DOI: 10.1534/genetics.110.116129

Evolution of a Distinct Genomic Domain in Drosophila: Comparative Analysis of the Dot Chromosome in *Drosophila melanogaster* and *Drosophila virilis*

Wilson Leung,* Christopher D. Shaffer,* Taylor Cordonnier,*,† Jeannette Wong,* Michelle S. Itano,*,‡ Elizabeth E. Slawson Tempel,* Elmer Kellmann,*,§ David Michael Desruisseau,* Carolyn Cain,*, Robert Carrasquillo,*,†† Tien M. Chusak,*,‡‡ Katazyna Falkowska,* Kelli D. Grim,*,§§ Rui Guan,*,*** Jacquelyn Honeybourne,* Sana Khan,*,††† Louis Lo,* Rebecca McGaha,*,‡‡‡ Jevon Plunkett,*,§§§ Justin M. Richner,*,**** Ryan Richt,* Leah Sabin,*,†††† Anita Shah,*,‡‡‡‡ Anushree Sharma,*,§§§§ Sonal Singhal,*,***** Fine Song,*,††††† Christopher Swope,* Craig B. Wilen,*,†††† Jeremy Buhler,‡‡‡‡‡ Elaine R. Mardis§§§§§ and Sarah C. R. Elgin*,[1]**

†*Ross University School of Medicine, Portsmouth, Commonwealth of Dominica, West Indies 00109-8000,* ‡*Department of Cell and Developmental Biology, University of North Carolina, Chapel Hill, North Carolina 27599,* §*Robller Vineyard Winery, New Haven, Missouri 63068,* ***Department of Human Genetics, University of Chicago, Chicago, Illinois 60637,* ††*Harvard Medical School, Boston, Massachusetts 02115,* ‡‡*Marshall School of Business, University of Southern California, Los Angeles, California 90033,* §§*Baylor College of Medicine, Houston, Texas 77054,* ****College of Medicine, University of Illinois, Chicago, IL 60612,* †††*Department of Obstetrics and Gynecology, University of Oklahoma Health Sciences Center, Oklahoma City, Oklahoma 73104,* ‡‡‡*Department of Epidemiology, M. D. Anderson Cancer Center, University of Texas, Houston, Texas 77030,* §§§*Department of Pediatrics, Vanderbilt University School of Medicine, Nashville, Tennessee 37232,* §§§*Human and Statistic Genetics Program, Washington University School of Medicine, St. Louis, Missouri 63110,* *****Department of Plant and Microbial Biology, University of California, Berkeley, California 94720,* ††††*Department of Microbiology, University of Pennsylvania, Philadelphia, Pennsylvania 19104,* ‡‡‡‡*Western University, Pomona, California 91766,* §§§§*Washington University School of Medicine, St. Louis, Missouri 63110,* ******Museum of Vertebrate Zoology, University of California, Berkeley, California 94720,* †††††*St. Louis University School of Medicine, St. Louis, Missouri 63104,* **Department of Biology, Washington University, St. Louis, Missouri 63130,* ‡‡‡‡‡*Department of Computer Science and Engineering, Washington University, St. Louis, Missouri 63130 and* §§§§§*Department of Genetics, The Genome Center at Washington University School of Medicine, St. Louis, Missouri 63108*

Manuscript received February 27, 2010
Accepted for publication May 9, 2010

ABSTRACT

The distal arm of the fourth ("dot") chromosome of *Drosophila melanogaster* is unusual in that it exhibits an amalgamation of heterochromatic properties (*e.g.*, dense packaging, late replication) and euchromatic properties (*e.g.*, gene density similar to euchromatic domains, replication during polytenization). To examine the evolution of this unusual domain, we undertook a comparative study by generating high-quality sequence data and manually curating gene models for the dot chromosome of *D. virilis* (Tucson strain 15010–1051.88). Our analysis shows that the dot chromosomes of *D. melanogaster* and *D. virilis* have higher repeat density, larger gene size, lower codon bias, and a higher rate of gene rearrangement compared to a reference euchromatic domain. Analysis of eight "wanderer" genes (present in a euchromatic chromosome arm in one species and on the dot chromosome in the other) shows that their characteristics are similar to other genes in the same domain, which suggests that these characteristics are features of the domain and are not required for these genes to function. Comparison of this strain of *D. virilis* with the strain sequenced by the Drosophila 12 Genomes Consortium (Tucson strain 15010–1051.87) indicates that most genes on the dot are under weak purifying selection. Collectively, despite the heterochromatin-like properties of this domain, genes on the dot evolve to maintain function while being responsive to changes in their local environment.

EUKARYOTIC genomes are packaged into two major types of chromatin: euchromatin is gene rich and has a diffuse appearance during interphase, while heterochromatin is gene poor and remains densely packaged throughout the cell cycle (Grewal and Elgin 2002). The distal 1.2 Mb of the fourth chromosome of *Drosophila melanogaster*, known as the dot chromosome or Muller *F* element, is unusual in exhibiting an amalgamation of heterochromatic and euchromatic properties. This domain has a gene density that is similar to the other autosomes (Bartolomé *et al.* 2002; Slawson *et al.*

Supporting information is available online at http://www.genetics.org/cgi/content/full/genetics.110.116129/DC1.

The improved *D. virilis* dot chromosome sequence data and gene annotations have been deposited with the NCBI Entrez Genome Project Database under project ID 41283.

[1]*Corresponding author:* Department of Biology, Campus Box 1137, One Brookings Drive, Washington University, St. Louis, MO 63130-4899. E-mail: selgin@biology.wustl.edu

2006). However, it appears heterochromatic by many criteria, including late replication and very low levels of meiotic recombination (Wang *et al.* 2002; Arguello *et al.* 2010). It exhibits high levels of association with heterochromatin protein 1 (HP1) and histone H3 di- and trimethylated at lysine 9 (H3K9me2/3), as shown by immunofluorescent staining of the polytene chromosomes (Riddle and Elgin 2006; Slawson *et al.* 2006). This association with heterochromatin marks has recently been confirmed by the modENCODE Project [N. C. Riddle, A. Minoda, P. V. Kharchenko, A. A. Alekseyenko, Y. B. Schwartz, M. Y. Tolstorukov, A. A. Gorchakov, C. Kennedy, D. Linder-Basso, J. D. Jaffe, G. Shanower, M. I. Kuroda, V. Pirrotta, P. J. Park, S. C. R. Elgin, G. H. Karpen, and the modENCODE Consortium (http://www.modencode.org), unpublished results]. To understand this unique domain and to examine the evolution of a region with very low levels of recombination, we have undertaken a comparative study using the dot chromosome of *D. virilis*, a species that diverged from *D. melanogaster* 40–60 million years ago (Powell and Desalle 1995). We sequenced and improved the assembly of the *D. virilis* dot chromosome and created a manually curated set of gene models to ensure that both the assembly and the gene annotations are at a quality comparable to those in *D. melanogaster*. We then compared the sequence organization and gene characteristics of the distal portion of the *D. virilis* dot chromosome with the corresponding region from the *D. melanogaster* dot chromosome.

In addition to examining the long-term dot chromosome evolution, we also investigated the short-term dot chromosome evolution by comparing the genomic sequences from two different strains of *D. virilis*. Agencourt Biosciences (AB) has previously produced a whole genome shotgun assembly of Tucson strain 15010–1051.87, while we have sequenced Tucson strain 15010–1051.88 of *D. virilis* [the Genomics Education Partnership (GEP) assembly]. The AB assembly has been improved by the Drosophila 12 Genomes Consortium and released as part of the comparative analysis freeze 1 (CAF1) assembly (Drosophila 12 Genomes Consortium *et al.* 2007).

Using the GEP and CAF1 assemblies from *D. virilis*, and the high-quality *D. melanogaster* assembly and its gene annotations from FlyBase (Crosby *et al.* 2007), we compared the gene properties and sequence organization of the dot chromosomes and reference euchromatic and heterochromatic domains. The dot chromosomes from *D. melanogaster* and *D. virilis* are distinct from the heterochromatic and euchromatic regions of the two genomes, both in organization (*e.g.*, repeat density) and in characteristics of the genes (*e.g.*, size, codon bias). The two dot chromosomes resemble each other by most criteria and differ only in the types of repetitive sequences present and in relative gene order and orientation. Despite the very low rate of meiotic recombination, comparison of the two *D. virilis* strains shows that dot chromosome genes are under weak purifying selection. Our analysis of genes that are present in a euchromatic chromosome arm in one species and on the dot chromosome in the other (the "wanderer" genes) shows that this set of genes evolves to maintain function while responding to the changes in the local chromosomal environment.

MATERIALS AND METHODS

Analysis overview: Sequence analyses were implemented using custom scripts and programs written in Perl, Ruby, R, and Bash shell scripts. Data were stored in subversion repositories, MySQL databases, and plain text files. Graphical and statistical analyses were done using R from the R Project for Statistical Computing and Microsoft Excel. Sequence analysis was carried out on a quad-core server with 8 GB of RAM running openSUSE 11.1. The custom repositories, databases, and scripts are available from the authors upon request.

Sequencing of *D. virilis* fosmid clones: The *D. virilis* fosmid library [Berkeley Drosophila Genome Project (BDGP)-DviF01] was obtained from the BACPAC Resource Center at Children's Hospital Oakland Research Institute (http://bacpac.chori.org/home.htm). Strategies used for isolating, sequencing, and finishing the *D. virilis* clones were previously documented (Slawson *et al.* 2006). All fosmids were improved to the quality standards used for the mouse genome and verified by restriction digests (see supporting information, File S1 for details). The nucleotide sequences and predicted protein sequences reported here have been deposited into the National Center for Biotechnology Information (NCBI) Entrez Genome Project Database under project ID 41283.

Curation strategy: The curation strategy has been documented previously (Slawson *et al.* 2006). Evidence tracks, which were used to create the gene models, included results from multiple gene predictors (Genscan, Twinscan, Geneid, SGP2, SNAP) (Burge and Karlin 1997; Korf *et al.* 2001; Parra *et al.* 2000, 2003; Korf 2004), splice site predictors (Genesplicer) (Pertea *et al.* 2001), and BLAST searches [WUBLAST 2.0MP-WashU (May 4, 2006), http://blast.wustl.edu] against annotated proteins in *D. melanogaster* (FlyBase 5.16) (Crosby *et al.* 2007). Student annotations were loaded into the Apollo Genome Annotation Curation Tool for final quality control and analysis (Lewis *et al.* 2002).

Repeat analysis: RepeatMasker (version open-3.2.7) was run at the most sensitive settings (-s) using the cross_match (version 0.990329) search engine (http://www.repeatmasker.org). Three repeat libraries were used in the analysis: the Drosophila repeats library in Repbase (release 14.03), the Superlibrary (repeats in Repbase release 14.03 with novel repeats identified by PILER-DF), and the species-specific RepeatModeler (beta open-1-0-3) libraries (http://www.repeatmasker.org/RepeatModeler.html). RepeatRunner was run using the Superlibrary and the default repeat protein database from the RepeatRunner package with default parameters (Jurka *et al.* 2005; Smith *et al.* 2007). RepeatModeler was run on the CAF1 *D. virilis* assembly and release 5 of the *D. melanogaster* assembly using the *de novo* repeat finder RECON (release 1.06) and RepeatScout (release 1.05) with default parameters (Bao and Eddy 2002; Price *et al.* 2005). The

repeats found by PILER-DF are available through the FlyBase FTP server (ftp://ftp.flybase.net/genomes/aaa/transposable_elements/PILER-DF/). The species-specific RepeatModeler libraries are available in File S2 (*D. melanogaster*) and File S3 (*D. virilis*).

Tandem repeats were identified using Tandem Repeats Finder (version 3.2.1) (Benson 1999), with the following parameters: matching weight = 2, mismatch penalty = 7, indel penalty = 7, match probability = 80, indel probability = 10, MinScore = 50, and maxPeriod = 2000.

Analysis of gene sizes, coding exon sizes, and intron sizes: In addition to our manual annotations, gene models for the most comprehensive isoform (*i.e.*, the isoform with the largest coding region) in *D. melanogaster* (release 5.16), and *D. virilis* GLEAN-R (Elsik *et al.* 2007) models (release 1.2), were extracted from the precomputed GFF files downloaded from FlyBase (see File S1 for analyses that justify the use of GLEAN-R models). *D. virilis* heterochromatic genes were not considered due to the small number of documented gene models available. To determine the intron size without repeats, intronic sequences were extracted from each model and repetitive sequences were identified using RepeatMasker with the species-specific RepeatModeler library. The unmasked bases were used to calculate the distribution of intron sizes with repeats removed. The significance threshold (alpha = 0.05) has been adjusted using the conservative Bonferroni correction to compensate for multiple pairwise comparisons (*i.e.*, 15): only raw *P*-values less than 3.33*E*-03 (*i.e.*, 0.05/15) are considered to be statistically significant in these analyses (Kolmogorov–Smirnov and Wilcoxon rank sum tests).

Codon bias analysis: In addition to our manual annotations, coding regions for the most comprehensive isoform in *D. melanogaster* and the GLEAN-R *D. virilis* models were extracted from the translation sequence files downloaded from FlyBase. The effective number-of-codons (Nc) statistic (Wright 1990) was calculated using the chips program in the EMBOSS package.

Synteny analysis: Each gene found on the *D. melanogaster* and *D. virilis* dot chromosomes and in the euchromatic reference regions was assigned a unique identifier with its relative orientation. This set of unique identifiers was analyzed using GRIMM with default parameters for the unichromosomal genome through the GRIMM Web interface (Tesler 2002).

