. D. grimshawi*, *D. grimshawi vs. D. willistoni*, *D. mojavensis vs. D. willistoni*, and *D. virilis vs. D. willistoni*). In each case, the breakpoint distances were estimated for the first species in the pairwise comparison. Data from Muller elements C, D, and E failed to find a significant difference in the length of breakpoint intervals for linkage chains of two *vs.* those greater than two (S. W. Schaeffer, unpublished data). These data provide no evidence to suggest that breakpoint sequence length affects reutilization.

**Syntenic block conservation among multiple Drosophila species:** Blocks of synteny conserved across nine Drosophila species were inferred. This analysis examined relaxed syntenic blocks that allowed localized scrambling of gene order and orientation within blocks (see materials and methods). A total of 2155 blocks of ≥2 genes were identified, 1296 (60%) of which contain ≥3 genes and 613 (28%) of which contain ≥5 genes (Figure 7, supplemental information). The largest block (31 genes; *CG6413–CG13623*) was found on Muller E (see discussion for further analysis of such blocks). One or more blocks containing >20 genes were found on all Muller elements except Muller A (where the largest block has 15 genes). This is consistent with the results of our rearrangement rate studies (see discussion), where Muller A appears to be the most rearranged across all species. In terms of nucleotides, the largest block size (found on Muller B with 12 genes, *CG5559–CG15153*; see supplemental material) spanned 491 kb in *D. melanogaster* (∼640 kb in *D. virilis*). The smallest block size in terms of *D. melanogaster* nucleotides was just under 1 kb and has 2 genes (*CG17996–CG31812*, on Muller B). Overall, 80% of the blocks were of length <50 kb in *D. melanogaster* (median value 15 kb) and only 10 blocks exceeded 250 kb. We also tested block conservation in outgroup species in addition to the nine genus Drosophila species considered here. We find, for example, that the 2-gene block mentioned above is conserved as an adjacent pair of genes in the mosquito *A. aegypti*, the Honeybee, and the red flour beetle. Additionally, within the aforementioned 31-gene block on Muller E (*CG6413–CG13623*),

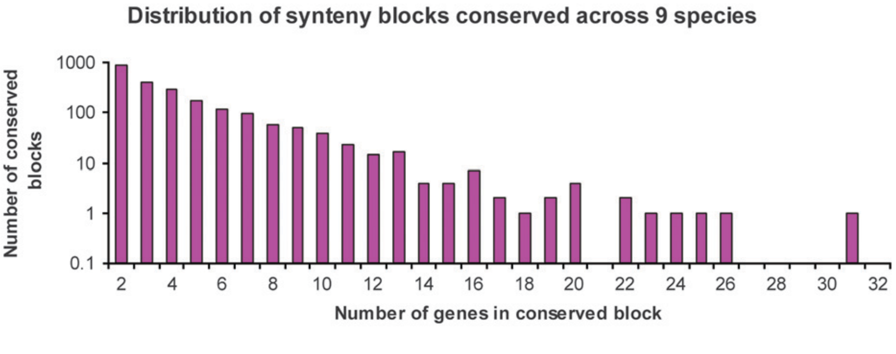

Figure 7.—Distribution of inferred multispecies conserved synteny blocks. The frequency of variously sized (number of genes) conserved blocks across nine species (excluding *D. sechellia*, *D. simulans*, and *D. persimilis*—see materials and methods) is shown (median = 3 genes, mode = 2 genes). The vertical axis (log scale) shows the number of blocks. Over 60% of the blocks have ≥3 genes. There are one or more large blocks of size ≥20 on each of the Muller elements, except Muller A (where the largest block has 15 genes). The largest block of 31 genes is found on Muller E.

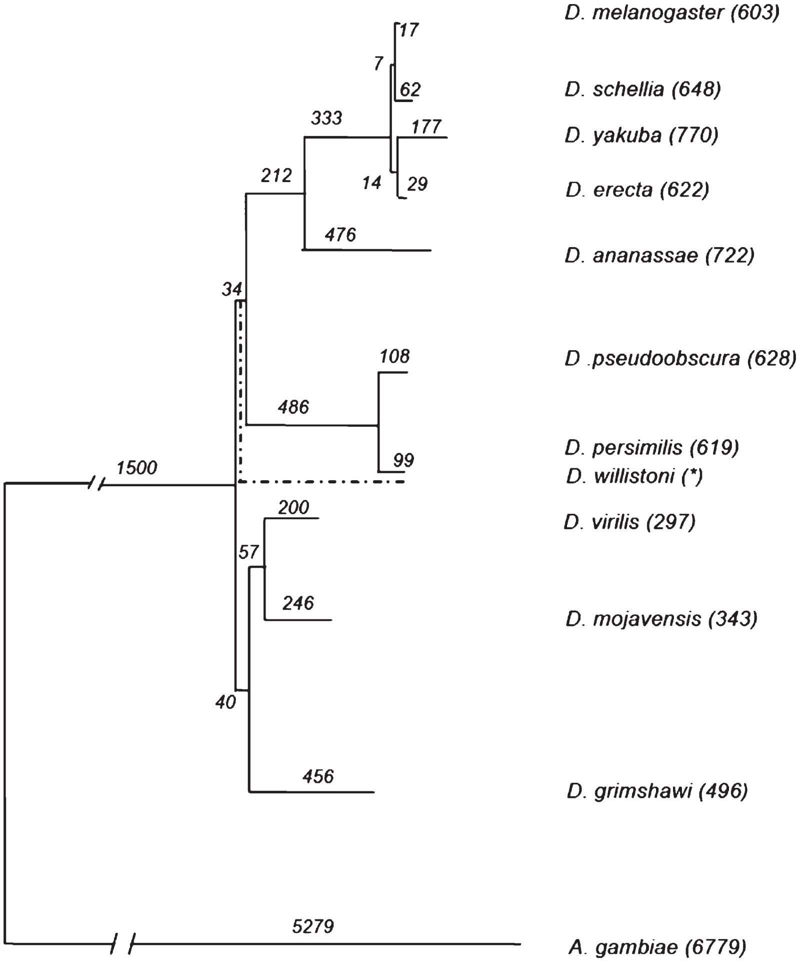

Figure 8.—Rearrangement phylogeny and estimated number of fixed chromosomal rearrangement breaks based on the NGP method (Bhutkar *et al.* 2007a), using ancestral adjacencies derived from *D. melanogaster* gene annotation. Each inferred break corresponds to a gene adjacency that existed at the immediate ancestor and was disrupted in that lineage. Paracentric inversions are assumed to be the dominant mechanism resulting in disruption of ancestral adjacencies and in shuffling the order of genes along a chromosome. This analysis includes macro-inversions as well as micro-inversions that shuffle the order and mutual transcriptional orientation between genes. Total counts of fixed breaks from the genus Drosophila root to each extant species (leaf node) are shown to the right of the figure. *An. gambiae* was used as the outgroup species to resolve ambiguities at the Drosophila genus root, wherever possible. Rearrangement counts for the subgenus Drosophila were found to be lower than those for the subgenus Sophophora. The low-coverage mosaic assembly for *D. simulans* was excluded from this analysis. See discussion for notes regarding the placement of *D. willistoni*.

17 genes are part of conserved segments within the various outgroup species (see supplemental information).

We also analyzed this data set of large conserved syntenic blocks for the presence of transposable elements within them, for functional clustering, and for correlation of embryonic gene expression data for these genes in *D. melanogaster* (Tomancak *et al.* 2002) (www.fruitfly.org) for tissue specificity and correlation of expression levels over developmental stages (see discussion). We do not find gene expression patterns for genes in a given block to be positively correlated with all others with respect to time-course expression data; however, we do see interesting patterns of high correlation values between various pairs of genes that span a block, despite being negatively correlated with others. We explore the possibility that some genes might be "trapped" within larger blocks and might be conserved in terms of position despite not being correlated with other genes in the block (see discussion).

**Inference of rearrangement phylogeny:** We employed the NGP method (Bhutkar *et al.* 2007a) to generate a phylogenetic clustering based on maximizing exclusive shared NGPs (see materials and methods). Our results expand on earlier analysis (Bhutkar *et al.* 2007a) that used a subset of the available species. The NGP approach leverages data both at the micro-inversion and at the macro-inversion levels as it analyzes the conservation and disruption of adjacent gene pairs (in a given mutual orientation). As a result, this analysis takes into account fine-scale changes to gene order that were ignored in our earlier inference of synteny blocks. The results of the phylogenetic clustering agree with the currently understood Drosophila phylogeny (Figure 8) except for the placement of *D. willistoni* (see discussion). The NGP algorithm was also used to compute rearrangement breaks across various branches of the inferred phylogeny. As opposed to the rearrangement rates calculated with respect to synteny with *D. melanogaster* gene order (see discussion), the NGP approach estimates total rearrangement breaks from the root of the genus Drosophila tree and also includes breaks from micro-inversions within syntenic blocks. Such a derivation from an inferred ancestral state leads to estimating rearrangement rates with a larger set of comparative markers rather than just comparing with *D. melanogaster.* We find these rearrangement rates to be higher in the subgenus Sophophora than those calculated using synteny data (based on macro-inversions; see discussion). This is consistent with earlier results (Bhutkar *et al.* 2007a) suggesting greater rearrangements

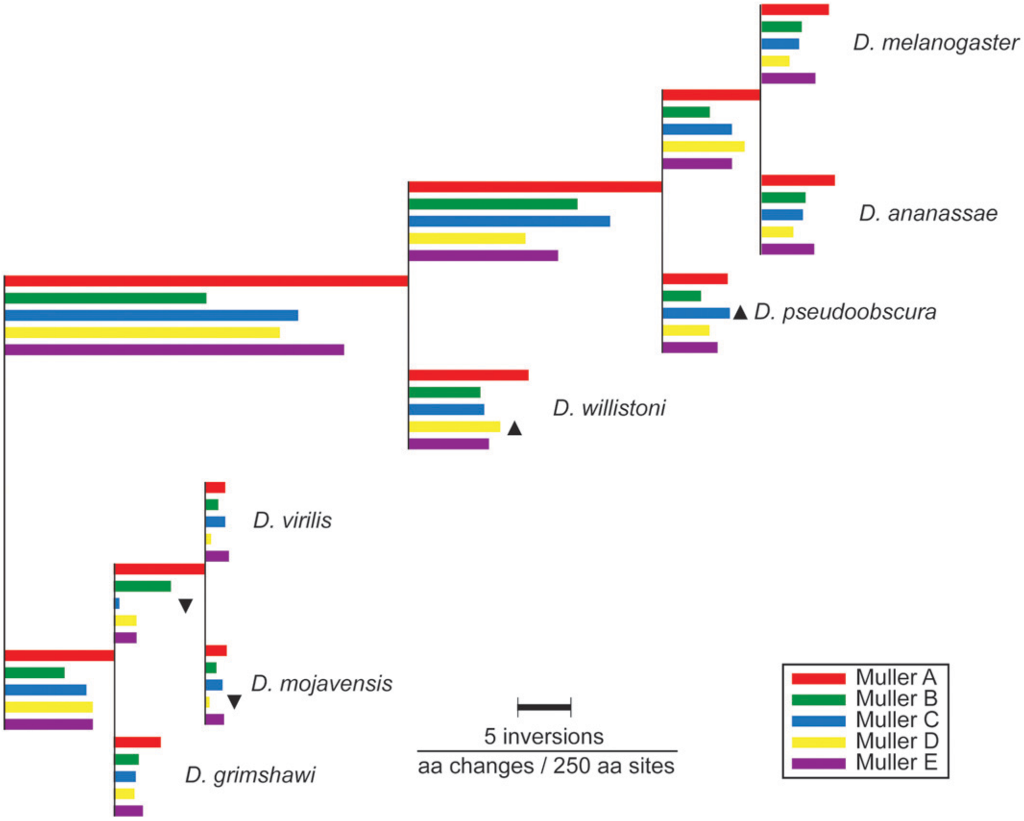

FIGURE 9.—Phylogeny of seven Drosophila species with the branch lengths based on inversion distances from pairwise linkage chain analysis. Each set of colored bars indicates the inversion rate for the five Muller elements for each branch. The triangles pointing up indicate that the number of inversions on this branch was greater than expected while the triangles pointing down indicate that the number of inversions was greater than expected. The expected values were determined on the basis of a chi-square test of homogeneity. The inversion rates were scaled to the second-codon position rate of change per 250 sites.

in the subgenus Sophophora than the subgenus Drosophila. It is also consistent with higher rates of rearrangement observed in the subgenus Sophophora (Papaceit *et al.* 2006; Gonzalez *et al.* 2007) and with the estimated distribution of intraspecific polymorphisms (Sperlich and Pfriem 1986). We obtain comparable results running the NGP approach with the *D. melanogaster* or *D. virilis* gene set as the starting point (data not shown).

We also used the inversion and breakpoint distances inferred from linkage chain analysis to estimate the branch lengths for the established Drosophila phylogeny (inversion tree shown in Figure 9). Inversion rates are presented as the number of inversions per second codon position change per 250 second codon position sites. Again, the rates of rearrangement are higher on the Sophophora lineages than on the Drosophila lineages even though the divergence times from the common ancestor for the two groups are estimated to be similar (Drosophila 12 Genomes Consortium 2007) (S. Kumar, personal communication). We tested whether the estimated branch lengths for the different Muller elements were heterogeneous. A chi-square test of homogeneity rejects the hypothesis that the numbers of inversions are the same for all chromosomes on all branches of the tree ($\chi^2 = 77.49$; d.f. $= 40$, $P < 0.0001$). Four branches on the tree contributed disproportionately to the rejection of homogeneity. Muller C on the *D. pseudoobscura* lineage and Muller D on the *D. willistoni* lineage had significant excesses of inversion events. Muller D on the *D. mojavensis* lineage and Muller C on the *D. virilis* lineage had significant deficiencies of inversions. The excess of inversions in *D. pseudoobscura* on Muller C is interesting because this chromosome has >30 gene arrangements segregating within *D. pseudoobscura* populations. This could suggest that this chromosome has experienced a recent elevation in inversion rate compared to the other chromosomes. The elevation of inversion rates on Muller D in *D. willistoni* is also interesting because Muller D fused with Muller A to become X-linked. Rearrangements on the X would be expected to be exposed to selection in hemizygous males.

We can ask what branches contribute to the reusage of breakpoints by using the number of inversions and breakpoints that map to each branch on the phylogeny to estimate *r* (Sankoff and Trinh 2005) (Table 4). Several trends emerge from the data. First, internal branches have higher breakpoint reusage values than terminal branches on all chromosomes (internal $r = 1.642$, terminal $r = 1.444$). This suggests that breakpoints uniquely used are on the more derived lineages. Second, the Sophophora lineages tend to be reused at a higher rate than the Drosophila lineages on all chromosomes (Sophophora $r = 1.502$, Drosophila $r = 1.278$). This is consistent with the overall higher rearrangement rates on the Sophophoran *vs.* Drosophila branches, which is inferred to lead to higher reusage

TABLE 4

**Breakpoint reusage for the five major Muller elements on different branches of the Drosophila phylogeny**

| Branch | A | B | C | D | E | Total |
|---|---|---|---|---|---|---|
| Internal | 1.955 | 1.589 | 1.508 | 1.455 | 1.658 | 1.642 |
| External | 1.566 | 1.394 | 1.387 | 1.429 | 1.403 | 1.444 |
| Sophophora | 1.678 | 1.428 | 1.430 | 1.459 | 1.464 | 1.502 |
| Drosophila | 1.362 | 1.273 | 1.207 | 1.114 | 1.323 | 1.278 |
| Total | 1.652 | 1.438 | 1.417 | 1.437 | 1.460 | 1.492 |

estimates. Finally, Muller A has a higher rate of reusage than the autosomal elements (Muller A $r = 1.652$, autosomes $r = 1.438$).

**Heterochromatin sequences and rearrangements:** Similar to anchoring scaffolds to the euchromatic portion of the genome (on the basis of majority hits from *D. melanogaster* gene markers in a candidate assembly—see MATERIALS AND METHODS), we were able to assign a number of scaffolds, in individual species, to heterochromatic portions of the genome (Schaeffer *et al.* 2008). We also find heterochromatic genes localized toward the edges of large terminal euchromatic scaffolds, as expected. For example, in the *D. ananassae* synteny report, 15 *D. melanogaster* heterochromatic genes have strong homologous placements at the edge of scaffold 12911 (GenBank accession no. *CH902623.1*), which is a terminal scaffold anchored on Muller E toward the centromere (Schaeffer *et al.* 2008). Additionally, this data set allowed us to identify genes that are heterochromatic in *D. melanogaster* but are found in largely euchromatic portions of other genomes, suggesting rearrangements involving heterochromatic sequences. These genes are included in the synteny table for each species (see supplemental material).

**Inference of repetitive sequences in breakpoint sequences:** Analysis of the sequences within the breakpoints of *D. pseudoobscura* revealed an enrichment for repetitive sequences (Richards *et al.* 2005). We asked whether the breakpoints in linkage chains of size greater than two had repetitive sequences as detected from a high degree of interbreakpoint matching. Breakpoints in linkage chains greater than two did not show a significant difference in the level of interbreakpoint matching than breakpoints in linkage chains of two.

## DISCUSSION

**Syntenic blocks and rearrangement rates:** An analysis of the distribution of syntenic blocks between various species (Figure 4) shows correspondence with the Drosophila phylogeny. Evolutionarily close species to *D. melanogaster* (*D. sechellia* and *D. erecta*, for example) show larger blocks of synteny where the distribution includes blocks in excess of 900 genes in each case. The distant species of the subgenus Drosophila, by contrast, do not include blocks in excess of 90 genes and the distribution is biased toward lower block sizes. The rearrangement process can be analyzed in terms of the number of conserved blocks of synteny in species at different evolutionary distances (Figure 10). Depending on the choice of estimated divergence times we find the breakup of syntenic blocks to approximate an exponential process or a linear process over time. With divergence estimates based on mutational clock analysis (Tamura *et al.* 2004) the number of synteny blocks approximates an exponential process over time ($R^2 = 0.9432$ with the best fit through the data values although a linear model has $R^2 = 0.9549$; however, it does not fit the data values as well as the exponential model). With alternate divergence estimates (Russo *et al.* 1995) the data best fit a linear model.