***D. virilis* strain comparisons:** The GEP dot chromosome was broken into nine smaller contigs on the basis of the locations of the gaps; these were aligned against the corresponding regions in the CAF1 dot chromosome using the global alignment algorithm stretcher in the EMBOSS package with default parameters (Rice *et al.* 2000). JalView was used to inspect and edit the alignments (Waterhouse *et al.* 2009).

K_a/K_s analysis: A custom BioPerl (Stajich *et al.* 2002) script used ClustalW (version 1.83) (Larkin *et al.* 2007) to generate the global alignments and codeml in the PAML package (version 3.14) to calculate the K_a/K_s ratios (Yang 2007).

RESULTS

The improved *D. virilis* dot chromosome assembly: We previously reported the analysis of 18 *D. virilis* fosmids, 11 (372,650 bp) from the dot chromosome and 7 (273,110 bp) from the major euchromatic chromosome arms (Slawson *et al.* 2006). We have since isolated, sequenced, and improved 29 additional dot chromosome fosmids, bringing the quality of the whole region to the quality standards used for the mouse genome (see File S1 for these criteria). (Undergraduate students carried out the sequence improvement and annotation under the sponsorship of the GEP.) Collectively, the 40 overlapping fosmids assemble into 1,240,624 nonoverlapping base pairs from the *D. virilis* dot chromosome; only eight gaps remain with an estimated total gap size of 14,728 bases (Figure S1). This improved assembly is orthologous to the banded region of the dot chromosome of *D. melanogaster*. A custom version of the University of California Santa Cruz (UCSC) Genome Browser, available at http://gander.wustl.edu (*D. virilis* Manuscript assembly), is used to host the sequence data and evidence tracks used in this study (Kent *et al.* 2002). We identified 81 genes on the dot chromosome of *D. virilis*; 74 have putative orthologs among the 83 genes on the *D. melanogaster* dot chromosome (FlyBase Release 5.16).

In situ hybridization results using several fosmids from the GEP assembly (Slawson *et al.* 2006) place the centromere to the left of the GEP *D. virilis* dot chromosome scaffold and the telomere to the right. The CAF1 scaffold_13052 is anchored in the same relative orientation (Schaeffer *et al.* 2008). Although the entire CAF1 scaffold_13052 consists of 2,019,633 bases, the initial 600,000 bases and the final 200,000 bases have very poor quality. These regions collectively account for 244,702 out of 246,340 (99.3%) of the gap bases in the scaffold and show no sequence homology to the banded portion of the *D. melanogaster* dot chromosome. Since we cannot discount the possibility that these regions of the CAF1 scaffold_13052 have been misassembled, we only analyzed the region that spans from the most proximal to the most distal annotated genes (from 600,384 to 1,826,586 bp). A dot plot analysis shows that this region corresponds to the portion of the GEP strain dot chromosome we have sequenced and annotated and that the two strains of *D. virilis* have a high degree of sequence similarity (Figure S2).

In addition to improving the *D. virilis* dot chromosome assembly, we also created manually curated gene models for this region using results from several gene predictors and homology to the putative *D. melanogaster* orthologs. For each gene, we have consistently annotated the most comprehensive isoform (*i.e.*, the isoform in *D. melanogaster* with the largest coding region). Our analysis has focused only on the coding regions because we cannot definitively annotate the untranslated regions in the *D. virilis* gene models, due to the sparse amount of expression data available.

As previously reported, the manually curated Repbase library has a strong bias toward repeats that are found in *D. melanogaster* (Slawson *et al.* 2006). To alleviate this bias, we generated three repeat libraries on the basis of the genomic assemblies of the two species. The first, Superlibrary, is a library of all previously reported

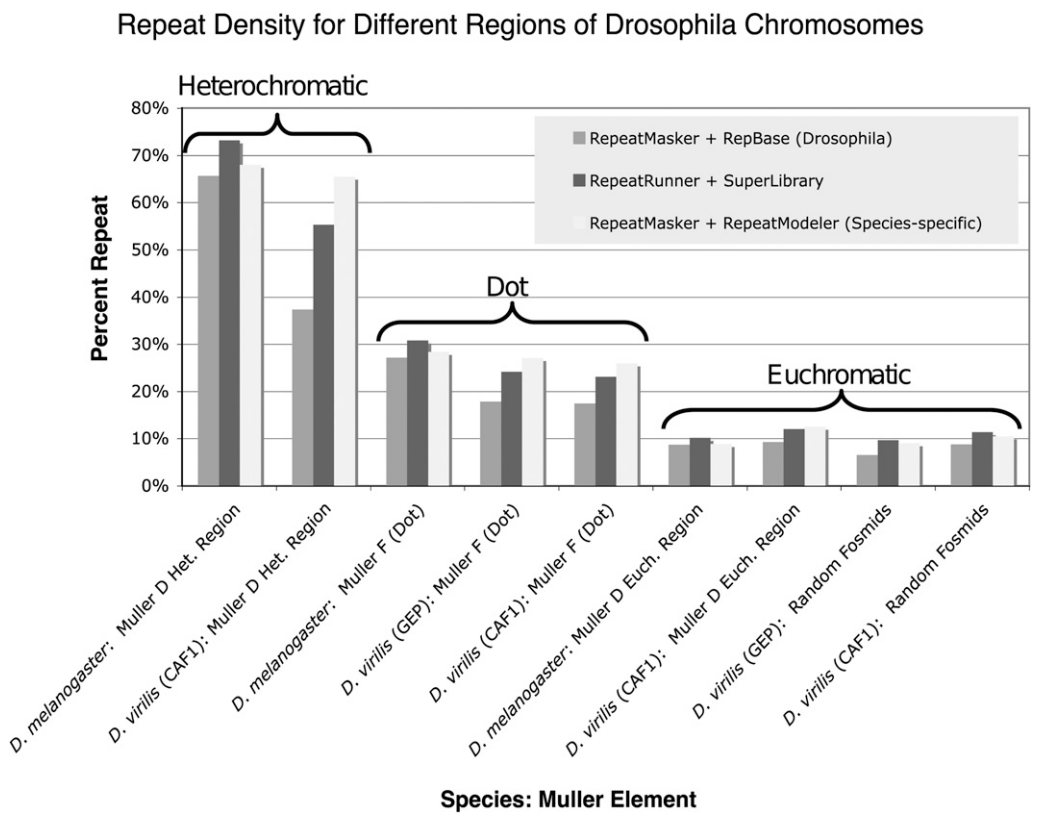


FIGURE 1.—Total repeat density for different domains in the *D. virilis* and *D. melanogaster* genomes. The dot chromosomes from *D. virilis* and *D. melanogaster* are similar to each other and have a higher repeat density than the euchromatic reference regions and a lower repeat density than the heterochromatic reference regions.

repeats, combining annotated repeats from the Drosophila Repbase Update with novel repeats in the 12 Drosophila species found by the *de novo* repeat finder PILER-DF (EDGAR and MYERS 2005; SMITH *et al.* 2007). Two species-specific repeat libraries were created with RepeatModeler (http://www.repeatmasker.org/RepeatModeler.html) and these later libraries were each used to analyze their respective genomes. However, despite our efforts to minimize bias in the repeat libraries, one cannot completely eliminate it using these approaches. The *D. melanogaster* assembly includes additional sequences from the Drosophila Heterochromatin Genome Project (HOSKINS *et al.* 2007). Because there has been no corresponding effort to improve the heterochromatic regions in *D. virilis*, the *D. melanogaster* repeat libraries may be more comprehensive than those for *D. virilis*.

Collectively, the improved assembly, the manually annotated gene set, and the custom repeat libraries provide a unique resource to study the organization and evolution of the Drosophila dot chromosome. Using these resources from *D. virilis* and the high-quality sequence and annotations available for *D. melanogaster*, we seek to characterize the properties of this unique genomic region and to identify the forces that impact the evolution of this domain and the genes that reside within it.

Defining the euchromatic and heterochromatic reference regions: To analyze the differences between the dot chromosomes and other regions of the Drosophila genome, we selected euchromatic and heterochromatic reference regions from both *D. melanogaster* and *D. virilis* for comparative study. We also utilized previously improved and annotated fosmids from euchromatic regions of the *D. virilis* genome and the corresponding regions from *D. melanogaster* (SLAWSON *et al.* 2006) to ensure that the reference regions we picked are representative of "typical" euchromatic regions in the two genomes.

Using the previously defined heterochromatin–euchromatin boundary for chromosome (chr) 3L (Muller *D* element) of *D. melanogaster* (HOSKINS *et al.* 2007), we extracted a 1.25-Mb region distal to that boundary (toward the telomere) as a representative euchromatic region (chr 3L: 21,705,576–22,955,575), and the adjacent proximal 1.25-Mb region (toward the centromere) as a representative heterochromatic region (chr 3L: 22,955,576–24,205,575). Because there is no defined heterochromatin–euchromatin boundary for the *D. virilis* CAF1 assembly, we selected a scaffold (scaffold_13049) mapped to the Muller *D* element and identified heterochromatic and euchromatic domains on the basis of differences in gene and repeat density (Figure S3). The changes in these parameters near the boundary were more gradual in *D. virilis* than in *D. melanogaster*. Therefore, we picked a region ~625 kb distal to the boundary to ensure that its properties reflect those of typical euchromatin in *D. virilis*. We also selected a 0.8-Mb region (from the boundary to the end of scaffold_13049) as a representative heterochromatic region. Both of these regions are high quality (12 gaps, total estimated gap size

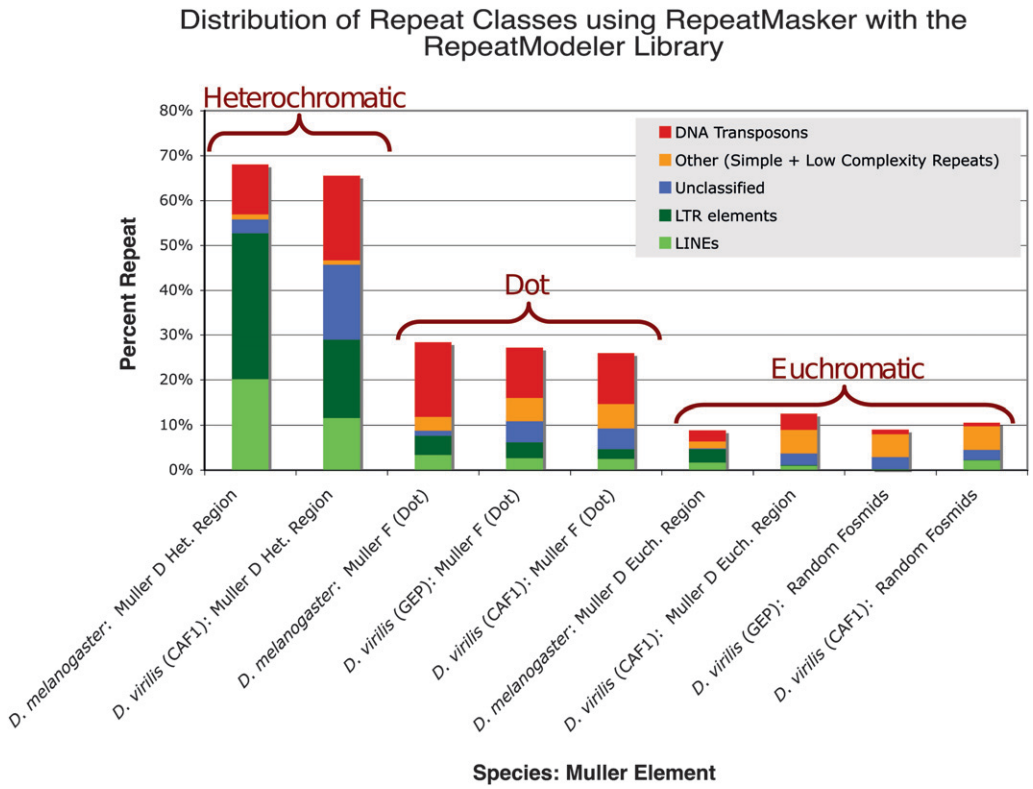


FIGURE 2.—Distribution of repeat classes using RepeatMasker with the RepeatModeler Library. A higher density of DNA transposons is present on the dot chromosomes than in the euchromatic regions for both *D. melanogaster* and *D. virilis*. In contrast, heterochromatic regions show a higher density of LINEs and LTR elements. The *D. virilis* dot chromosome and euchromatic reference regions exhibit a higher density of simple and low complexity repeats compared to the corresponding regions in *D. melanogaster*.

13,612 bases in the euchromatic region; 12 gaps, total estimated gap size 37,211 bases in the heterochromatic region). Using these sequences as reference points, we asked whether the dot chromosomes more closely resemble the heterochromatic or the euchromatic domains for each property under investigation.

The dot chromosomes have a repeat density intermediate between that found in euchromatin and that in pericentric heterochromatin: A well-established characteristic of heterochromatin is a high level of repetitive DNA (GREWAL and ELGIN 2002). The *D. melanogaster* dot chromosome is unusual in that, while its gene density is similar to the other autosomes, its repeat density is much higher than the other nonpericentromeric regions in the *D. melanogaster* genome (BERGMAN *et al.* 2006). To determine whether the *D. virilis* dot chromosome has a similar repeat density, we used RepeatMasker with three custom repeat libraries (described above) to analyze it and the reference heterochromatic and euchromatic regions.

Results from RepeatMasker using the Repbase library initially suggested that the dot chromosome of *D. melanogaster* has a higher repeat density than that of *D. virilis*. However, the *D. melanogaster* and *D. virilis* dot chromosomes show similar overall repeat densities (Figure 1, Table S1) when we used the more comprehensive repeat libraries (*i.e.*, RepeatRunner with the Superlibrary or RepeatMasker with the species-specific RepeatModeler libraries). These findings suggest that the Repbase library has a strong bias toward repeats in *D. melanogaster*.