We also looked at the distribution of the size of chromosomal segments (in kilobases) that contain various syntenic blocks, in the context of chromosomal breakage models. The random breakage model was first proposed by Ohno (1973) and later formalized by Nadeau and Taylor (1984). It has since been explored further by a number of other studies (Nadeau and Sankoff 1998a,b; Schoen 2000; Waddington *et al.* 2000). We attempted to test the suitability of the Nadeau–Taylor model to our synteny data. The distribution of the lengths of syntenic segments is expected to be exponential under this model. We analyzed the distribution of the lengths of syntenic segments (in kilobases) between *D. melanogaster* and *D. virilis* (across all Muller elements after synteny blocks were merged across assembly scaffold gaps wherever possible). Only scaffolds anchored to chromosomes were considered (Schaeffer *et al.* 2008). We also checked to make sure that the large number of small blocks (<10, 20, or 30 kb) was not an artifact of assembly fragmentation. Only blocks on corresponding Muller elements were considered (transpositions were filtered out) and singleton gene blocks were similarly ignored. The results do not fit an exponential curve, but they fit a power law, which is a straight line on a log–log graph (Figure 10c). This suggests that the distribution of breakpoints on a chromosome is not random for these species. A large number of small syntenic segments bias the distribution and there seem to be a number of large conserved segments across species (as we discuss later). These observations are similar to other studies of insects and mammals (Pevzner and Tesler 2003; Zdobnov and Bork 2007).

We also investigated the distribution of synteny blocks (in number of genes) in individual chromosomes of each species (Figure 11). In general, we observe Muller element A to be the most fragmented with respect to the *D. melanogaster* gene order across all species (except *D. sechellia*). In the subgenus Drosophila, for example,

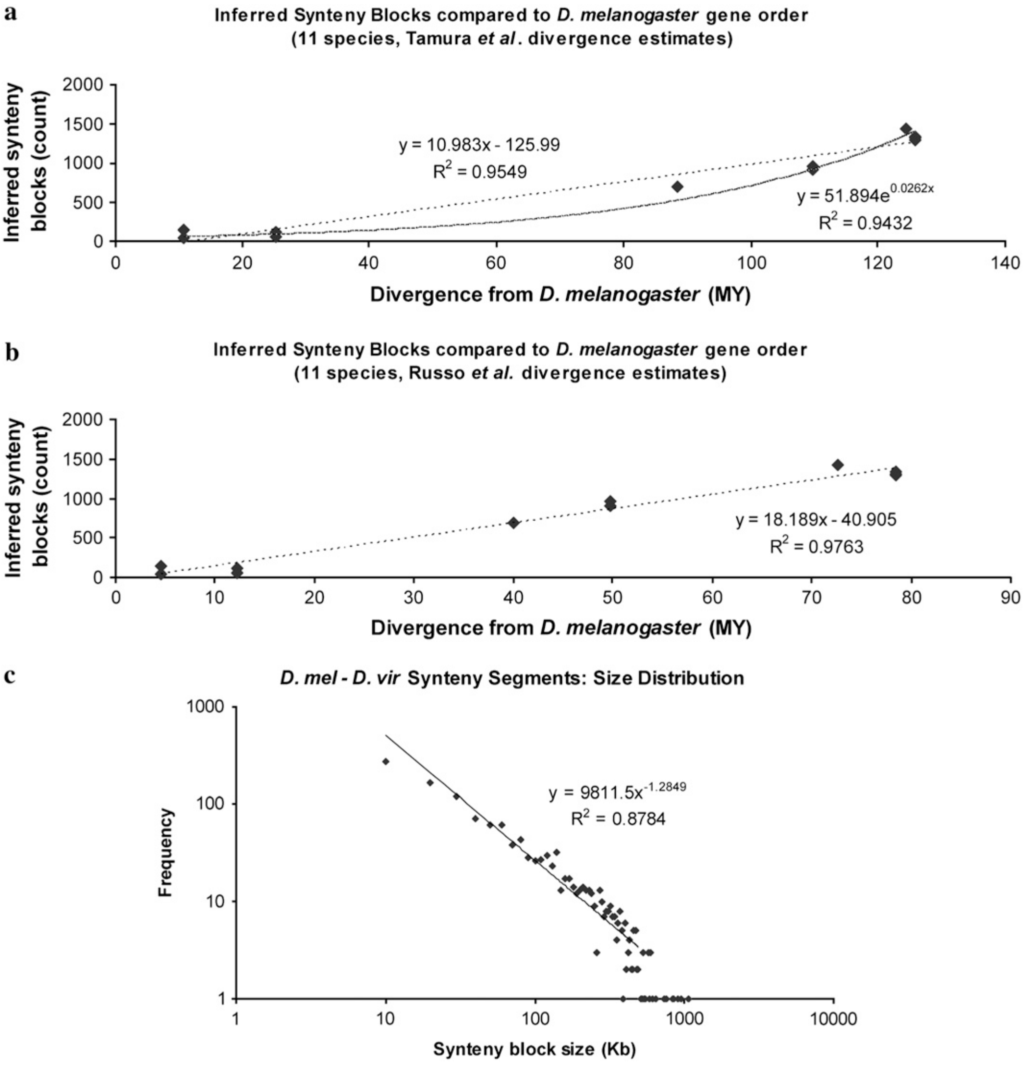

FIGURE 10.—Synteny disruption over evolutionary time. (a) Synteny disruption based on divergence estimates from TAMURA *et al.* (2004). The horizontal axis shows total independent evolutionary time between *D. melanogaster* and a given species (which is twice the time since their last common ancestor). The vertical axis shows the number of inferred synteny blocks with respect to the *D. melanogaster* gene order. An exponential model fits this data set well. A linear model (with comparable $R^2$ value) is also shown but it does not contain the data points as well as the exponential model. (b) Synteny disruption based on divergence estimates from RUSSO *et al.* (1995). A linear model fits this data set best. (c) An example of the distribution of synteny block lengths (in kilobases) between *D. melanogaster* and *D. virilis*. The vertical axis (log scale) shows the frequency of variously sized syntenic segments (in kilobases on the horizontal axis; log scale). A power law fits this distribution well, suggesting a bias toward a large number of small segments with fewer well-conserved larger segments. See text for a discussion of possible chromosomal breakage models.

Muller elements B–E show similar sizes of average block sizes with Muller A being the outlier. The total number of blocks on each Muller element, the maximum block size, and total number of genes in these blocks are generally in line with the phylogenetic distribution of these species and the increased fragmentation of Muller A (see supplemental material).

There are two explanations for why Muller A has a high inversion rate. The first possibility is that scaffold order for Muller A is artificially shuffled due to assembly errors. DNA for these genomic sequencing projects was derived from males and females so that the X would be expected to have three-quarters the sequence coverage of the autosomes, leading a greater number of scaffolds for the X. The other one-quarter sequence coverage would be sequences for the Y chromosome. More scaffolds may lead to greater potential for misordering of the contigs and artificially elevated inversion rates and breakpoint reusage. Lower sequence coverage of the X can be ruled out because the metacentric X of *D. pseudoobscura* includes genes from Muller A and D. Thus, one would expect both arms of the X in these species to suffer from the misassembly and scaffold-ordering problem. Linkage chain analysis of Muller D involving comparisons of *D. pseudoobscura* does not show evidence for significant excess of breakpoint reusage.

The second explanation for the elevated rate of breakpoint reusage could be that selection for X-linked variation is enhanced in males. Muller A has the greatest number of inversions *vs.* any of the autosomes even when chromosome length is used to standardize the number of rearrangements. This elevation of inversion rate could inflate the inversion rate and the levels of reusage. This result raises the intriguing possibility that the evolution of Muller A is being enhanced by adaptive fixations of rearrangements. If selection in males is driving up the rate of rearrangement on the X, then this suggests that rearranged chromosomes harbor recessive alleles within inversions that contribute to local adaptation (KIRKPATRICK and BARTON 2006).

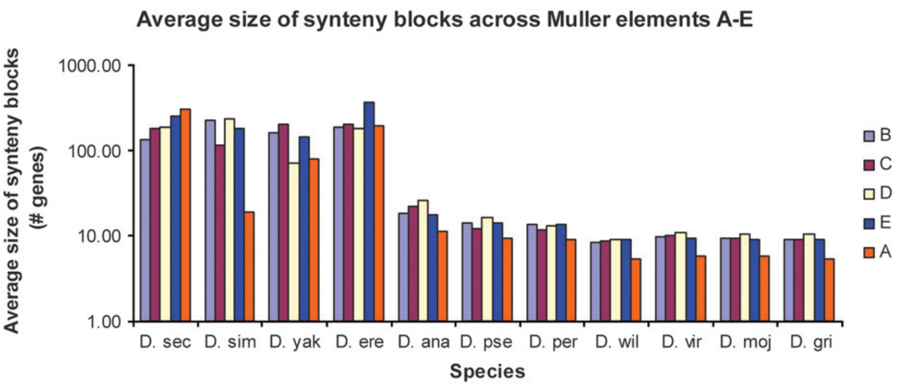

Figure 11.—Distribution of the average size (in number of genes) of synteny blocks across the Muller elements (A–E) of various species based on the *D. melanogaster* euchromatic gene order. The vertical axis shows the average size of the blocks (log scale). Outside the *melanogaster* subgroup Muller A shows a trend of being more fragmented than the other arms. Species: *D. sec*, *D. sechellia*; *D. sim*, *D. simulans*; *D. yak*, *D. yakuba*; *D. ere*, *D. erecta*; *D. ana*, *D. ananassae*; *D. pse*, *D. pseudoobscura*; *D. per*, *D. persimilis*; *D. wil*, *D. willistoni*; *D. vir*, *D. virilis*; *D. moj*, *D. mojavensis*; *D. gri*, *D. grimshawi*.

This data set also lends itself to an analysis of rates of chromosomal evolution within the genus Drosophila. Using the number of syntenic blocks between species as a lower bound of the chromosomal disruptions during their total time of divergence, we calculated the rates in terms of disruptions per megabase per MY (Ranz *et al.* 2001; Gonzalez *et al.* 2002; Bartolome and Charlesworth 2006). Rates were calculated for two estimates of the divergence time between Drosophilids (Russo *et al.* 1995; Tamura *et al.* 2004) for each chromosome in each species (Table 5). These represent the first such attempt to calculate these rates across these species on the basis of direct counts of rearrangement events on the whole-genome scale. Taking Muller element E as an example (Table 5), we find rearrangement rates higher than those estimated before (Ranz *et al.* 2001; Gonzalez *et al.* 2002; Bartolome and Charlesworth 2006) but largely within the same order of magnitude depending on the estimates of divergence time, chromosome size, and sets of species chosen for comparison. For example, our whole-genome direct-count results for rates (Muller E) calculated using the *D. melanogaster*–*D. pseudoobscura* species pair show 0.071–0.157 disruptions/Mb/MY (Table 5) for divergence estimates from Tamura *et al.* (2004) and Russo *et al.* (1995), respectively. Earlier studies estimated lower rates for Muller E in using this pair of species: 0.013 (support limits 0.007–0.029) (Bartolome and Charlesworth 2006), 0.014 (±0.060) (Ranz *et al.* 1997), and 0.0848 (Ranz *et al.* 2001) disruptions/Mb/MY using various divergence estimates. The last estimate is based on a chromosome size of 22 Mb and divergence estimate of 30 MY between these species, using which our analysis gives a rate of 0.165 disruptions/Mb/MY. The higher end of our estimates is found to be of the same order of magnitude as the lower end of estimates (0.4–1.0 disruptions/Mb/MY) reported previously for yeast (Coghlan and Wolfe 2002), similar to (but higher than) earlier analysis

**TABLE 5**

**Chromosomal rearrangement rates across various Drosophilids for Muller element E**

| Species | No. synteny blocks | Chr. (kb) *D. mel*, species | Divergence A, B (MY) | A: estimated rearrangement rate (disruptions/Mb/MY) | | B: estimated rearrangement rate (disruptions/Mb/MY) | |
|---|---|---|---|---|---|---|---|
| *D. sechellia* | 12 | 27,894, 27,691 | 5.4, 2.3 | 0.03983 | 0.04013 | 0.09352 | 0.09421 |
| *D. simulans* | 18 | 27,894, 27,519 | 5.4, 2.3 | 0.05975 | 0.06056 | 0.14028 | 0.14219 |
| *D. yakuba* | 23 | 27,894, 28,834 | 12.6, 6.1 | 0.03272 | 0.03165 | 0.06759 | 0.06538 |
| *D. erecta* | 11 | 27,894, 28,221 | 12.6, 6.1 | 0.01565 | 0.01547 | 0.03232 | 0.03195 |
| *D. ananassae* | 183 | 27,894, 33,167 | 44.2, 20 | 0.07421 | 0.06242 | 0.16401 | 0.13794 |
| *D. pseudoobscura* | 218 | 27,894, 30,794 | 54.9, 24.3 | 0.07118 | 0.06447 | 0.15693 | 0.14215 |
| *D. persimilis* | 224 | 27,894, 31,655 | 54.9, 24.3 | 0.07314 | 0.06445 | 0.16125 | 0.14209 |
| *D. willistoni* | 341 | 27,894, 31,708 | 62.2, 36.3 | 0.09827 | 0.08645 | 0.16839 | 0.14813 |
| *D. virilis* | 313 | 27,894, 35,495 | 62.9, 39.2 | 0.08920 | 0.07010 | 0.14313 | 0.11248 |
| *D. mojavensis* | 325 | 27,894, 34,148 | 62.9, 39.2 | 0.09262 | 0.07565 | 0.14861 | 0.12140 |
| *D. grimshawi* | 325 | 27,894, 34,503 | 62.9, 39.2 | 0.09262 | 0.07488 | 0.14861 | 0.12015 |

All rates were calculated using *D. melanogaster* as the reference species. Synteny blocks with respect to *D. melanogaster* (*D. mel*) are used to estimate a lower bound for the number of chromosomal disruptions on Muller element E. Rates are calculated using two estimates for chromosome sizes (one for *D. melanogaster* and the other for each individual species based on scaffolds mapped to that chromosome) (Schaeffer *et al.* 2008) and two estimates for species divergence times: A (Tamura *et al.* 2004) and B (Russo *et al.* 1995). For each divergence time estimate (A, B), two sets of rates are shown, first using the *D. melanogaster* chromosome size and then second using the species' estimated chromosome size.

**TABLE 6**

**Chromosomal rearrangement rates for four representative species across all Muller elements**

| Species | Muller A | | Muller B | | Muller C | | Muller D | | Muller E | |
|---|---|---|---|---|---|---|---|---|---|---|
| *D. willistoni* | 0.12777 | 0.10045 | 0.08604 | 0.05916 | 0.10248 | 0.07007 | 0.08468 | 0.06637 | 0.09827 | 0.08645 |
| *D. virilis* | 0.11259 | 0.07619 | 0.08008 | 0.06208 | 0.09388 | 0.06950 | 0.07207 | 0.06244 | 0.08920 | 0.07010 |
| *D. mojavensis* | 0.11332 | 0.07772 | 0.07722 | 0.05239 | 0.09702 | 0.07308 | 0.07241 | 0.06138 | 0.09262 | 0.07565 |
| *D. grimshawi* | 0.11911 | 0.09897 | 0.07936 | 0.06601 | 0.09781 | 0.08420 | 0.07276 | 0.06860 | 0.09262 | 0.07488 |

All rates are calculated with respect to *D. melanogaster* as the reference species. Rates shown are calculated using divergence estimates (Tamura *et al.* 2004) of 62.2 MY and 62.9 MY, respectively, for *D. willistoni* and the subgenus Sophophora species with respect to *D. melanogaster*. The first number for each element uses the *D. melanogaster* chromosome size to estimate the rate and the second number uses the size for each species based on the genome assembly (similar to Table 5). We observe chromosomal evolution rates for Muller A (*D. melanogaster* chromosome X) to be consistently higher compared to other Muller elements.

(Bartolome and Charlesworth 2006). Within subgroups, we observe discordant rearrangement rates in species equidistant from *D. melanogaster* (*D. yakuba* and *D. erecta*, for example), similar to those observed in a smaller study (Vieira *et al.* 1997). We observe similar rearrangement rate patterns across all chromosomes (data not shown). Our observations for Muller element A confirm earlier studies regarding a higher rate of evolution for the *D. melanogaster* X chromosome (Charlesworth *et al.* 1987). We find the average size of synteny blocks to be lower, as mentioned before. Additionally, we find rearrangement rates calculated for this element to be higher than those for other elements (Table 6).

We observed a major difference in the rate of rearrangement between the Sophophora and the Drosophila subgenera. This may result from differences in polymorphism levels within species of these subgenera (Sperlich and Pfriem 1986). Members of the *obscura* and *willistoni* groups are quite polymorphic for paracentric inversions (Da Cunha *et al.* 1950; Da Cunha and Dobzhansky 1954; Dobzhansky and Sturtevant 1938; Valente and Morales 1985; Krimbas 1992; Valente *et al.* 1993). Thus, the fixed rearrangement differences may result from higher levels of rearrangement polymorphism within these lineages. Another factor that could explain the elevated fixation rates within the Sophophora is that Sophophora have a shorter generation time (Markow and O'Grady 2007) compared to Drosophila species.

Alternatively, inferred ancestral gene adjacency disruption information from the NGP method was used to infer rearrangement rates (see materials and methods). These rates are within the same order of magnitude as our results using syntenic block information (supplemental Table S4).

**Large conserved synteny blocks:** Our list of multispecies conserved blocks includes some previously studied blocks. For example, members of the *Osiris* gene family (Dorer *et al.* 2003) are inferred to be part of large conserved blocks on Muller element E (block nos. 47 and 48, *CG1154–CG31559* and *CG15595–CG18048*) along with strong gene order conservation evidence from various outgroup insect species (see supplemental material). Additionally, these genes show a high level of embryonic expression correlation, *in D. melanogaster* (Tomancak *et al.* 2002), with other genes in their blocks. These two blocks are conserved as a single large block in all species (and hence the complete *Osiris* gene family of 20 genes is conserved as a contiguous block). However, in *D. willistoni*, the intermediate genes (*CG15597*, *CG15594*) between *Osi12* and *Osi13* could not be placed in syntenic locations, resulting in these genes being reported as part of two blocks (see materials and methods).

In an earlier study, Zdobnov and Bork (2007) assumed random gene order between genomes and estimated the probability of a minimal conserved syntenic block (two orthologs next to each other with at most a single gene separating them) for a data set of 4632 orthologs across various insect species. They reported the probability to be in the range of $P < 4 \times 10^{-3}$ for two-way synteny and $P < 4 \times 10^{-6}$ for three-way synteny. In a data set of 8967 genes (Bhutkar *et al.* 2007a) that we found strong orthology for across all Drosophila species, if we assume random gene order between genomes we estimate the probability of a gene having the same neighbor in the other genomes as $P < 1 \times 10^{-4}$ for two-way synteny and $P < 1 \times 10^{-8}$ for three-way synteny.