Both dot chromosomes show repeat densities (26–28%) that are higher than the euchromatic reference regions (9–13%) and lower than the heterochromatic reference regions (66–68%) (Figure 1). This difference is consistent with our previous report based on a limited set of *D. virilis* fosmids (SLAWSON *et al.* 2006). The euchromatic reference region from *D. virilis* has a slightly higher repeat density (13%) than that from *D. melanogaster* (9%). This difference in the euchromatic reference regions is most pronounced when we use the least biased (RepeatModeler) species-specific libraries. Due to the difficulties of accurately defining repeat boundaries and heuristics used by repeat-finding algorithms (BAO and EDDY 2002), we cannot discern whether these small differences in repeat density are significant. However, our results are consistent with those obtained by the Drosophila 12 Genomes Consortium, which reported a higher overall repeat density in the *D. virilis* genome assembly compared to *D. melanogaster* (DROSOPHILA 12 GENOMES CONSORTIUM *et al.* 2007). The CAF1 and GEP dot chromosomes show a similar overall repeat density (27 *vs.* 28%), even though the CAF1 dot chromosome is unfinished while the GEP dot chromosome is improved to the mouse genome standard.

We used the classifications from the species-specific RepeatModeler libraries to analyze the distribution of different classes of repeats, since these libraries have the least bias. The *D. melanogaster* dot is enriched in DNA transposons and retroelements, and the *D. virilis* dot is enriched in simple repeats and low complexity sequences (Figure 2, Table S2). The difference in total repeat

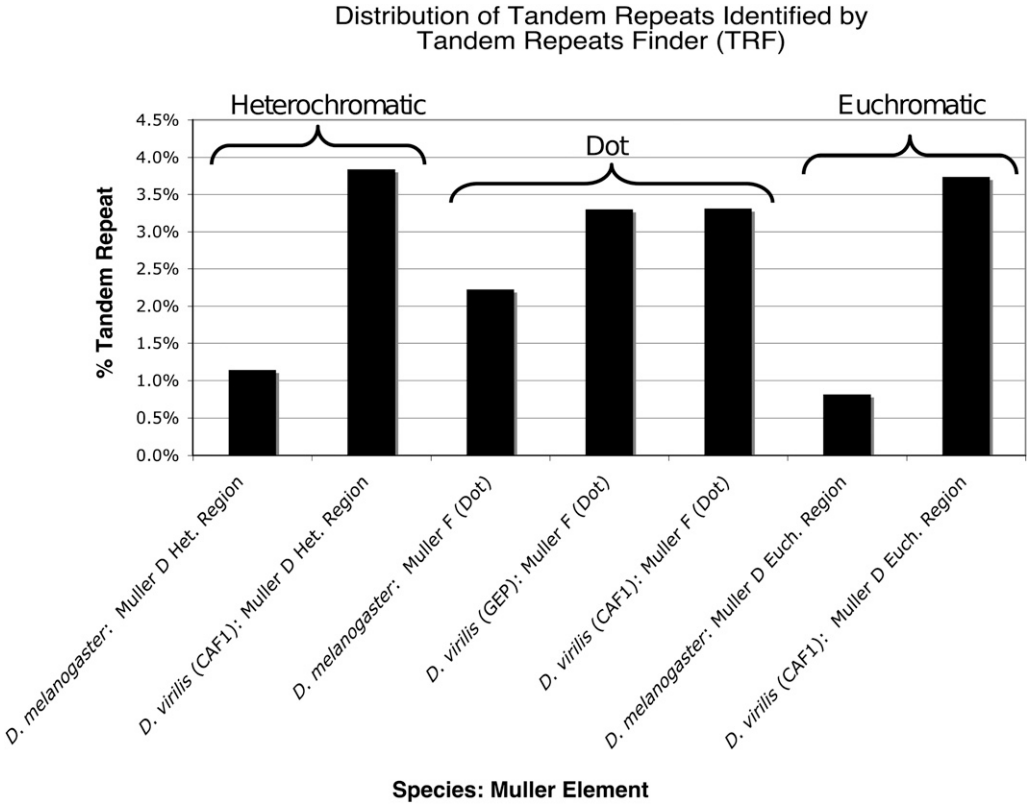


FIGURE 3.—Distribution of tandem repeats identified by Tandem Repeats Finder (TRF). A higher density of tandem repeats is identified in all three regions (heterochromatic reference, dot chromosome, and euchromatic reference) in the *D. virilis* genome compared to the corresponding regions in the *D. melanogaster* genome. The *D. melanogaster* dot chromosome has a higher density of tandem repeats compared to its heterochromatic and euchromatic reference regions.

density between the dot chromosomes and the euchromatic regions can be attributed to the higher density of DNA transposons and retroelements on the dot chromosomes. In contrast, the difference in total repeat density between the heterochromatic reference regions and the dot chromosomes can primarily be attributed to an increase in retroelements. Given the high density of simple and low complexity sequences in *D. virilis*, we next investigated tandem repeats using Tandem Repeats Finder (TRF) (BENSON 1999). We found that all domains in *D. virilis* have higher densities of tandem repeats compared to the corresponding regions in *D. melanogaster* (Figure 3, Table S3).

The dot chromosomes have larger genes, reflecting larger introns, compared to euchromatic domains: The higher repeat density on the dot chromosomes suggests the possibility of larger genes as a consequence of larger introns. A cumulative distribution plot of gene sizes (limited to the region from the start codon to the stop codon) show larger gene sizes on the *D. melanogaster* and *D. virilis* dot chromosomes compared to the euchromatic reference regions (Figure 4A; summary statistics in Table S4). Side-by-side boxplots show that the median gene size and the interquartile range (IQR) are larger on the dot chromosomes compared to the euchromatic reference regions (Figure S4A). The nonparametric Kolmogorov–Smirnov (KS) test shows that this difference in gene size is statistically significant (see raw *P*-values in Table S5). The difference in the distribution of gene sizes between the dot chromosome and the heterochromatic reference region was not statistically significant, which might be partially attributed to the smaller number of genes (21) documented in the latter domain. The distribution of gene sizes between the genes on the *D. melanogaster* and the *D. virilis* dot chromosomes is not significantly different (KS test raw *P*-value = 0.91). Similarly, we found no significant differences (raw *P*-value = 0.60) in the distribution of gene sizes between the manually annotated gene models for the GEP strain and the computationally predicted GLEAN-R gene models for the CAF1 strain. Thus the dot chromosome genes from the two species are similar to each other, and significantly larger than euchromatic genes.

To investigate factors that might contribute to the differences observed in gene size, we examined the size distributions of both the individual coding DNA sequences (CDSs) that make up the translated exons for each gene and the introns. The cumulative distribution plot of CDS sizes shows that the CDSs on the dot chromosome tend to be smaller than those in the euchromatic reference regions and larger than those in the *D. melanogaster* heterochromatic reference region (Figure 4B, summary statistics in Table S6). Differences in CDS sizes between the euchromatic reference regions and the dot chromosomes are statistically significant using the KS test (see boxplots in Figure S4B, raw *P*-values in Table S7). There are no significant differences in CDS sizes between the GEP and CAF1 *D. virilis* dot chromosomes or between the *D.*

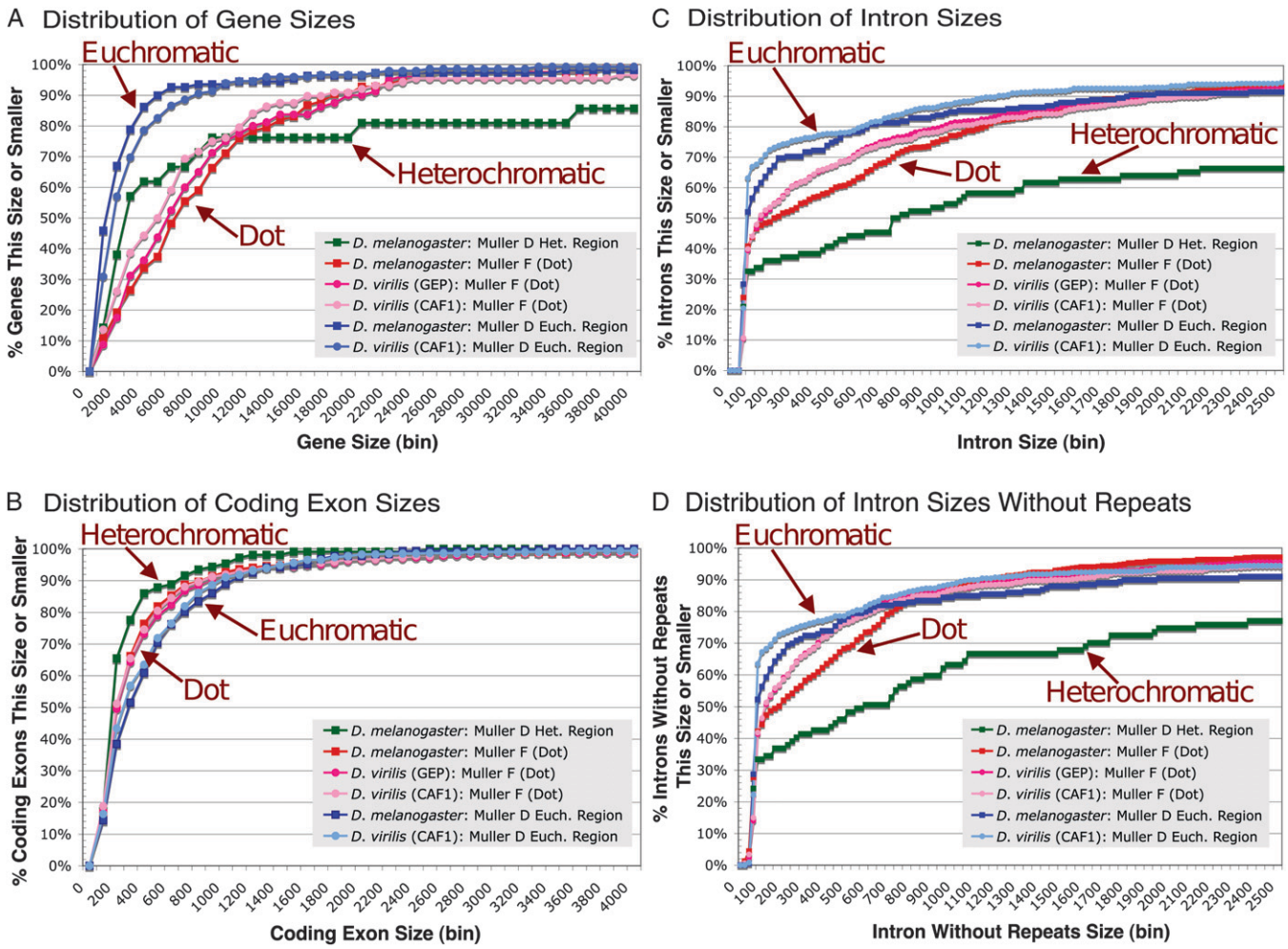


FIGURE 4.—Distribution of gene sizes, coding exon sizes, intron sizes, and intron sizes without repeats. The graphs show empirical cumulative distribution plots for these features on the *D. melanogaster* and *D. virilis* dot chromosomes as well as the euchromatic and heterochromatic reference regions. (A) Genes on the dot chromosomes (from the start codon to the stop codon) are larger than genes from the euchromatic reference regions. (B) Coding exons on the dot chromosome tend to be slightly larger than coding exons in the heterochromatic reference region and slightly smaller than the coding exons in the euchromatic reference regions. (C) Introns on the dot chromosome are significantly smaller than the introns in the heterochromatic reference region and larger than introns in the euchromatic reference region. (D) Removing the repeats from introns reduces but does not eliminate this difference.

melanogaster and GEP *D. virilis* dot chromosomes (raw *P*-values = 0.99 and 0.72, respectively). Because CDSs on the dot chromosomes are generally smaller than those in the euchromatic reference regions, the larger overall gene size on the dot chromosomes must reflect larger intron sizes.

The cumulative distribution plot of intron sizes shows that the introns on the dot chromosomes are generally larger than the introns in the euchromatic reference regions, but smaller than the introns in the heterochromatic reference region (Figure 4C, summary statistics in Table S8). The one-sided Wilcoxon rank sum tests show these differences to be significant (see boxplot in Figure S4C, raw *P*-values in Table S9). There is no significant difference in intron sizes between the *D. melanogaster* and the *D. virilis* dot chromosomes (raw *P*-value = 0.90). Hence differences in intron sizes contribute to the differences in gene sizes observed between euchromatic and dot chromosome domains.

To ascertain whether the difference in intron sizes could be explained by the higher repeat density in the dot chromosomes and the heterochromatic reference regions compared to the euchromatic reference regions, we analyzed the size of the introns after repeats are removed. The cumulative distribution plot (Figure 4D) suggests that the differences in intron sizes are generally less pronounced (increase in *P*-value) but remain statistically significant (summary statistics in Table S10, boxplot in Figure S4D, Wilcoxon rank sum tests of significance in Table S11). The exception is the comparison between the introns for the *D. melanogaster* dot chromosome and the euchromatic reference regions from both *D. melanogaster* and *D. virilis*, where the raw *P*-values rose above the threshold considered to be

statistically significant (from 2.18*E*-03 to 6.90*E*-02 and 3.32*E*-05 to 3.61*E*-03, respectively). Therefore, the higher repeat density within introns is one of the factors that contribute to the larger intron sizes observed on the dot chromosomes and in the heterochromatic reference region compared to the euchromatic reference regions, but is not the sole factor that leads to this result.