Analysis of the cross-genus conserved blocks suggests various hypotheses for their conservation. We see few cases of functional clustering on the basis of available Gene Ontology annotation (Gene Ontology Consortium 2001), where all or some of the genes in a block have similar annotation for molecular or biological processes: for example (see supplemental material for a list of genes in numbered blocks), Muller A, block 294 (6 genes, *CG9676–CG4678*) and block 4 (7 genes, *CG3796–CG3923*); Muller B, block 169 (9 genes, *CG8419–CG8282*, conserved through outgroup species); Muller D, block 81 (3 genes, *Cpr64Aa*, *Cpr64Ab*, *Cpr64Ac*; all involved in larval cuticle structure, conserved through outgroup species, similar to block 112); and Muller E, block 55 (13 genes, *Taf1-Ccp84Aa*, 8 involved in larval cuticle structure). On the other hand, there are conserved blocks composed of genes with seemingly unrelated functional (gene ontology, GO) annotation: for example, Muller B, block 122 (7 genes, *CG9553–CG9098*). We also explored expression correlation of genes within a syntenic block, utilizing previously published embryonic gene expression studies in *D. melanogaster.* We find ∼60% of these genes to have a correlation coefficient >0 when compared to their preceding gene in a block. The percentage of genes that have a positive average correlation coefficient with all other genes in their block is also ∼60%. This is lower than the ∼80% of genes that exhibit positive correlation, reported in a previous study (Stolc *et al.* 2004) utilizing full life-cycle expression data. Our analysis might be limited by embryonic expression data. However, when we look at individual blocks we see various patterns of expression correlation (Figure 12) within the large conserved blocks on each Muller element. If expression correlation is indeed a driving factor in block conservation, these patterns might suggest islands of strong expression correlation (or networks of strong correlation) within conserved blocks. It might be possible that a number of genes that do not share any functional or expressional commonality with other genes might be trapped within these islands (or networks) and hence might be part of these conserved blocks. The boundaries between syntenic blocks may include gene expression insulator sequences (Dorman *et al.* 2007). Another attribute we studied, tissue specificity during different stages of embryonic development, does not appear to be a dominating common factor between genes in the same block (see supplemental information).

We also looked at the prevalence of natural transposons (Kaminker *et al.* 2002) and *P*-insertion elements (Spradling *et al.* 1995; Bellen *et al.* 2004) in *D. melanogaster* within such conserved blocks. We find both classes of elements (insertion sites) within conserved blocks across the size spectrum. We focused on the prevalence of *P*-element insertion sites in some of the large blocks. Selecting the top two blocks (in length) from each Muller we get 14 large blocks (there are multiple blocks of the same size on the same Muller in some cases) of which 8 either have no *P*-element insertion sites that disrupt the span of a gene or have insertion sites that disrupt genes only at the edges of these blocks (see supplemental information). The remaining blocks have a small number of genes (up to three) that are not on the edges of these blocks that are disrupted by *P*-element insertion sites. These observations raise the question of whether genes in conserved blocks resist disruption by insertion elements. Genes on the edges that seem to be interrupted by insertion elements, as well as the few genes within blocks that are interrupted, might be trapped within the islands (or networks) of strong expression correlation within these blocks. Further study with gene expression data from multiple species would be needed to explore this possibility.

It is possible that one or more of these hypotheses might contribute to the strong conservation of gene order across these species: Functional clustering and expression correlation (islands or networks within blocks) might in part be responsible for the multispecies conservation. This analysis is limited by the availability of expression data for various species. Further gene expression studies in the non-*D. melanogaster* species are necessary to demonstrate that genes within these clusters have maintained correlated regulation patterns.

In a previous study, Spellman and Rubin (2002) identified 3199 genes in *D. melanogaster* that are part of 210 separate blocks of neighboring genes with similar expression patterns (embryonic and adult). When we compare these *D. melanogaster* results with our multispecies data set, we find that >50% of the genes in 155 (of 210) of these *D. melanogaster* (Spellman and Rubin 2002) blocks are part of our multispecies conserved blocks (>80% of the genes in 70 blocks). Overall, 1823 of the 3199 genes were found to exist in some multispecies conserved blocks. One of the hypotheses put forth by Spellman and Rubin is that there might be a core set of genes that are adjacent as they might need to be transcribed together, and the rest of the genes in a block might just happen to get transcribed as chromatin remodeling in that vicinity might enable this to happen. A hypothesis resulting from their analysis is that if these blocks containing genes with correlated expression patterns are conserved across species, then there might be an advantage to maintaining such gene adjacencies. We observe many of these blocks to be conserved across species, supporting the hypothesis that there might indeed be an evolutionary advantage. Furthermore, we find the boundaries of these blocks to include genes beyond the Spellman–Rubin blocks in *D. melanogaster,* implying that the edges of the blocks might contain genes that are "along for the ride" (complementary to the notion put forth by Spellman and Rubin for genes beyond the core set of expression-correlated genes) from a conservation perspective (and might not necessarily be correlated from an expression perspective). For example, block 413 (Figure 12) has a core set of 9 genes from *CG13616* to *CG13613* that are part of a Spellman–Rubin block of correlated expression in *D. melanogaster.* We find an additional 9 genes to be conserved across species on either side of the block. Similarly, for multispecies conserved block 280 (Figure 12) we find 1 gene (*CG31839*) on one side and 9 genes (*CG15288–CG4501*) on the other side of a core block of 4 genes (*CG8930–CG16873*) in the Spellman–Rubin *D. melanogaster* set. All of block 457 is found in both data sets. However, interestingly, none of the multispecies conserved block 47 (*Osiris* cluster in Figure 12) was identified as a block of expression-correlated adjacent genes in the Spellman–Rubin *D. melanogaster* set.

***D. willistoni* and the rearrangement phylogeny:** The currently understood placement of *D. willistoni,* a

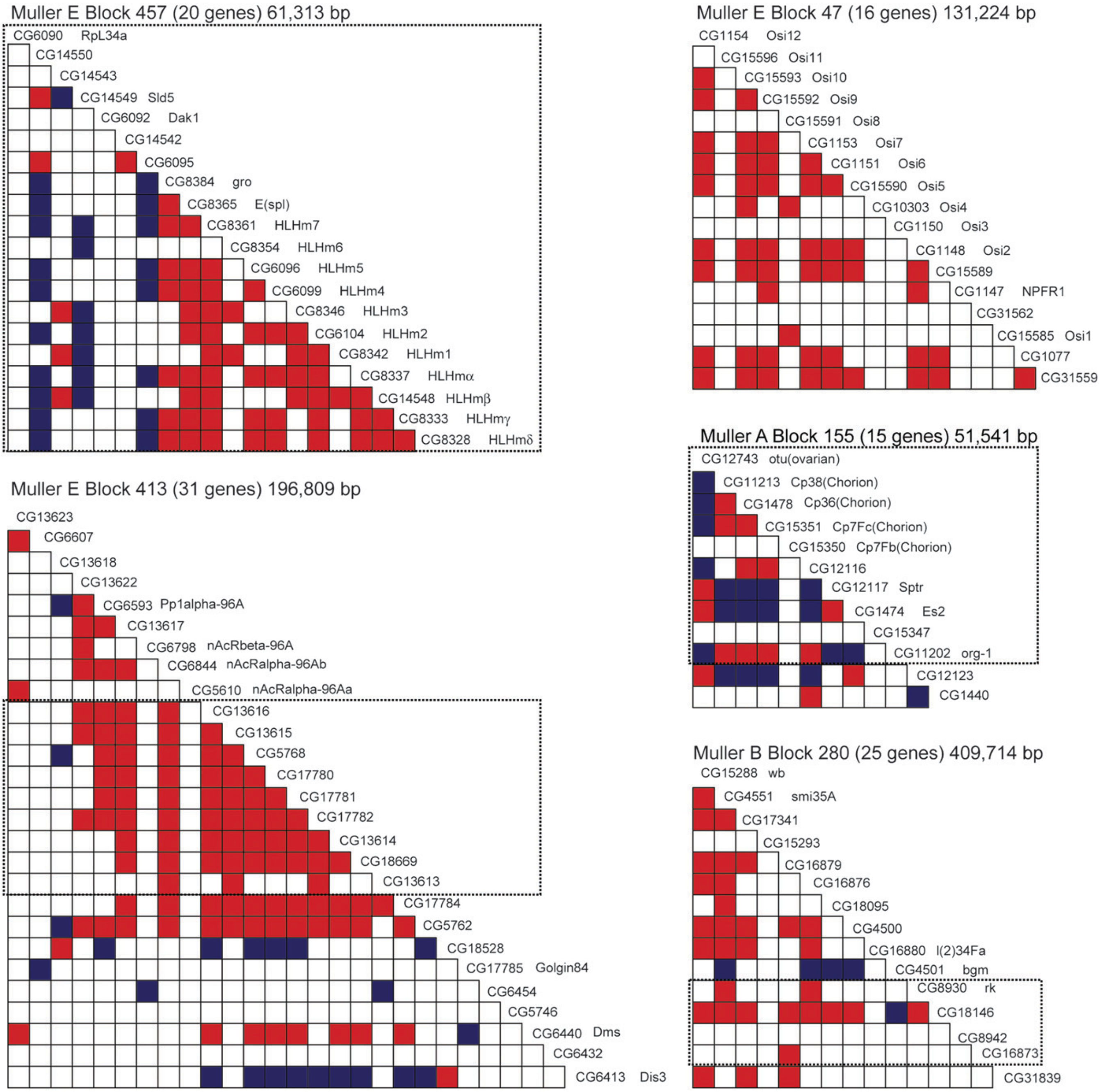

Figure 12.—Embryonic gene expression correlation within some of the larger multispecies conserved blocks for each Muller element. Red squares indicate significant positive correlation ($P < 0.05$) and blue squares indicate significant negative correlation. Block identifiers refer to their tags in the supplemental material. Although some blocks exhibit high levels of positive correlation (like Muller E block 47, which contains the *Osiris* gene cluster; see discussion), others show different patterns. These plots raise the possibility of islands and networks of positive gene correlation that might select for these genes to be conserved in their order or proximity to each other (see discussion for details). For example, blocks 457 and 413 on Muller E have a major chunk each of positively correlated genes devoid of any negative correlation in these chunks. Further, they show positive correlation with some genes outside these chunks, which might hold the whole block together. Blocks 155 (Muller A) and 280 (Muller B) show patterns where some negatively correlated genes might be trapped within blocks (for example *bgm* in block 280). We find areas of overlap between our set of cross-species conserved blocks and blocks determined by Spellman and Rubin (2002) in a study based on expression analysis in *D. melanogaster.* Areas of overlap are shown boxed (in dashed lines). See text for an analysis of the overlap and discussion related to cross-species conservation and expression profiles.

species of the subgenus Sophophora, is very close to the split (62.9 MYA) between the two subgenera (Sophophora and Drosophila) of the genus Drosophila (Tamura *et al.* 2004). *D. willistoni* is known to have extensive gene arrangement polymorphism on all chromosomes (Da Cunha *et al.* 1950, 1959; Da Cunha and Dobzhansky 1954; Valente and Araujo 1985, 1986; Valente *et al.* 1993, 2001, 2003; Rohde 2000; Rohde *et al.* 2005) as observed from chromosomal variability in natural populations.

As a result of its higher rate of intraspecific polymorphism and significantly large independent evolutionary time compared to other species, computational methods seem to demonstrate some ambiguity in its phylogenetic placement. The ambiguity arises because of the elevated level of sequence and gene-order evolution that leads to a significantly long lineage leading to *D. willistoni.* Most phylogenetic reconstruction software tends to force *D. willistoni* to be the outgroup on the basis of elevated evolutionary rates. For example, gene trees with PHYLIP (Felsenstein 1989) constructed using concatenated SRP54 and SRP19 amino acid sequences (genes thought to be under minimal species-specific selection) for various species (Drosophila 12 Genomes Consortium 2007) (rana.lbl.gov/drosophila) do not match the currently accepted phylogeny (Powell 1997) (supplemental Figure S2). Rearrangement-based trees are similarly affected as a result of its early divergence from all other species in our set. This also highlights one of the limitations of such computational approaches in dealing with species that have large divergence times and where we lack the genome sequence of an evolutionarily close species for comparison. The NGP clustering algorithm based on maximizing exclusively shared NGPs suggests two solutions for the placement of *D. willistoni.* The first solution places *D. willistoni* as an outgroup to all other Drosophila species (which are clustered with 687 NGPs). The second solution clusters it with the subgenus Sophophora species on the basis of 396 NGPs. The weaker solution matches the currently accepted phylogenetic partitioning of the genus Drosophila. Additionally, a genome assembly error was uncovered in *D. willistoni* on the basis of syntenic analysis (involving Muller elements C and D) and the Muller elements E and F have undergone fusion in these species. As a result of these issues, we eliminated *D. willistoni* from the set of species used to estimate rearrangement rates on the basis of the micro-inversion-based method employed by the NGP approach. Similarly, the low-coverage mosaic assembly of *D. simulans* was also set aside for this NGP comparison. In line with these observations, *D. willistoni* also exhibits a higher rate of rearrangement compared to other species (Tables 5 and 6).

**Rearrangement breakpoints are reused:** Pairwise comparisons of gene order among the seven species in the genus Drosophila using linkage chain analysis demonstrate that there is a high degree of breakpoint reusage. All pairwise comparisons clearly reject a breakage model that assumes that breakpoints are introduced on the basis of a uniform distribution. This does not, however, rule out the possibility that some other distribution other than the uniform may underlie the introduction of breakpoints. On the other hand, most pairwise comparisons fail to reject a breakpoint hot/cold-spot model using a uniform sampling of a subset of sites along the chromosome. The hot- and cold-spot model can be explained either by how inversion mutations are introduced or by the process that new inversions are fixed in populations. The mutation hypothesis suggests that particular sites on the chromosome are more susceptible to double-strand breaks either because of repetitive sequences (Richards *et al.* 2005) or because of transcriptional initiation of double-strand breaks.

Several mechanisms have been suggested for the generation of genome rearrangements in Drosophila. The first model of rearrangement was popularized with the discovery of transposable elements in Drosophila genomes (Engels and Preston 1984; Lim 1988). The transposable-element model proposed that these repetitive sequences act as sites for ectopic exchange within chromosomal arms, leading to rearrangement. Sequences at inversion breakpoints have (Mathiopoulos and Lanzaro 1995; Mathiopoulos *et al.* 1998, 1999; Cáceres *et al.* 1999, 2001) and have not found repetitive sequences at the breakpoints of gene arrangements (Wesley and Eanes 1994; Andolfatto and Kreitman 2000; Anderson *et al.* 2005; Matzkin *et al.* 2005). Several studies of rearrangement breakpoints have shown that the gene duplications often accompany the rearrangement event (Matzkin *et al.* 2005; Ranz *et al.* 2007). Repeat sequences are observed at rearrangement breakpoints; however, not all species have repeats and we did not find evidence for an elevation of repeat sequences in reutilized breakpoints. Extant breakpoint sequences may not, however, reflect what repeats existed at the time of the rearrangement events in an ancestral lineage.

The second mutational mechanism is based on the observation that derived inversions tend to be more distal to the ancestral rearrangements (Novitski 1946). Novitski (1946) thought that the sites of breakpoint pairing in gene-arrangement heterozygotes would be susceptible to new double-strand breaks, creating a new arrangement with breakpoints more distal to the ancestral ones. This model is consistent with observations of the coincidence or reusage of rearrangement breakpoints on cytological maps within and among species (Lemeunier and Ashburner 1976; Olvera *et al.* 1979). Modest support for this model was found in species with extensive inversion polymorphism as well as from knowledge of ancestry of the different karyotypes (Dobzhansky 1944; Novitski 1946). To adequately test this model, we need to know if the sequence of rearrangement events that converted the common ancestral gene order into the current gene order used increasingly more distal breakpoints. At this point, the history of inversion events on the different Drosophila lineages is not clear. Thus, approaches that infer the sequence of inversion down each lineage may provide valuable insights about the mutational mechanisms.

The fixation hypothesis leaves open the possibility that all sites are free to be an inversion breakpoint, but differential fixation of inversions leads to the appearance of nonrandom distribution of breakpoints. Purifying selection could remove new inversion mutations if one or both breakpoints occur within the boundary of a gene or disrupt regulatory sequences. Careful investigation of the sequences at syntenic blocks is needed to address this possibility. Another possibility is that selection could remove inversions that break up coordinately expressed genes. Positive Darwinian selection could promote the fixation of new chromosomal inversions (Charlesworth and Charlesworth 1973; Charlesworth 1974) by capturing sets of genes that allow local adaptation (Kirkpatrick and Barton 2006), by capturing suites of epistatically interacting genes (Dobzhansky 1950), or by a combination of both mechanisms.

**Concluding remarks:** The Drosophila 12 genomes project has provided a glimpse into the potential forces that shape the organization of genes on chromosomes and into the rates of chromosomal evolution. Micro- and macro-inversions have acted to reorganize genes within the genome. Purifying selection appears to be the predominant force that acts on gene-order changes. An important finding of this research is that information at syntenic block boundaries carries information about the past history of rearrangements and offers the possibility to easily reconstruct that history. It also highlights conservation of gene-ordered blocks across species, their expression correlation, and the evolutionary process of gene-order scrambling.

S.W.S. thanks Richard P. Meisel and Hiroshi Akashi for helpful discussions about the methods developed in this article and Michael Mollenhauer and Gillian Halter for help with preliminary analyses. We also thank FlyBase for *D. melanogaster* data sets and computing support, the research computing group in the Division of Life Sciences at Harvard University for systems support, BioMolecular Engineering Research Center (Boston University) computing and support staff including Nancy Sands, Venky Iyer (Eisen Lab at University of California, Berkeley) and the Assembly/Alignment/Annotation coordinating committee for releasing GLEAN-R gene predictions, and Stanley Letovsky for discussions and helpful suggestions. This work was supported by a subcontract to Boston University from Harvard University under National Institutes of Health grant HG000739.

## LITERATURE CITED

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

Anderson, A. R., A. A. Hoffmann, S. W. McKechnie, P. A. Umina and A. R. Weeks, 2005 The latitudinal cline in the *In(3R)Payne* inversion polymorphism has shifted in the last 20 years in Australian *Drosophila melanogaster* populations. Mol. Ecol. **14:** 851–858.

Andolfatto, P., and M. Kreitman, 2000 Molecular variation at the *In(2L)t* proximal breakpoint site in natural populations of *Drosophila melanogaster* and *D. simulans.* Genetics **154:** 1681–1691.

Bartolome, C., and B. Charlesworth, 2006 Rates and patterns of chromosomal evolution in *Drosophila pseudoobscura* and *D. miranda.* Genetics **173:** 779–791.