Dot chromosome genes exhibit low codon bias: The effective Nc is a simple metric for measuring the usage of synonymous codons (Wright 1990); its value ranges from 61 (all synonymous codons are used equally) to 20 (1 codon used exclusively for each amino acid). Genes with high codon bias have a low Nc while genes with low codon bias have a high Nc.

Side-by-side boxplots of Nc show that genes on the dot chromosome have a significantly lower codon bias than genes in the euchromatic reference regions (Figure 5, summary statistics in Table S12). Two-sided KS tests show that this difference is statistically significant (see raw *P*-values in Table S13). Genes on the *D. virilis* and the *D. melanogaster* dot chromosomes also have a statistically significant different codon bias (raw *P*-value = 1.53 *E*-06). Side-by-side boxplots of Nc show that genes in the *D. melanogaster* heterochromatic domain appear to have an intermediate distribution of codon usage between that seen for the dot chromosome genes and that seen for euchromatic genes, with the former but not the latter difference being statistically significant (raw *P*-values = 3.42*E*-05, 2.01*E*-04 compared to raw *P*-values = 3.61*E*-02, 5.37*E*-02). Our observations of low codon bias on the dot chromosomes are consistent with previous reports on codon bias in Drosophila (Singh *et al.* 2005; Drosophila 12 Genomes Consortium *et al.* 2007) and may reflect the low level of recombination in this domain (see discussion).

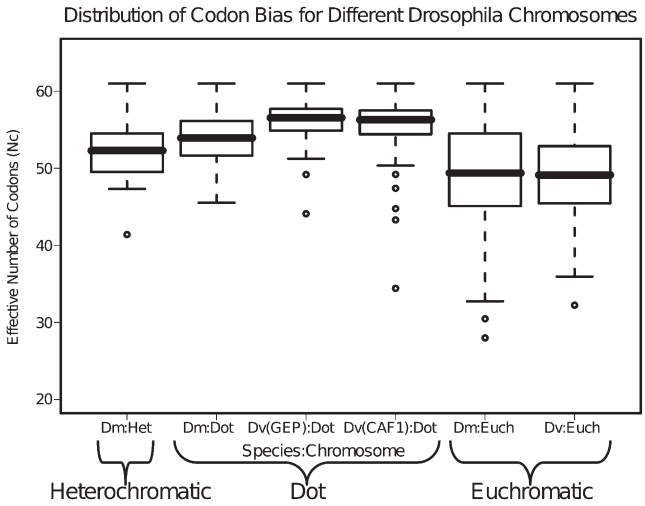


Figure 5.—Side-by-side boxplots of codon bias (as measured by effective number of codons [Nc]). Genes on the dot chromosomes exhibit lower codon bias than those in the euchromatic reference regions. The boxplots also show that genes on the *D. virilis* dot chromosome have lower codon bias than genes on the *D. melanogaster* dot chromosome. The box in each boxplot represents the interquartile range (IQR) and is the difference between the third (Q3) and first (Q1) quartile (IQR = Q3–Q1). Outliers are defined by Nc values that are more than 1.5 × IQR below Q1 or 1.5 × IQR larger than Q3 and are represented as dots in the boxplots. The smallest and largest values that are not outliers are shown as whiskers in the boxplots.

Differences in gene order and orientation between the dot chromosomes of *D. melanogaster* and *D. virilis* indicate rearrangements within the chromosomes: Because we have assembled the *D. virilis* dot chromosome from a set of overlapping fosmids, each finished to high quality and verified by restriction digests, we can approach an analysis of synteny (gene order and orientation) with confidence. Specifically, we can estimate the minimum number of inversions (*i.e.*, the reversal distance) required to transform the *D. melanogaster* dot chromosome into the *D. virilis* dot chromosome using the program GRIMM (Tesler 2002) and can identify genes that are located on the dot chromosome in one species and on another chromosome in the other. Using the set of 74 genes that can be found on both the *D. virilis* and *D. melanogaster* dot chromosomes, GRIMM predicts that a minimum of 33 inversions are required to transform the gene order and relative orientation on the *D. melanogaster* dot chromosome into that observed on the *D. virilis* dot chromosome (Figure 6).

To determine whether this number of inversions is unusual, we utilized the GLEAN-R ortholog assignments from FlyBase for *D. virilis* and calculated the ratio of the reversal distance to the number of genes in a euchromatic domain. We first ascertained whether the GLEAN-R ortholog assignments are adequate for this type of analysis by comparing the dot chromosome GRIMM results obtained with the manually curated dataset (GEP assembly) to results obtained using the GLEAN-R ortholog assignments (CAF1 assembly). Of the 69 genes on the *D. melanogaster* dot chromosome where GLEAN-R ortholog assignments are available in *D. virilis*, 66 remain on the *D. virilis* dot chromosome. GRIMM estimated that a minimum of 27 inversions are needed to transform the gene order and orientation for this subset of genes from that of *D. melanogaster* to that of *D. virilis*. This reversal distance to gene ratio (27/66 = 0.41) is similar to the ratio obtained using the complete set of genes (33/74 = 0.45), establishing that the GLEAN-R ortholog assignments are adequate for this type of analysis.

For the 1906 genes found (using the GLEAN-R ortholog assignment) on both the *D. melanogaster* and *D. virilis* Muller *D* elements (~25-Mb region), GRIMM estimated that a minimum of 385 inversions is required to transform the gene order and orientation from that of *D. melanogaster* to that observed in *D. virilis*. Hence the reversal distance to gene ratio (385/1906 = 0.20) on the Muller *D* element is much lower than the ratio observed on the Muller *F* element (dot chromosome).

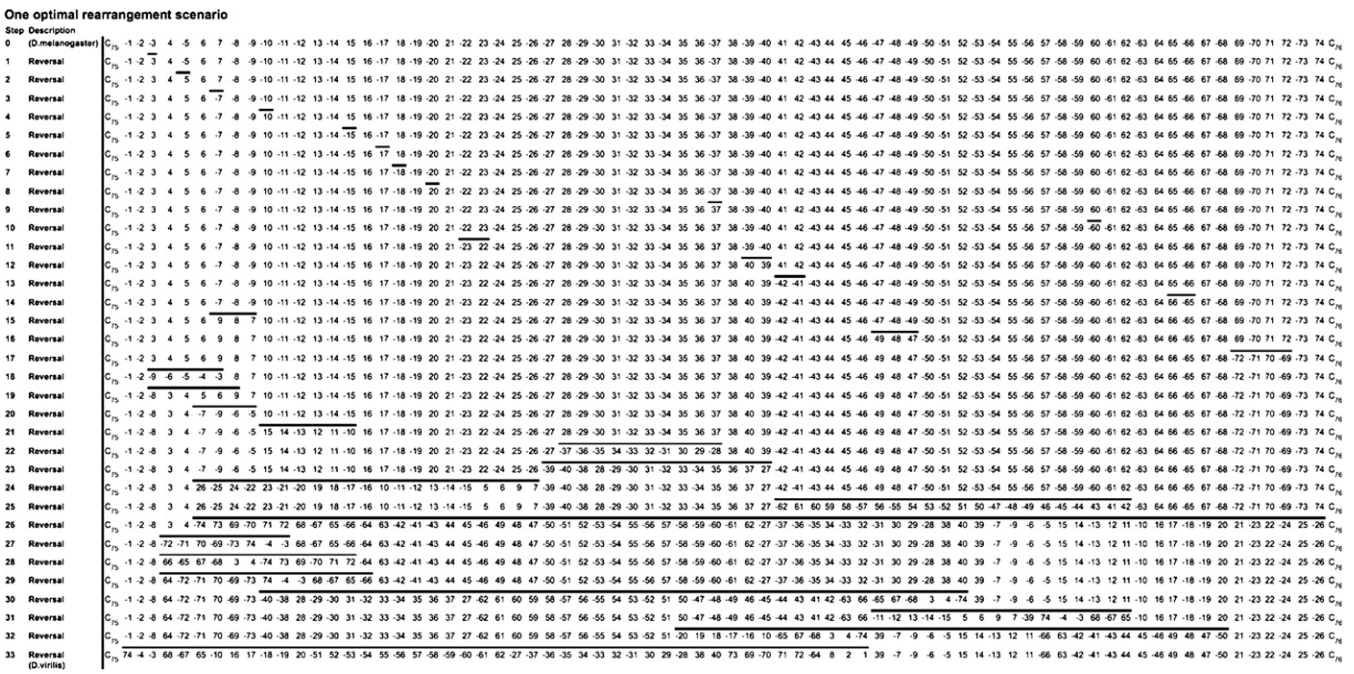


FIGURE 6.—One possible gene rearrangement scenario predicted by GRIMM. A minimum of 33 inversions is required to transform the *D. melanogaster* dot chromosome (top) into the *D. virilis* dot chromosome (bottom). Lines between each inversion step in the GRIMM output indicate the list of genes that is inverted in each step in one optimal rearrangement scenario. Alternative rearrangement scenarios may exist that would also result in this minimum number of inversions.

To confirm this striking difference between the dot chromosome and other elements, we determined the size of the syntenic blocks by visual inspection. The majority of the syntenic blocks on the dot chromosome are small, with an average block size of 1.9. We found 14 syntenic blocks with block sizes of at least 2; the largest syntenic block is 9. In contrast, the syntenic blocks on the Muller *D* element are much larger, with the largest syntenic block being 42 and an average block size of 8.8. The smaller syntenic blocks on the dot chromosomes are consistent with its higher rate of gene rearrangements.

Genes that have moved between the dot and euchromatic domains show a shift in gene characteristics that reflects the local chromatin environment. Our manual annotations identified three potentially novel genes on the *D. virilis* dot chromosome: A putative paralog of *D. melanogaster CG16719* (*CG16719-alpha*), a putative paralog of *D. melanogaster eIF-5A* (*eIF-5A-beta*), and a novel gene (*GEP001*). See File S1 for additional details on the annotation of these proposed novel genes.

We also identified four putative orthologs on the *D. virilis* dot chromosome that are located elsewhere in the *D. melanogaster* genome: One gene on the *D. melanogaster* Muller *D* (chr 3L) element (*CG5262*), two on Muller *B* (chr 2L) element (*CG5367* and *rho-5*), and one on Muller *C* (chr 2R) element (*CG4038*). Conversely, nine genes annotated on the *D. melanogaster* dot chromosome cannot be found on the *D. virilis* dot chromosome. Of these, four (*CG11076*, *CG11077*, *CG1732*, and *CG9935*) can be mapped to other *D. virilis* Muller elements in the CAF1 assembly. Of the remaining five, *JY-alpha* is an incomplete gene on the *D. melanogaster* dot chromosome and cannot be definitively mapped onto the *D. virilis* CAF1 assembly. Three other proposed genes (*CG11231*, *CG11260*, and *CG32021*) are likely to be remnants of repetitive elements that have been incorrectly annotated as genes. The remaining gene, *CG33797*, cannot be mapped to the *D. virilis* CAF1 assembly by sequence similarity. Hence *CG33797* could be a *D. melanogaster* specific gene; it could be present in other species but in regions that are not part of the CAF1 assembly (*e.g.*, in gaps or heterochromatic regions), or it could be an error in the *D. melanogaster* annotation. See File S1 for details on the annotation of these missing genes.

Thus among the above cases there are eight genes from *D. melanogaster* and *D. virilis* that can be unambiguously determined to reside on the dot chromosome in one species and on a non-dot chromosome in the other species (Figure 7). We can ask whether these wanderer genes (*CG11076*, *CG11077*, *CG1732*, *CG4038*, *CG5262*, *CG5367*, *rho-5*, and *CG9935*) show altered properties in the two species as a consequence of residing in the dot chromosome or in a euchromatic domain (Table 1). Consistent with the overall characteristics of the dot chromosome genes reported above, when these genes are found on the dot chromosomes of either species they have a larger average gene size (3099 bp *vs.* 2375 bp), a larger average intron size (1476 bp *vs.* 756 bp), a lower codon bias (Nc = 56.1 *vs.* 51.7), and a higher surrounding repeat density (19.4 *vs.* 6.0%). Thus these wanderer genes have evolved to maintain function in

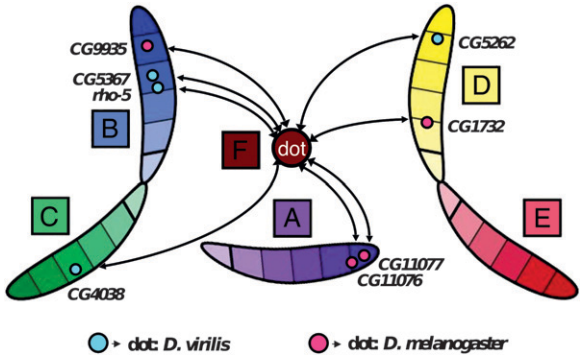


FIGURE 7.—Eight "wanderer" genes can be mapped to the dot chromosome in one species and to a euchromatic region in the other species. The Muller elements are identified by schematic chromosomes that are color coded and labeled A–F. Blue dots indicate genes on the *D. virilis* dot chromosome that are mapped to a different Muller element in *D. melanogaster*. Pink dots indicate genes on the *D. melanogaster* dot chromosome that are mapped to a euchromatic arm in *D. virilis*.

their new local environment while acquiring the characteristics of genes in that environment.

Comparison of two different strains of *D. virilis* shows transposon movement and small indels: Previous studies have demonstrated that the lack of crossing over in the *D. melanogaster* dot chromosome results in a less effective adaptive response to selection (MARAIS and CHARLESWORTH 2003; HADDRILL *et al.* 2007). To see whether the same evolutionary pattern exists in the *D. virilis* dot chromosome, we compared the genomic sequences of the GEP and CAF1 strains. Because the ancestral sequence cannot be determined without a third outgroup, sequences found only in either the GEP (Tucson strain 15010–1051.88) or the CAF1 (Tucson strain 15010–1051.87) strains are labeled as indels (insertions or deletions).

Sequence comparison of the GEP and CAF1 strains of *D. virilis* shows that >70% of the indels are small (<10 bp) (Figure 8). All of the large indels (>1000 bp) have sequence similarity with LTR retroelements and DNA transposons in the Superlibrary (Table 2). Because the long terminal repeats of an LTR retroelement are identical at the time of integration, the age of an LTR insertion event can be estimated using the percent identity of the terminal repeats (SANMIGUEL *et al.* 1998; LAMB *et al.* 2007). The long terminal repeats can be identified in six of these indels (four in the GEP strain and two in the CAF1 strain): Five have perfect sequence

TABLE 1

Characteristics of wanderer genes on the dot and non-dot chromosomes

Gene	Species	No. of exons	Coding gene size (aa)	Total gene length (nt)	Total intron size (nt)	Repeat density (%)	Codon bias (Nc)	GC content
A. Wanderer genes on the dot chromosome								
CG11076[a]	*D. mel.*	1	280	840	0	11.22	61.000	0.40
CG11077[a]	*D. mel.*	1	168	504	0	8.10	50.493	0.40
CG1732	*D. mel.*	10	636	4676	2774	34.86	57.385	0.36
CG4038[b]	*D. vir.*	3	208	960	336	10.78	57.410	0.40
CG5262	*D. vir.*	4	505	2016	501	1.44	58.310	0.34
CG5367	*D. vir.*	5	336	3401	2394	39.40	57.647	0.35
rho-5[b]	*D. vir.*	6	1531	5458	867		54.564	0.45
CG9935	*D. mel.*	10	669	6940	4939	30.00	51.936	0.35
Average		5.00	541.63	3099.38	1476.38	19.40	56.093	0.38
B. Wanderer genes on other Muller elements (euchromatic regions)								
CG11076[a]	*D. vir.*	2	317	1008	55	0.00	47.906	0.43
CG11077[a]	*D. vir.*	1	169	507	0	0.00	46.021	0.40
CG1732	*D. vir.*	11	635	4140	2243	27.18	58.412	0.35
CG4038	*D. mel.*	3	237	1105	393	2.52	49.557	0.41
CG5262	*D. mel.*	4	509	2011	484	0.99	54.573	0.40
CG5367	*D. mel.*	5	338	1467	454	2.35	52.942	0.36
rho-5	*D. mel.*	6	1429	4722	438	0.68	46.912	0.49
CG9935	*D. vir.*	10	688	4039	1981	14.48	56.968	0.36
Average		5.25	540.25	2374.88	756.00	6.03	51.661	0.40

Characteristics of the set of eight genes from *D. melanogaster* and *D. virilis* that can be unambiguously mapped to the dot chromosome (A) in one species and to a non-dot chromosome (B) in the other species show that this set of genes has conformed to its local environment. The genes on the dot chromosomes for both species show lower average codon bias, higher repeat density, larger introns, and larger gene size compared to the genes in the other (euchromatic) regions. Note that the gene model for CG4038 is nested within the gene model for *rho-5* in the *D. virilis* dot chromosome (A). Therefore, the repeat density for *rho-5* has been omitted to avoid double counting the repeats found in this region. *D. mel.*, *D. melanogaster*; *D. vir.*, *D. virilis*.

[a] *CG11076* and *CG11077* appear to have moved as a pair.

[b] The gene model for *CG4038* is nested within the gene model for *rho-5* in *D. virilis*; therefore, the repeat density for this region is only counted once.

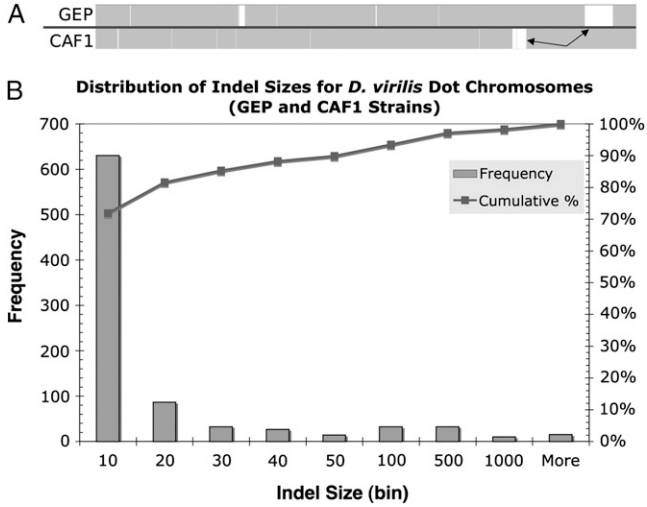


Figure 8.—Distribution of indel sizes for the *D. virilis* dot chromosomes. (A) Examples of large insertions and deletions in the GEP and CAF1 strains in the JalView overview window. Gaps (white spaces) in the overview window represent indels in either *D. virilis* strain. The arrows indicate large indels that are in either the GEP strain (top) or the CAF1 strain (bottom). (B) The cumulative distribution of indel sizes shows that the majority of the indels in the two strains of *D. virilis* are <10 bp.

identity and one has 99.7% (2113/2120 bp) sequence identity (Table 2, Figure S5). While we cannot generate an accurate estimate of the age of these large LTR indel events (because of the small number of mismatches), these large indels likely reflect recent evolutionary events while others (*e.g.*, where no terminal repeats can be identified) may be older.

The six indels with identifiable long terminal repeats have sequence similarity to five different consensus sequences in the Superlibrary (*Ulysses_I*, *dvir.2.37.centroid*, *dvir.2.53.centroid*, *dvir.5.67.centroid*, and *dvir.35.83.centroid*). Four of the five consensus sequences belong to the Gypsy family of LTR retrotransposons (see File S1 for the annotation of these consensus sequences). Hence our analysis suggests a recent invasion of Gypsy LTR retrotransposons into the two *D. virilis* genomes and that insertion and excision of transposons is an integral part of the evolution of the dot chromosomes in *D. virilis*.

Indel to mismatch ratios in the two *D. virilis* strains: The global alignments also reveal many base mismatches between the two strains of *D. virilis* (Table 3). A total of 88.6% of the mismatched bases (3141/3547) in the CAF1 dot chromosome have phred scores ≥30 (*i.e.*, with an estimated error rate of <1 in 1000 bases) (Figure S6). Therefore the majority of the mismatches are genuine differences between the two strains. Previous work used a non-LTR retrotransposon (*Helena*) to estimate the neutral mutation rates in *D. virilis*; an estimate of 0.16 deletions per substitution with a 95% confidence interval of 0.09–0.26 deletions per substitution was obtained (Petrov *et al.* 1996). An extension of this analysis using five different non-LTR retrotransposons that are "dead-on-arrival" generated a neutral mutation rate of 0.174 deletions per substitution (95% confidence interval: 0.133–0.244) (Blumenstiel *et al.* 2002). The two strains of *D. virilis* used here show an overall indel-to-mismatch ratio of 0.247 on the dot chromosome compared to 0.200 on the random set of fosmids (previously finished and annotated in Slawson *et al.* 2006) from the other chromosome arms (Table 3). We also found that indels and mismatches are nonuniformly distributed across the *D. virilis* dot chromosome,

TABLE 2

List of indels >1 kb in the two strains of *D. virilis*

Strain	Strain difference ID	Repeat length	Best hit to Superlibrary	Repeat class	LTR length	% terminal repeat identities
GEP	gep_contig9_155839	6811	dvir.2.37.centroid	LTR	412	100.0
GEP	gep_contig9_314579	6809	dvir.2.37.centroid	LTR	412	100.0
GEP	gep_contig4_259601	6437	dvir.35.83.centroid	LTR	427	100.0
GEP	gep_contig5_83074	5409	dvir.5.67.centroid	LTR	227	100.0
GEP	gep_contig5_186788	2539	dvir.16.2.centroid	DNA	NA	NA
GEP	gep_contig4_40674	2489	dvir.16.2.centroid	DNA	NA	NA
GEP	gep_contig5_139357	1704	dvir.35.83.centroid	LTR	NA	NA
GEP	gep_contig4_166960	1616	Helitron-1N1_DVir	DNA/Helitron	NA	NA
GEP	gep_contig5_136754	1422	dvir.35.83.centroid	LTR	NA	NA
GEP	gep_contig8_155919	1343	dvir.3.94.centroid	LTR	NA	NA
GEP	gep_contig5_135417	1329	dvir.35.83.centroid	LTR	NA	NA
GEP	gep_contig5_138182	1167	dvir.35.83.centroid	LTR	NA	NA
CAF1	caf_contig8_177230	10,606	Ulysses_I	LTR/Gypsy	2120	99.7
CAF1	caf_contig5_181062	6394	dvir.2.53.centroid	LTR	412	100.0
CAF1	caf_contig8_53382	2123	dvir.21.15.centroid	LTR	NA	NA

Analysis of indels >1 kb in the two strains (GEP and CAF1) of *D. virilis* shows that most of the large indels have sequence similarity to LTR retrotransposons and DNA transposons. All of the indels >3 kb are classified as LTR retroelements in the Superlibrary; five of the six terminal repeats that can be identified have perfect sequence identity.

TABLE 3

Indels to mismatch ratios on the dot chromosome and random fosmids from other chromosomes for three different types of sequences (overall, coding exon, and introns)

Type	Region	No. mismatches	No. indels	No. indels/no. mismatches
Overall	Dot chromosome	3547	877	0.247
	Random fosmids	2125	424	0.200
Coding exons	Dot chromosome	282	22	0.078
	Random fosmids	87	6	0.069
Introns	Dot chromosome	1335	289	0.216
	Random fosmids	419	53	0.126

The intronic regions are used as a proxy to estimate the neutral mutation rate on the dot chromosomes and the random fosmids. Higher ratios of indels to mismatches were observed on the dot chromosomes in all three types of sequences compared to the random fosmids, which suggests less effective selection on the dot chromosome compared to other euchromatic regions.

with higher numbers of indels and mismatches near the centromere (Figure S7).

The ratio of indels to substitutions will change depending on the region's functional constraints (Chen *et al.* 2009): Coding regions have lower indel-to-substitution ratios because purifying selection removes indels that lead to frameshifts (Taylor *et al.* 2004). To determine whether the same constraints exist in the *D. virilis* dot chromosome, we first identified the analogous coding regions in the CAF1 strain by mapping our manually curated coding exons from the GEP strain onto the CAF1 dot chromosome. A total of 593 out of 594 exons showed full-length alignment (see File S1 for analysis of exons that show partial or poor alignments) with an average percentage of sequence identity of 99.9% and a standard deviation of 0.3%. After filtering mismatches that were caused by errors in the alignment or in the consensus sequence, we found a higher indel-to-mismatch ratio on the coding exons of the dot chromosome (0.078) compared to the random set of fosmids (0.069) (Table 3). To estimate the neutral mutation rate in both regions, we also analyzed the indel-to-mismatch ratios within introns. The dot chromosome again shows a higher indel-to-mismatch ratio (0.216) compared to random set of fosmids (0.126) (Table 3).

Collectively, our analysis shows that the dot chromosome has an elevated indel-to-mismatch ratio compared to the random set of fosmids from the other *D. virilis* chromosomes. The higher indel-to-mismatch ratio may reflect less effective selection on the dot chromosome compared to other regions of the *D. virilis* genome.

K_a/K_s analysis shows weak purifying selection for most of the genes on the dot chromosome: We calculated the ratio of the number of substitutions per nonsynonymous site to the number of substitutions per synonymous site (K_a/K_s ratio) as a metric for the degree of functional constraint (*i.e.*, purifying or directional selection) (Hurst 2002). Among the 76 genes on the GEP strain that can be mapped unambiguously onto the CAF1 strain, 22 genes had no base mismatches within the coding region, 48 genes had fewer than 10 mismatches, and the remaining 6 genes had more than 10 mismatches. Among the 54 genes with at least 1 mismatch, 20 genes contained only synonymous substitutions and 12 genes contained only nonsynonymous changes. For the 22 genes that contained both synonymous and nonsynonymous changes, the median K_a/K_s ratio is 0.302 and the mean is 0.398. The K_a/K_s ratio ranged from a maximum of 1.094 for CG32016 (which suggests that the gene is under no selective constraint) to the minimum of 0.047 for CG11093 (which suggests that the gene is under purifying selection). Therefore, despite the unique environment of the dot chromosome, most of the genes on the *D. virilis* dot chromosome are undergoing purifying selection (see Table S14 for the K_a/K_s ratio of each gene).

To determine whether the K_a/K_s ratio on the dot chromosome is unusual, we also analyzed the genes on the set of random fosmids that we have previously annotated (Slawson *et al.* 2006). Of the 20 partial and complete genes found on these fosmids, 17 can be mapped unambiguously onto the CAF1 assembly. For 9 of these genes that contain both synonymous and nonsynonymous changes, the median K_a/K_s ratio is 0.137 and the mean is 0.250 (Table S15). The higher K_a/K_s ratio on the dot chromosome compared to this random set of fosmids suggests that selection is less effective on the dot chromosome than in the euchromatic regions of the genome.