Bellen, H. J., R. W. Levis, G. Liao, Y. He, J. W. Carlson *et al.*, 2004 The BDGP gene disruption project: single transposon insertions associated with 40% of Drosophila genes. Genetics **167:** 761–781.

Beverley, S. M., and A. C. Wilson, 1984 Molecular evolution in Drosophila and the higher Diptera II. A time scale for fly evolution. J. Mol. Evol. **21:** 1–13.

Bhutkar, A., S. Russo, T. F. Smith and W. M. Gelbart, 2006 Techniques for multi-genome synteny analysis to overcome assembly limitations. Genome Inform. **17:** 152–161.

Bhutkar, A., W. M. Gelbart and T. F. Smith, 2007a Inferring genome-scale rearrangement phylogeny and ancestral gene order: a Drosophila case study. Genome Biol. **8:** R236.

Bhutkar, A., S. M. Russo, T. F. Smith and W. M. Gelbart, 2007b Genome scale analysis of positionally relocated genes. Genome Res. **17:** 1880–1887.

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

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

Charlesworth, B., 1974 Inversion polymorphism in a two-locus genetic system. Genet. Res. **23:** 259–280.

Charlesworth, B., and D. Charlesworth, 1973 Selection of new inversion in multi-locus genetic systems. Genet. Res. **21:** 167–183.

Charlesworth, B., J. A. Coyne and N. H. Barton, 1987 The relative rates of evolution of sex chromosomes and autosomes. Am. Nat. **130:** 113–146.

Clark, J. B., P. C. Kim and M. G. Kidwell, 1998 Molecular evolution of P transposable elements in the genus Drosophila. III. The melanogaster species group. Mol. Biol. Evol. **15:** 746–755.

Coghlan, A., and K. H. Wolfe, 2002 Fourfold faster rate of genome rearrangement in nematodes than in Drosophila. Genome Res. **12:** 857–867.

Crew, F. A. E., and R. Lamy, 1935 Linkage groups in *Drosophila pseudo-obscura.* With special notes on homology and the nature of gene action. J. Genet. **30:** 15–29.

Crosby, M. A., J. L. Goodman, V. B. Strelets, P. Zhang and W. M. Gelbart, 2007 FlyBase: genomes by the dozen. Nucleic Acids Res. **35:** D486–D491.

da Cunha, A. B., and T. Dobzhansky, 1954 A further study of chromosomal polymorphism in *Drosophila willistoni* in its relation to the environment. Evolution **8:** 119–134.

da Cunha, A. B., H. Burla and T. Dobzhansky, 1950 Adaptive chromosomal polymorphism in *Drosophila willistoni.* Evolution **4:** 212–235.

da Cunha, A. B., T. Dobzhansky, O. A. Pavlovsky and B. Spassky, 1959 Genetics of natural populations. XXVIII. Supplementary data on the chromosomal polymorphism in *Drosophila willistoni* in its relation to the environment. Evolution **13:** 389–404.

Dobzhansky, T., 1944 Chromosomal races in *Drosophila pseudoobscura* and *Drosophila persimilis.* Carnegie Inst. Wash. Publ. **554:** 47–144.

Dobzhansky, T., 1950 The genetics of natural populations. XIX. Origin of heterosis through natural selection in populations of *Drosophila pseudoobscura.* Genetics **35:** 288–302.

Dobzhansky, T., and A. H. Sturtevant, 1938 Inversions in the chromosomes of *Drosophila pseudoobscura.* Genetics **23:** 28–64.

Donald, H. P., 1936 On the genetical constitution of *Drosophila pseudo-obscura,* race A. J. Genet. **33:** 103–122.

Dorer, D. R., J. A. Rudnick, E. N. Moriyama and A. C. Christensen, 2003 A family of genes clustered at the Triplo-lethal locus of *Drosophila melanogaster* has an unusual evolutionary history and significant synteny with *Anopheles gambiae*. Genetics **165:** 613–621.

Dorman, E. R., A. M. Bushey and V. G. Corces, 2007 The role of insulator elements in large-scale chromatin structure in interphase. Semin. Cell Dev. Biol. **18:** 682–690.

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

Ehrlich, J., D. Sankoff and J. H. Nadeau, 1997 Synteny conservation and chromosome rearrangements during mammalian evolution. Genetics **147:** 289–296.

Engels, W. R., and C. R. Preston, 1984 Formation of chromosome rearrangements by *P* factors in Drosophila. Genetics **107:** 657–678.

Felsenstein, J., 1989 PHYLIP—Phylogeny Inference Package (Version 3.2). Cladistics **5:** 164–166.

Gene Ontology Consortium, 2001 Creating the gene ontology resource: design and implementation. Genome Res. **11:** 1425–1433.

Gonzalez, J., J. M. Ranz and A. Ruiz, 2002 Chromosomal elements evolve at different rates in the Drosophila genome. Genetics **161:** 1137–1154.

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

Goto, S. G., and M. T. Kimura, 2001 Phylogenetic utility of mitochondrial COI and nuclear Gpdh genes in Drosophila. Mol. Phylogenet. Evol. **18:** 404–422.

Harr, B., B. Zangerl and C. Schlotterer, 2000 Removal of microsatellite interruptions by DNA replication slippage: phylogenetic evidence from Drosophila. Mol. Biol. Evol. **17:** 1001–1009.

HGSC, 2006 Insights into social insects from the genome of the honeybee Apis mellifera. Nature **443:** 931–949.

Holt, R. A., G. M. Subramanian, A. Halpern, G. G. Sutton, R. Charlab *et al.*, 2002 The genome sequence of the malaria mosquito Anopheles gambiae. Science **298:** 129–149.

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

Kirkpatrick, M., and N. Barton, 2006 Chromosome inversions, local adaptation and speciation. Genetics **173:** 419–434.

Krimbas, C. B., 1992 The inversion polymorphism of *Drosophila subobscura*, pp. 127–220 in *Drosophila Inversion Polymorphism*, edited by C. B. Krimbas and J. R. Powell. CRC Press, Boca Raton, FL.

Kumar, S., K. Tamura and M. Nei, 2004 MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief. Bioinform. **5:** 150–163.

Lemeunier, F., and M. Ashburner, 1976 Relationships within the melanogaster species subgroup of the genus *Drosophila* (Sophophora). II. Phylogenetic relationships between six species based upon polytene chromosome banding sequences. Proc. R. Soc. Lond. Ser. B **193:** 275–294.

Lim, J. K., 1988 Intrachromosomal rearrangements mediated by hobo transposons in Drosophila melanogaster. Proc. Natl. Acad. Sci. USA **85:** 9153–9157.

Loukas, M., C. B. Krimbas, P. Mavragani-Tsipidou and C. D. Kastritsis, 1979 Genetics of *Drosophila subobscura* populations. VII. Allozyme loci and their chromosomal maps. J. Hered. **70:** 17–26.

Ma, J., L. Zhang, B. B. Suh, B. J. Raney, R. C. Burhans *et al.*, 2006 Reconstructing contiguous regions of an ancestral genome. Genome Res. **16:** 1557–1565.

Machado, C. A., T. S. Haselkorn and M. A. Noor, 2007 Evaluation of the genomic extent of effects of fixed inversion differences on intraspecific variation and interspecific gene flow in *Drosophila pseudoobscura* and *D. persimilis.* Genetics **175:** 1289–1306.

Markow, T. A., and P. M. O'Grady, 2007 *Drosophila* biology in the genomic age. Genetics **177:** 1269–1276.

Mathiopoulos, K. D., and G. C. Lanzaro, 1995 Distribution of genetic diversity in relation to chromosomal inversions in the malaria mosquito Anopheles gambiae. J. Mol. Evol. **40:** 578–584.

Mathiopoulos, K. D., A. Della Torre, V. Predazzi and V. Petrarca, 1998 Cloning inversion breakpoints in the *Anopholes gambiae* complex traces a transposable element at the inversion junction. Proc. Natl. Acad. Sci. USA **95:** 12444–12449.

Mathiopoulos, K. D., A. della Torre, F. Santolamazza, V. Predazzi, V. Petrarca *et al.*, 1999 Are chromosomal inversions induced by transposable elements? A paradigm from the malaria mosquito Anopheles gambiae. Parassitologia **41:** 119–123.

Matzkin, L. M., T. J. Merritt, C. T. Zhu and W. F. Eanes, 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.

Moore, B. C., and C. E. Taylor, 1986 *Drosophila* of southern California. III. Gene arrangements of *Drosophila persimilis.* J. Hered. **77:** 313–323.

Muller, H. J., 1940 Bearings of the 'Drosophila' work on systematics, pp. 185–268 in *The New Systematics*, edited by J. Huxley. Clarendon Press, Oxford.

Nadeau, J. H., and D. Sankoff, 1998a Counting on comparative maps. Trends Genet. **14:** 495–501.

Nadeau, J. H., and D. Sankoff, 1998b The lengths of undiscovered conserved segments in comparative maps. Mamm. Genome **9:** 491–495.

Nadeau, J. H., and B. A. Taylor, 1984 Lengths of chromosomal segments conserved since divergence of man and mouse. Proc. Natl. Acad. Sci. USA **81:** 814–818.

Nene, V., J. R. Wortman, D. Lawson, B. Haas, C. Kodira *et al.*, 2007 Genome sequence of Aedes aegypti, a major arbovirus vector. Science **316:** 1718–1723.