DISCUSSION

In this study, we have generated a high-quality *D. virilis* dot chromosome sequence and manually curated gene models to examine the characteristics and evolution of the dot chromosome in different Drosophila species (*D. melanogaster* and *D. virilis*) and in different *D. virilis* strains (GEP and CAF1). Our analysis consistently shows that the *D. melanogaster* and *D. virilis* dot chromosomes are more similar to each other than to the reference heterochromatic and euchromatic regions of both species.

Dot chromosomes have distinct distribution of repetitive elements: The total repeat densities of the *D. melanogaster* and *D. virilis* dot chromosomes are similar, intermediate between that of euchromatin and heterochromatin (Figure 1). The main difference is in the types of repeats present, with the *D. melanogaster* dot enriched in DNA transposons and retroelements (Figure 2, Table S2). Previous studies using position effect variegation (PEV) in *D. melanogaster* as a readout of chromatin packaging have characterized the dot chromosome as largely heterochromatic and have also suggested that both proximity to certain transposable elements (*e.g.*, the DNA transposon *1360*) and overall repeat density may both play a role in heterochromatin formation and maintenance (Sun *et al.* 2004; Haynes *et al.* 2006; Riddle *et al.* 2008). If, as has been suggested, transposable elements are a better target for heterochromatin formation in Drosophila (Huisinga *et al.* 2006), differences in the distribution of classes of repeats may alter effective heterochromatin formation on the two dot chromosomes under some circumstances. This difference in repeat type, one of the few observed, might explain the difference reported earlier in polytene chromosome immunofluorescent staining, where the *D. virilis* dot chromosome fails to show the prominent association with HP1a and H3K9me2/3 seen in *D. melanogaster* (James *et al.* 1989; Slawson *et al.* 2006). However, genes on the *D. virilis* and *D. melanogaster* dot chromosome have similar characteristics (*e.g.*, large size, low codon bias), which argues that these genes have evolved in a similar heterochromatin-like domain in both species and must be similarly packaged in germ line cells.

Our analysis also shows a higher abundance of tandem repeats in the *D. virilis* dot chromosome, as well as euchromatin and heterochromatin (Figure 3). The majority of the tandem repeats identified here by TRF overlap with simple and low complexity repeats identified by RepeatMasker, in agreement with previous findings of Schlotterer and Harr (2000). The expansion of these types of low complexity sequences both on the dot chromosome and in the euchromatic and heterochromatic reference regions may have contributed to the larger euchromatic genome size in *D. virilis* compared to *D. melanogaster* (150 Mb *vs.* 110 Mb; Moriyama *et al.* 1998), albeit recognizable tandem repeats only account for a small percentage of the two genomes.

Gene characteristics reflect the low levels of recombination: Another well-established property of heterochromatic domains is a lack of recombination (Grewal and Elgin 2002). Previous reports have shown low levels of recombination on both the *D. melanogaster* and *D. virilis* dot chromosomes (Chino and Kikkawa 1933; Bridges 1935; Wang *et al.* 2002; Arguello *et al.* 2010). Work by others has found that both very short and very long introns are associated with regions of low recombination (Carvalho and Clark 1999; Comeron and Kreitman 2000). An earlier study by Haddrill *et al.* (2007) also found that a lack of recombination could be correlated with an increase in gene length. Therefore, if both the *D. melanogaster* and *D. virilis* dot chromosomes have low levels of recombination, they should have similar distributions of intron sizes, as we have observed. The higher density of repetitive elements on the dot chromosomes contributes to the larger gene and intron sizes on the dot chromosomes compared to the euchromatic reference regions (Figure 4, A and C). However, recognizable repetitive elements within introns are not the sole factor leading to larger gene and intron sizes, because the differences in intron sizes between the dot chromosomes and the euchromatic reference regions remain statistically significant even after recognizable repeats are removed (Figure 4D).

Codon usage bias has previously been shown to be negatively correlated with protein length and positively correlated with levels of recombination (Powell and Moriyama 1997; Haddrill *et al.* 2007). The positive correlation between codon bias and recombination rate can be attributed to the Hill–Robertson effect (*i.e.*, regions with a low rate of recombination show less effective response to selection) (Hill and Robertson 1966). Selection may be at work in the positive correlation of codon bias with gene expression levels, observed generally in *D. melanogaster* (Duret and Mouchiroud 1999). Differences in expression levels are unlikely to be a major contributor to differences in codon bias between the dot chromosomes and the euchromatic reference regions, since Betancourt and colleagues (using expression data generated by Zhang *et al.* 2007) have previously shown that the difference in gene expression levels for the dot and non-dot loci are not statistically significant in *D. virilis* (Betancourt *et al.* 2009).

Our codon bias results (Figure 5) are consistent with the low rate of recombination reported for the heterochromatin-like *D. melanogaster* dot chromosome and suggest a similar evolution of the *D. virilis* dot chromosome. Our findings further suggest that the rate of recombination may be a more important determinant of codon usage bias on the dot chromosomes than protein length or level of expression. The codon bias in the *D. melanogaster* heterochromatic reference region is higher (lower Nc value) than the codon bias in the *D. virilis* dot chromosome, even though the pericentric heterochromatin has a significantly higher repeat density and is thought to have a similar low level of recombination.

Gene order and orientation indicate a high rate of inversions in a domain with low recombination: While the recombination rate is significantly lower, we observe an approximately twofold higher rate of gene rearrangements on the dot chromosome compared to a euchromatic domain (Figure 6); this higher rate may reflect the higher repeat density, assuming that these elements promote inversions (Casals and Navarro 2007). Bhutkar *et al.* (2008) have previously observed a higher rate of gene rearrangements on the Muller *A* element

(X chromosome) compared to Muller elements *B–E* and suggested that the rate of gene rearrangement may play a role in the evolution of the X chromosome. The higher rate of inversions on the dot chromosome compared to the Muller *D* element suggests that, similar to the Muller *A* element, gene rearrangements may play an important role in the evolution of the dot chromosome.

Wanderer genes exhibit the properties common in the domain they inhabit: Despite the large number of gene rearrangements, ~90% of the genes (74/83) can be found on both the *D. melanogaster* and *D. virilis* dot chromosomes. Our results are consistent with the previous findings by Bhutkar *et al.* (2008) who estimated that 95% of the genes in Drosophila are localized to the same Muller element across different Drosophila species. We identified eight wanderer genes that are present in a euchromatic domain in one species and heterochromatic (dot chromosome) in the other (Figure 7). These genes exhibit the characteristics of other genes in the same environment (Table 1), which suggests that characteristics such as gene size, codon bias, and repeat density are properties of the domain, and are not required for either set of genes to function *per se*. Our results are consistent with a previous study of the *lt* gene cluster in different Drosophila species, which shows that genes that transition from a euchromatic domain to a heterochromatic domain will reflect the properties of their local environment (*i.e.*, increase in gene size due to accumulation of transposable elements in the heterochromatic domain) (Yasuhara *et al.* 2005). The movement of genes from one chromosome to another is widely observed in Drosophila (Drosophila 12 Genomes Consortium *et al.* 2007), but the mechanism remains obscure; these events do not appear to be the consequence of recombination or of retroviral action through a cDNA.

Strain differences indicate that indels contribute significantly to change: Comparison of the GEP and CAF1 strains of *D. virilis* shows a large number of differences (*e.g.*, base mismatches, insertions, and deletions). These differences include a few large indels of transposable elements (primarily LTR and DNA transposons; Table 2), although the majority of the indels are short (Figure 8). We found that most of the large indels with conserved long terminal repeats can be classified as members of the gypsy family, which suggests a recent invasion of gypsy elements into the genomes of *D. virilis*. Our results are consistent with previous reports that show gypsy retroelements to be actively transcribed in *D. virilis* and are also consistent with reports that show variation in the distribution of gypsy elements in different strains of *D. virilis* (Mizrokhi and Mazo 1991; Mejlumian *et al.* 2002).

Previous studies have suggested that indels play an important role in the evolution of eukaryotic genomes, and have postulated that indels account for the majority of the sequence differences in closely related DNA samples (Britten *et al.* 2003). Our analysis shows that the total number of mismatches exceeds the number of indel events in this case (Table 3). However, as each indel on average introduces a difference of two or more bases (including large transposon insertions and deletions), these types of events contribute more to the difference between the dot chromosomes (a total of 81,715 indel bases for the two strains combined), as postulated (Britten *et al.* 2003).

Strain differences point to weak purifying selection in the dot chromosome domain: Previous studies on polymorphisms within the coding regions of *D. melanogaster* have shown much lower levels of both nonsynonymous and synonymous changes on the dot chromosome compared to the genome average (Berry *et al.* 1991; Sheldahl *et al.* 2003). Analysis of the K_a/K_s ratio observed here suggests that most of the genes on the *D. virilis* dot chromosome are undergoing weak purifying selection compared to the genes on a random set of fosmids from other, euchromatic regions. Analysis of *D. americana*, a species closely related to *D. virilis*, also suggested weak purifying positive selection on the dot chromosome, presumably a consequence of its lower level of recombination (Betancourt *et al.* 2009).

Future studies: The dot chromosome is unusual compared to other gene-rich (euchromatic) regions of the Drosophila genome because of the high density of repetitive elements. However, in this regard the dot chromosome actually resembles a typical euchromatic region of a mammalian genome, where one observes repeat densities of ~30–40%, with remnants of transposable elements interspersed within and between genes that are actively transcribed. How is gene activity maintained in the midst of repetitious elements, elements that are thought to serve as targets for heterochromatin formation and gene silencing? Future investigation should examine the transcription start sites of the dot chromosome genes through a comprehensive study of the distribution of histone modifications and chromosomal proteins surrounding these regions. In conjunction with the publicly available data released by the modENCODE Project for *D. melanogaster* [N. C. Riddle, A. Minoda, P. V. Kharchenko, A. A. Alekseyenko, Y. B. Schwartz, M. Y. Tolstorukov, A. A. Gorchakov, C. Kennedy, D. Linder-Basso, J. D. Jaffe, G. Shanower, M. I. Kuroda, V. Pirrotta, P. J. Park, S. C. R. Elgin, G. H. Karpen, and the modENCODE Consortium (http://www.modencode.org), unpublished results]. a comparative study with mapping data from multiple Drosophila species may reveal common sequence motifs that regulate gene expression and chromatin packaging in this genomic environment. The unique properties of the dot chromosome provide an opportunity to examine the impact of chromatin packaging on the evolution of genomes and the control of gene expression, making it worthy of further study.

We thank Casey Bergman, Andrew Clark, Kenneth Olsen, and two anonymous referees for valuable comments and suggestions on this manuscript. Members of the Elgin lab contributed throughout the process with criticisms and suggestions. We thank the Washington University Genome Center for generating raw sequences and providing training and support for many of the coauthors. This work was supported by grant no. 52005780 from the Howard Hughes Medical Institute (HHMI) to Washington University (to S.C.R.E.) with additional funding for data analysis from National Institutes of Health (NIH) grant R01 GM068388 (to S.C.R.E.). The content of this article is solely the responsibility of the authors and does not necessarily represent the official views of HHMI, the National Institute of General Medical Sciences, or the NIH.

LITERATURE CITED

Arguello, J. R., Y. Zhang, T. Kado, C. Fan, R. Zhao *et al.*, 2010 Recombination yet inefficient selection along the *Drosophila melanogaster* subgroup's fourth chromosome. Mol. Biol. Evol. **27**(4): 848–861.

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

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

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

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

Berry, A., J. Ajioka and M. Kreitman, 1991 Lack of polymorphism on the *Drosophila* fourth chromosome resulting from selection. Genetics **129:** 1111–1117.

Betancourt, A. J., J. J. Welch and B. Charlesworth, 2009 Reduced effectiveness of selection caused by a lack of recombination. Curr. Biol. **19:** 655–660.

Bhutkar, A., S. W. Schaeffer, S. M. Russo, M. Xu, T. F. Smith *et al.*, 2008 Chromosomal rearrangement inferred from comparisons of 12 Drosophila genomes. Genetics **179:** 1657.

Blumenstiel, J. P., D. L. Hartl and E. R. Lozovsky, 2002 Patterns of insertion and deletion in contrasting chromatin domains. Mol. Biol. Evol. **19:** 2211–2225.

Bridges, C. B., 1935 The mutants and linkage data of chromosome four of *Drosophila melanogaster*. Biol. Zh. (Moscow) **4:** 401–420.

Britten, R. J., L. Rowen, J. Williams and R. A. Cameron, 2003 Majority of divergence between closely related DNA samples is due to indels. Proc Natl. Acad. Sci. USA **100:** 4661–4665.

Burge, C., and S. Karlin, 1997 Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. **268:** 78–94.

Carvalho, A. B., and A. G. Clark, 1999 Intron size and natural selection. Nature **401:** 344.

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

Chen, J., Y. Wu, H. Yang, J. Bergelson, M. Kreitman *et al.*, 2009 Variation in the ratio of nucleotide substitution and indel rates across genomes in mammals and bacteria. Mol. Biol. Evol. **26:** 1523–1531.

Chino, M., and H. Kikkawa, 1933 Mutants and crossing over in the dot-like chromosome of *Drosophila virilis*. Genetics **18:** 111–116.

Comeron, J. M., and M. Kreitman, 2000 The correlation between intron length and recombination in Drosophila dynamic equilibrium between mutational and selective forces. Genetics **156:** 1175–1190.

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

Drosophila 12 Genomes Consortium, A. G. Clark, M. B. Eisen, D. R. Smith, C. M. Bergman *et al.*, 2007 Evolution of genes and genomes on the Drosophila phylogeny. Nature **450:** 203–218.

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

Edgar, R. C., and E. W. Myers, 2005 PILER: identification and classification of genomic repeats. Bioinformatics **21:** 152–158.

Elsik, C. G., A. J. Mackey, J. T. Reese, N. V. Milshina, D. S. Roos *et al.*, 2007 Creating a honey bee consensus gene set. Genome Biol. **8:** R13.

Grewal, S. I. S., and S. C. R. Elgin, 2002 Heterochromatin: new possibilities for the inheritance of structure. Curr. Opin. Genet. Dev. **12:** 178–187.

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

Haynes, K. A., A. A. Caudy, L. Collins and S. C. R. Elgin, 2006 Element 1360 and RNAi components contribute to HP1-dependent silencing of a pericentric reporter. Curr. Biol. **16:** 2222–2227.

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

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

Huisinga, K. L., B. Brower-Toland and S. C. R. Elgin, 2006 The contradictory definitions of heterochromatin: transcription and silencing. Chromosoma **115:** 110–122.

Hurst, L. D., 2002 The Ka/Ks ratio: diagnosing the form of sequence evolution. Trends Genet. **18:** 486–487.

James, T. C., J. C. Eissenberg, C. Craig, V. Dietrich, A. Hobson *et al.*, 1989 Distribution patterns of HP1, a heterochromatin-associated nonhistone chromosomal protein of *Drosophila*. Eur. J. Cell Biol. **50:** 170–180.

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

Kent, W. J., C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle *et al.*, 2002 The human genome browser at UCSC. Genome Res. **12:** 996–1006.

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

Korf, I., P. Flicek, D. Duan and M. R. Brent, 2001 Integrating genomic homology into gene structure prediction. Bioinformatics **17:** 140–148.

Lamb, J. C., N. C. Riddle, Y. Cheng, J. Theuri and J. A. Birchler, 2007 Localization and transcription of a retrotransposon-derived element on the maize B chromosome. Chromosome Res. **15:** 383–398.

Larkin, M. A., G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan *et al.*, 2007 Clustal W and clustal X version 2.0. Bioinformatics **23:** 2947–2948.

Lewis, S., S. Searle, N. Harris, M. Gibson, V. Lyer *et al*, 2002 Apollo: a sequence annotation editor. Genome Biol. **3:** 0082–0081.

Marais, G., and B. Charlesworth, 2003 Genome evolution: recombination speeds up adaptive evolution. Curr. Biol. **13:** R68–R70.

Mejlumian, L., A. Pelisson, A. Bucheton and C. Terzian, 2002 Comparative and functional studies of Drosophila species invasion by the *gypsy* endogenous retrovirus. Genetics **160:** 201–209.

Mizrokhi, L. J., and A. M. Mazo, 1991 Cloning and analysis of the mobile element *gypsy* from *D. virilis*. Nucleic Acids Res. **19:** 913–916.

Moriyama, E., D. Petrov and D. Hartl, 1998 Genome size and intron size in *Drosophila*. Mol. Biol. Evol. **15:** 770.

Parra, G., E. Blanco and R. Guigó, 2000 Geneid in *Drosophila*. Genome Res. **10:** 511–515.

Parra, G., P. Agarwal, J. F. Abril, T. Wiehe, J. W. Fickett *et al.*, 2003 Comparative gene prediction in human and mouse. Genome Res. **13:** 108–117.

Pertea, M., X. Lin and S. L. Salzberg, 2001 GeneSplicer: a new computational method for splice site prediction. Nucleic Acids Res. **29:** 1185.

Petrov, D., E. Lozovskaya and D. Hartl, 1996 High intrinsic rate of DNA loss in *Drosophila.* Nature **384:** 346.

Powell, J. R., and R. DeSalle, 1995 Drosophila molecular phylogenies and their uses. Evol. Biol. **28:** 87–138.

Powell, J. R., and E. N. Moriyama, 1997 Evolution of codon usage bias in *Drosophila.* Proc. Natl. Acad. Sci. USA **94:** 7784–7790.

Price, A. L., N. C. Jones and P. A. Pevzner, 2005 De novo identification of repeat families in large genomes. Bioinformatics **21:** 351–358.

Rice, P., I. Longden, A. Bleasby, and others, 2000 EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. **16:** 276–277.

Riddle, N. C., and S. C. R. Elgin, 2006 The dot chromosome of Drosophila: insights into chromatin states and their change over evolutionary time. Chromosome Res. **14:** 405–416.

Riddle, N. C., W. Leung, K. A. Haynes, H. Granok, J. Wuller *et al.*, 2008 An investigation of heterochromatin domains on the fourth chromosome of *Drosophila melanogaster.* Genetics **178:** 1177.

SanMiguel, P., B. S. Gaut, A. Tikhonov, Y. Nakajima and J. L. Bennetzen, 1998 The paleontology of intergene retrotransposons of maize. Nat. Genet. **20:** 43–45.

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.

Schlotterer, C., and B. Harr, 2000 *Drosophila virilis* has long and highly polymorphic microsatellites. Mol. Biol. Evol. **17:** 1641–1646.

Sheldahl, L. A., D. M. Weinreich and D. M. Rand, 2003 Recombination, dominance and selection on amino acid polymorphism in the *Drosophila* genome contrasting patterns on the X and fourth chromosomes. Genetics **165:** 1195–1208.

Singh, N. D., J. C. Davis and D. A. Petrov, 2005 X-linked genes evolve higher codon bias in Drosophila and Caenorhabditis. Genetics **171:** 145–155.

Slawson, E. E., C. D. Shaffer, C. D. Malone, W. Leung, E. Kellmann *et al.*, 2006 Comparison of dot chromosome sequences from *D. melanogaster* and *D. virilis* reveals an enrichment of DNA transposon sequences in heterochromatic domains. Genome Biol. **7:** R15.

Smith, C. D., R. C. Edgar, M. D. Yandell, D. R. Smith, S. E. Celniker *et al.*, 2007 Improved repeat identification and masking in dipterans. Gene **389:** 1–9.

Stajich, J. E., D. Block, K. Boulez, S. E. Brenner, S. A. Chervitz *et al.*, 2002 The BioPerl toolkit: Perl modules for the life sciences. Genome Res. **12:** 1611–1618.

Sun, F., K. Haynes, C. L. Simpson, S. D. Lee, L. Collins *et al.*, 2004 Cis-acting determinants of heterochromatin formation on *Drosophila melanogaster* chromosome four. Mol. Cell. Biol. **24:** 8210–8220.

Taylor, M. S., C. P. Ponting and R. R. Copley, 2004 Occurrence and consequences of coding sequence insertions and deletions in mammalian genomes. Genome Res. **14:** 555–566.

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

Wang, W., K. Thornton, A. Berry and M. Long, 2002 Nucleotide variation along the *Drosophila melanogaster* fourth chromosome. Science **295:** 134–137.

Waterhouse, A. M., J. B. Procter, D. M. A. Martin, M. Clamp and G. J. Barton, 2009 Jalview version 2: a multiple sequence alignment editor and analysis workbench. Bioinformatics **25:** 1189–1191.

Wright, F., 1990 The 'effective number of codons' used in a gene. Gene **87:** 23–29.

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

Yasuhara, J. C., C. H. DeCrease and B. T. Wakimoto, 2005 Evolution of heterochromatic genes of Drosophila. Proc.Natl. Acad. Sci. USA **102:** 10958–10963.

Zhang, Y., D. Sturgill, M. Parisi, S. Kumar and B. Oliver, 2007 Constraint and turnover in sex-biased gene expression in the genus Drosophila. Nature **450:** 233–237.

Communicating editor: N. Perrimon

GENETICS

Supporting Information
http://www.genetics.org/cgi/content/full/genetics.110.116129/DC1

Evolution of a Distinct Genomic Domain in Drosophila: Comparative Analysis of the Dot Chromosome in *Drosophila melanogaster* and *Drosophila virilis*

Wilson Leung, Christopher D. Shaffer, Taylor Cordonnier, Jeannette Wong, Michelle S. Itano, Elizabeth E. Slawson Tempel, Elmer Kellmann, David Michael Desruisseau, Carolyn Cain, Robert Carrasquillo, Tien M. Chusak, Katazyna Falkowska, Kelli D. Grim, Rui Guan, Jacquelyn Honeybourne, Sana Khan, Louis Lo, Rebecca McGaha, Jevon Plunkett, Justin M. Richner, Ryan Richt, Leah Sabin, Anita Shah, Anushree Sharma, Sonal Singhal, Fine Song, Christopher Swope, Craig B. Wilen, Jeremy Buhler, Elaine R. Mardis and Sarah C. R. Elgin

DOI: 10.1534/genetics.110.116129

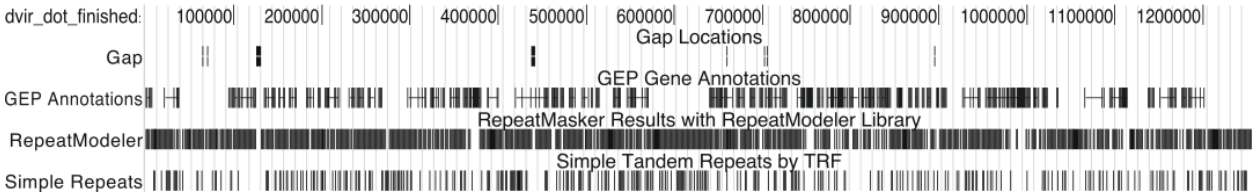


FIGURE S1.—Improved *D. virilis* (Tucson strain 15010-1051.88) dot chromosome on the custom *UCSC Genome Browser* (http://gander.wustl.edu; *D. virilis* assembly). The improved dot chromosome consists of 1,240,624 non-overlapping base pairs of high quality sequence with 8 remaining gaps (estimated total gap size: 14,728 bases). The *D. virilis* custom genome browser contains evidence tracks for gene annotations, recognizable repeats, and UCSC Net alignments referenced to the *D. melanogaster* genome assembly.

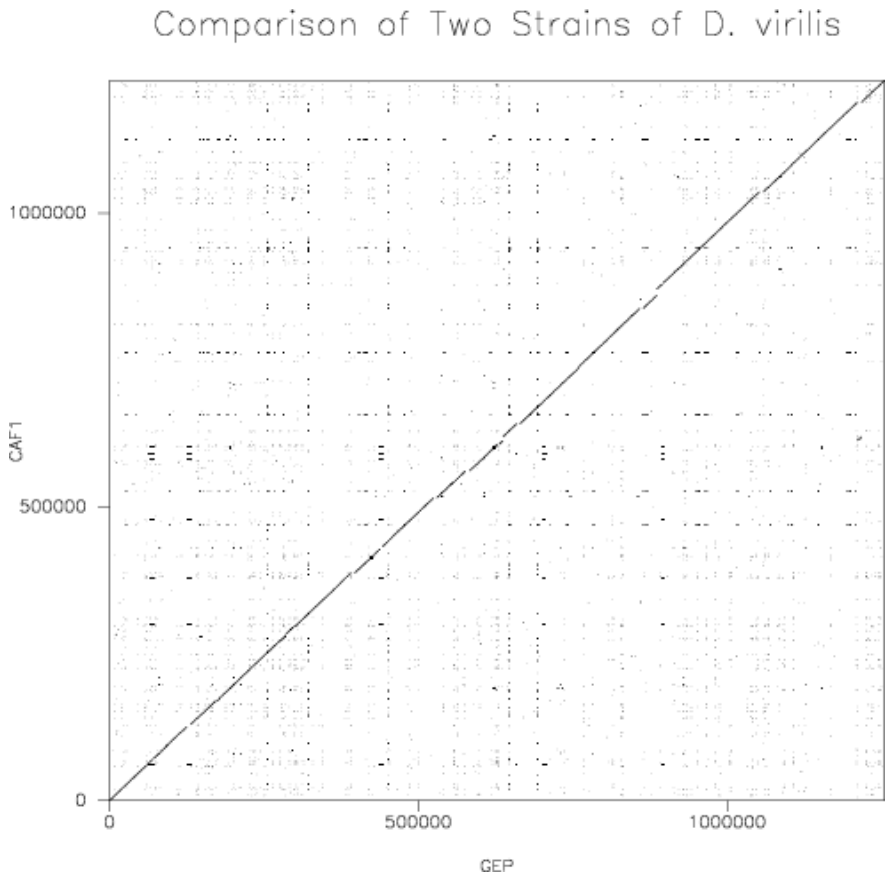
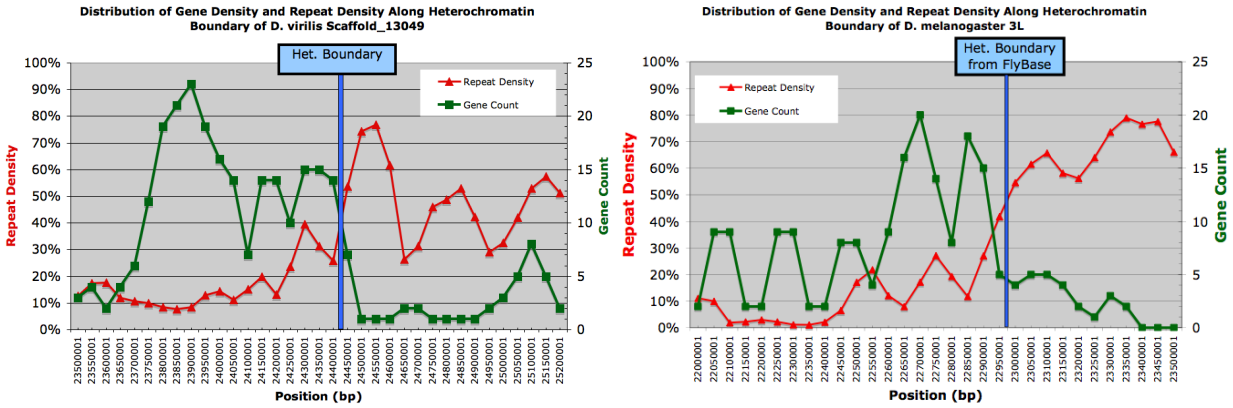


FIGURE S2.—Dot plot comparison of the dot chromosomes from two strains of *D. virilis*. A dot plot alignment comparing the dot chromosome of the GEP strain (Tucson strain 15010-1051.88) (X-axis) with the dot chromosome of the CAF1 strain (Tucson strain 15010-1051.87; 1,226,203 bases with 7 gaps estimated at 1,638 bases) (Y-axis) shows that the two *D. virilis* dot chromosomes have a high degree of sequence identity. Regions of sequence identity are shown by a dot; thus the diagonal lines in the dot plot indicate that the two chromosomes are essentially collinear.