Noor, M. A. F., D. A. Garfield, S. W. Schaeffer and C. A. Machado, 2007 Divergence between the *Drosophila pseudoobscura* and *D. persimilis* genome sequences in relation to chromosomal inversions. Genetics **177:** 1417–1428.

Novitski, E., 1946 Chromosomal variation in *Drosophila athabasca.* Genetics **31:** 508–524.

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

Olvera, O., J. R. Powell, M. E. De La Rosa, V. M. Salceda, M. I. Gaso *et al.*, 1979 Population genetics of Mexican Drosophila VI. Cytogenetic aspects of the inversion polymorphism in *Drosophila pseudoobscura.* Evolution **33:** 381–395.

Ohno, S., 1973 Ancient linkage groups and frozen accidents. Nature **244:** 259–262.

Painter, T. S., 1934 A new method for the study of chromosomal aberrations and the plotting of chromosomal maps in *Drosophila melanogaster.* Genetics **19:** 175–188.

Papaceit, M., and E. Juan, 1998 Fate of dot chromosome genes in *Drosophila willistoni* and *Scaptodrosophila lebanonensis* determined by *in situ* hybridization. Chromosome Res. **6:** 49–54.

Papaceit, M., M. Aguade and C. Segarra, 2006 Chromosomal evolution of elements B and C in the Sophophora subgenus of Drosophila: evolutionary rate and polymorphism. Evolution **60:** 768–781.

Pelandakis, M., and M. Solignac, 1993 Molecular phylogeny of Drosophila based on ribosomal RNA sequences. J. Mol. Evol. **37:** 525–543.

Pelandakis, M., D. G. Higgins and M. Solignac, 1991 Molecular phylogeny of the subgenus Sophophora of Drosophila derived from large subunit of ribosomal RNA sequences. Genetica **84:** 87–94.

Pevzner, P., and G. Tesler, 2003 Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. Proc. Natl. Acad. Sci. USA **100:** 7672–7677.

Powell, J. R., 1997 *Progress and Prospects in Evolutionary Biology: The Drosophila Model.* Oxford University Press, New York.

Powell, J. R., and R. deSalle, 1995 Drosophila molecular phylogenies and their uses, pp. 87–138 in *Evolutionary Biology*, edited by M. K. Hecht, R. J. MacIntyre and M. T. Clegg. Plenum Press, New York.

Ranz, J. M., F. Casals and A. Ruiz, 2001 How malleable is the eukaryotic genome? Extreme rate of chromosomal rearrangement in the genus Drosophila. Genome Res. **11:** 230–239.

Ranz, J. M., J. Gonzalez, F. Casals and A. Ruiz, 2003 Low occurrence of gene transposition events during the evolution of the genus Drosophila. Evolution **57:** 1325–1335.

Ranz, J. M., C. Segarra and A. Ruiz, 1997 Chromosomal homology and molecular organization of Muller's elements D and E in the *Drosophila repleta* species group. Genetics **145:** 281–295.

Ranz, J. M., D. Maurin, Y. S. Chan, M. V. Grotthuss, L. W. Hillier *et al.*, 2007 Principles of genome evolution in the *Drosophila melanogaster* species group. PLoS Biol. **5:** 1366–1381.

Rice, W. R., 1989 Analyzing tables of statistical tests. Evolution **43:** 223–225.

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

Rohde, C., 2000 *Polimorfismo Cromossômico e Elementos Transponíveis em Drosophila willistoni.* Ph.D. Thesis, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil.

Rohde, C., T. H. Degrandi, D. C. de Toni and V. L. S. Valente, 2005 *Drosophila willistoni* polytene chromosomes. I. Pericentric inversion on X chromosome. Caryologia **58:** 249–254.

Russo, C. A., N. Takezaki and M. Nei, 1995 Molecular phylogeny and divergence times of drosophilid species. Mol. Biol. Evol. **12:** 391–404.

Saitou, N., and M. Nei, 1987 The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4:** 406–425.

Sankoff, D., and P. Trinh, 2005 Chromosomal breakpoint reuse in genome sequence rearrangement. J. Comput. Biol. **12:** 812–821.

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

Schoen, D. J., 2000 Comparative genomics, marker density and statistical analysis of chromosome rearrangements. Genetics **154:** 943–952.

Segarra, C., E. R. Lozovskaya, G. Ribo, M. Aguade and D. L. Hartl, 1995 P1 clones from *Drosophila melanogaster* as markers to study the chromosomal evolution of Muller's A element in two species of the obscura group of Drosophila. Chromosoma **104:** 129–136.

Segarra, C., G. Ribó and M. Aguadé, 1996 Differentiation of Muller's chromosomal elements D and E in the obscura group of Drosophila. Genetics **144:** 139–146.

Spassky, B., and T. Dobzhansky, 1950 Comparative genetics of *Drosophila willistoni.* Heredity **4:** 201–215.

Spellman, P. T., and G. M. Rubin, 2002 Evidence for large domains of similarly expressed genes in the Drosophila genome. J. Biol. **1:** 5.

Sperlich, D., and P. Pfriem, 1986 Chromosomal polymorphism in natural and experimental populations, pp. 257–309 in *The Genetics and Biology of Drosophila*, edited by M. Ashburner, H. L. Carson and J. N. Thomson. Academic Press, New York.

Spradling, A. C., D. M. Stern, I. Kiss, J. Roote, T. Laverty *et al.*, 1995 Gene disruptions using P transposable elements: an integral component of the Drosophila genome project. Proc. Natl. Acad. Sci. USA **92:** 10824–10830.

Steinemann, M., W. Pinsker and D. Sperlich, 1984 Chromosome homologies within the *Drosophila obscura* group probed by *in situ* hybridization. Chromosoma **91:** 46–53.

Stolc, V., Z. Gauhar, C. Mason, G. Halasz, M. F. van Batenburg *et al.*, 2004 A gene expression map for the euchromatic genome of *Drosophila melanogaster.* Science **306:** 655–660.

Sturtevant, A. H., and E. Novitski, 1941 The homologies of the chromosome elements in the genus Drosophila. Genetics **26:** 517–541.

Sturtevant, A. H., and C. C. Tan, 1937 The comparative genetics of *Drosophila pseudoobscura* and *D. melanogaster.* J. Genet. **34:** 415–432.

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

Tomancak, P., A. Beaton, R. Weiszmann, E. Kwan, S. Shu *et al.*, 2002 Systematic determination of patterns of gene expression during Drosophila embryogenesis. Genome Biol. **3:** RESEARCH0088.

Tribolium Genome Sequencing Consortium, 2007 The genome of the model beetle and pest *Tribolium castaneum.* Nature **452:** 949–955.

Valente, V. L. S., and A. M. Araujo, 1985 Observations on the chromosomal polymorphism of natural populations of *Drosophila willistoni* and its association with the choice of feeding and breeding sites. Rev. Brasil. Genet. **8:** 271–284.

Valente, V. L. S., and A. M. Araujo, 1986 Chromosomal polymorphism, climatic factors, and variation in population size of *Drosophila willistoni* in southern Brazil. Heredity **57:** 149–159.

Valente, V. L. S., and N. B. Morales, 1985 New inversions and qualitative description of inversion heterozygotes in natural populations of Drosophila willistoni inhabiting two different regions in the State of Rio Grande do Sul, Brazil. Rev. Brasil. Genet. **8:** 167–173.

Valente, V. L. S., A. Ruszczyk and R. A. dos Santos, 1993 Chromosomal polymorphism in urban *Drosophila willistoni.* Rev. Brasil. Genet. **16:** 307–319.

Valente, V. L. S., C. Rohde, V. H. Valiati, N. B. Morales and B. Goni, 2001 Chromosome inversions occurring in Uruguayan populations of *Drosophila willistoni.* Dros. Inf. Serv. **84:** 55–59.

Valente, V. L. S., B. Goni, V. H. Valiati, C. Rohde and N. B. Morales, 2003 Chromosomal polymorphism in Drosophila willistoni populations from Uruguay. Genet. Mol. Biol. **26:** 163–173.

Vieira, J., C. P. Vieira, D. L. Hartl and E. R. Lozovskaya, 1997 Discordant rates of chromosome evolution in the *Drosophila virilis* species group. Genetics **147:** 223–230.

Waddington, D., A. J. Springbett and D. W. Burt, 2000 A chromosome-based model for estimating the number of conserved segments between pairs of species from comparative genetic maps. Genetics **155:** 993.

Wesley, C. S., and W. F. Eanes, 1994 Isolation and analysis of the breakpoint sequences of chromosome inversion *In*(*3L*)*Payne* in *Drosophila melanogaster.* Proc. Natl. Acad. Sci. USA **91:** 3132–3136.

Whiting, J. H., M. D. Pliley, J. L. Farmer and D. E. Jeffery, 1989 *In situ* hybridization analysis of chromosomal homologies in *Drosophila melanogaster* and *Drosophila virilis.* Genetics **122:** 99–109.

Yang, Y., Y. P. Zhang, Y. H. Qian and Q. T. Zeng, 2004 Phylogenetic relationships of Drosophila melanogaster species group deduced from spacer regions of histone gene H2A–H2B. Mol. Phylogenet. Evol. **30:** 336–343.

York, T. L., R. Durrett and R. Nielsen, 2007 Dependence of paracentric inversion rate on tract length. BMC Bioinform. **8:** 115.

Zdobnov, E. M., and P. Bork, 2007 Quantification of insect genome divergence. Trends Genet. **23:** 16–20.

Communicating editor: D. M. Rand