A.



B.

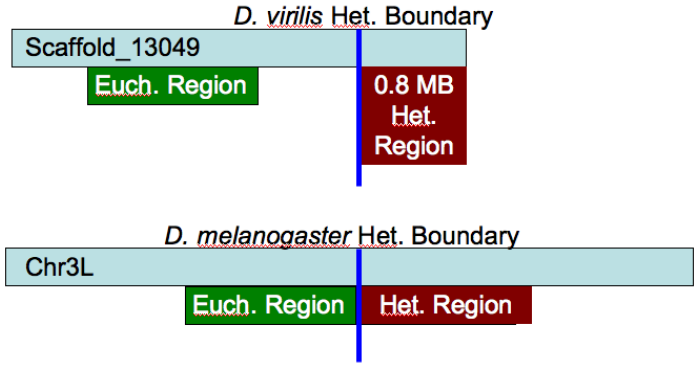


FIGURE S3.—Defining the heterochromatic and euchromatic reference regions for *D. melanogaster* and *D. virilis*. (A) The heterochromatic-euchromatic boundary for *D. virilis* scaffold_13049 has been assigned for this analysis based on changes in repeat density and gene density. (B) Schematic of reference regions used in the sequence analysis. For *D. virilis*, the euchromatic reference region from scaffold_13049 spans from 22,575,001 to 23,825,000 bp and the heterochromatic reference region from scaffold_13049 spans from 24,450,001 to 25,233,164 bp. For *D. melanogaster*, the euchromatic reference region from chr3L spans from 21,705,576 to 22,955,575 bp and the heterochromatic reference region spans from 22,955,576 to 24,205,575 bp.

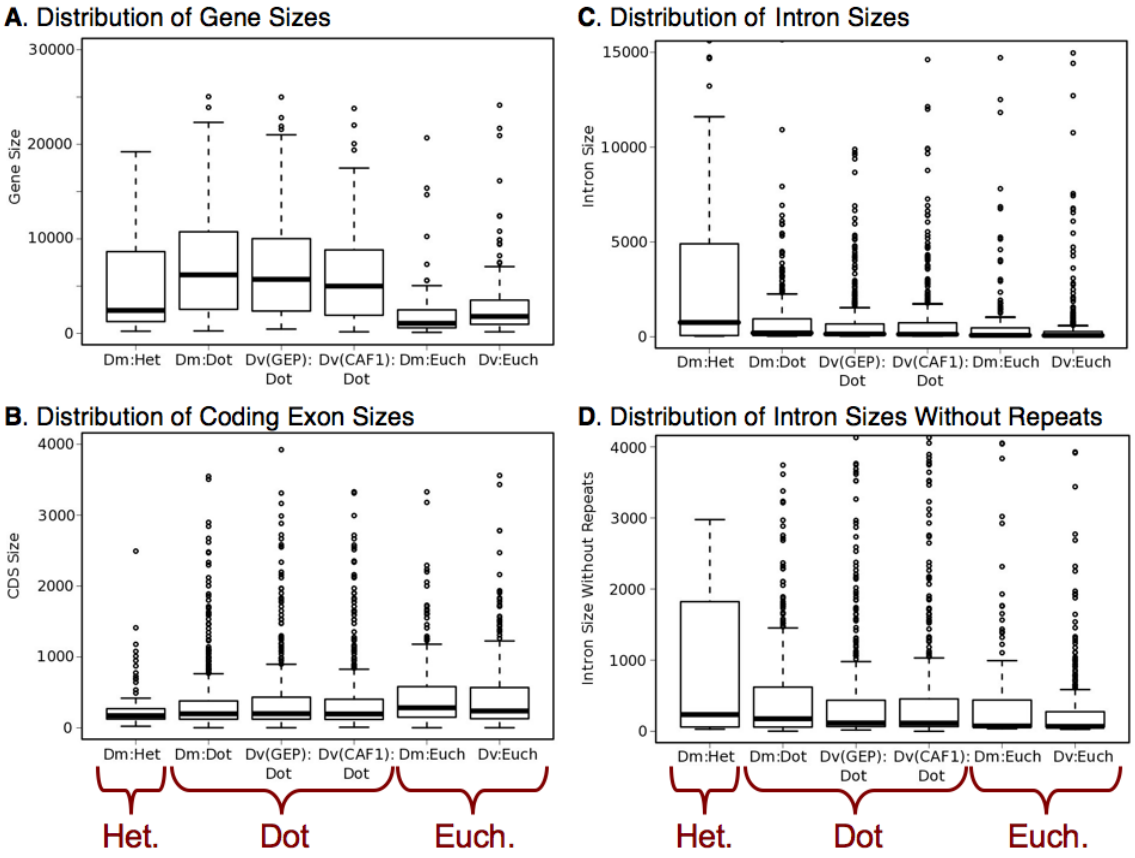


FIGURE S4.—Side-by-side boxplots of size distributions for genes, exons, introns, and introns without repeats. Results for various features on the *D. melanogaster* and *D. virilis* dot chromosomes as well as the euchromatic and heterochromatic reference regions are shown. The box in each boxplot represents the interquartile range (IQR) and is the difference between the third (Q3) and first (Q1) quartile (IQR = Q3-Q1). The line within each box in the boxplot represents the median. Outliers are defined as sizes that are more than 1.5*IQR below Q1 or 1.5*IQR larger than Q3 and are represented as dots in the boxplot. The smallest and largest values that are not outliers are shown as "whiskers" in the boxplots. (A) Side-by-side boxplots of gene sizes (from the start codon to the stop codon) show that genes on the dot chromosomes are larger than genes from the euchromatic reference regions. (B) Side-by-side boxplots show that coding exons on the dot chromosome tend to be larger than coding exons on the heterochromatic reference region and smaller than the coding exons on the euchromatic reference regions. (C) Side-by-side boxplots of intron sizes show that introns on the dot chromosome are smaller than the introns in the heterochromatic reference region and larger than introns in the euchromatic reference region. (D) Side-by-side boxplots of intron sizes after the repeats are removed show a decrease in intron size differences among the dot chromosomes and the euchromatic reference regions, suggesting that repeats within introns are one of the factors that lead to larger intron sizes on the dot chromosomes.

```
BLASTN 2.0MP-WashU [04-May-2006] [macosx-10.4-x64-I32LPF64 2006-09-21T14:14:58]

Copyright (C) 1996-2006 Washington University, Saint Louis, Missouri USA.
All Rights Reserved.

Reference:  Gish, W. (1996-2006) http://blast.wustl.edu

Notice:  this program and its default parameter settings are optimized to find
nearly identical sequences rapidly.  To identify weak protein similarities
encoded in nucleic acid, use BLASTX, TBLASTN or TBLASTX.

Query=  caf1_scaffold_13052_contig8_177230-187835
        (10,606 letters)

Database:  superlibrary_v2.lib
           1320 sequences; 2,075,090 total letters.
Searching....10....20....30....40....50....60....70....80....90....100% done
```

Sequences producing High-scoring Segment Pairs:	High Score	Smallest Sum Probability P(N)	N
Ulysses_I#LTR/Gypsy RepbaseID: ULYSSES_I	27234	0.	1
dvir.21.15.centroid#LTR	10283	0.	1
Ulysses_LTR#LTR/Gypsy RepbaseID: ULYSSES_LTR	5911	2.2e-263	1
OSVALDO_I#LTR/Gypsy RepbaseID: OSVALDO_I	1015	7.2e-36	2
dvir.23.26.centroid#LTR	515	7.8e-18	1
TIRANT_I#LTR/Gypsy RepbaseID: TIRANT_I	572	1.5e-17	1
dmel.6.25.centroid#LTR	572	1.6e-17	1
dmel.6.30.centroid#LTR	572	1.6e-17	1

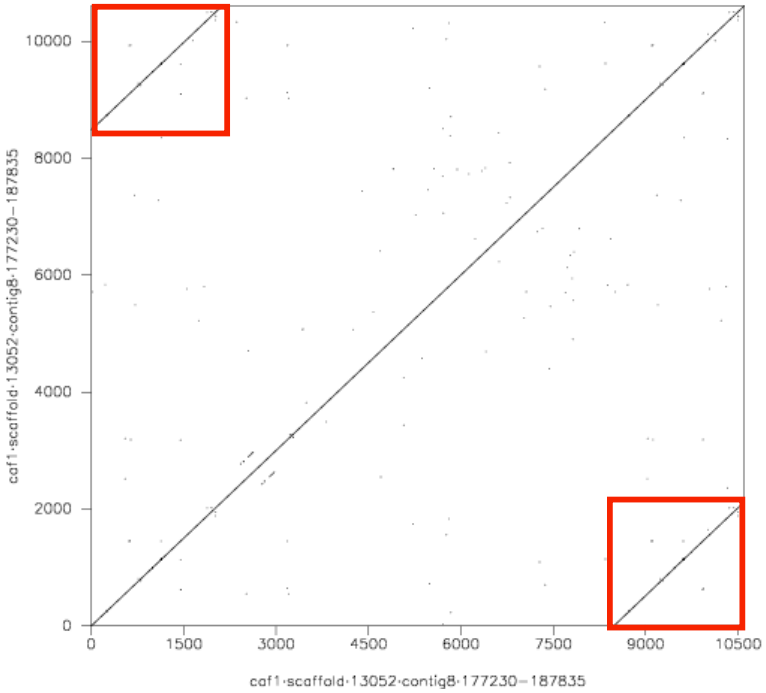


FIGURE S5.—Large indels occur in the dot chromosome from two strains of *D. virilis*. Example of a large indel (a Ulysses_I LTR retrotransposon) identified in the GEP strain of the *D. virilis* dot chromosome when compared to the CAF1 strain. Dot plot of this 10 kb LTR retroelement aligned against itself shows the locations of the 2kb long terminal repeat (highlighted by the red squares).

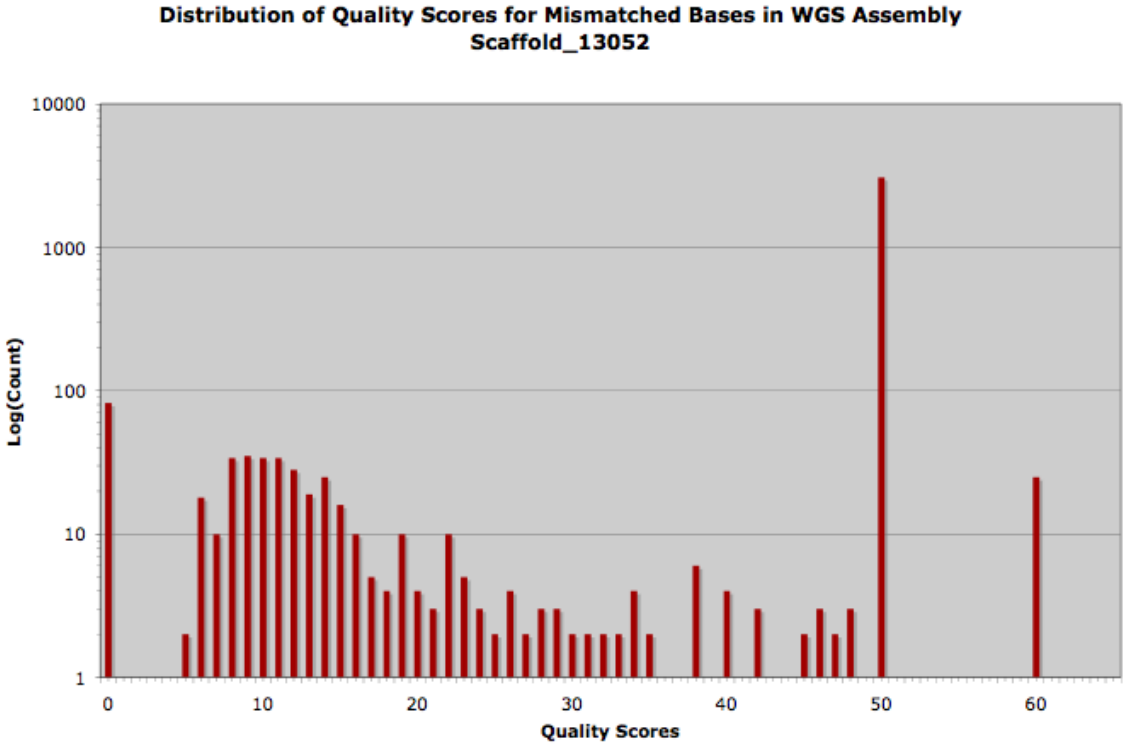


FIGURE S6.—Consensus quality scores for mismatched bases in the CAF1 strain of the *D. virilis* dot chromosome. Distribution of the consensus quality scores for the mismatched bases in scaffold_13052 of the CAF1 assembly shows that the majority of the mismatched bases are of high quality. 88.6% have a *phred* quality score of 30 or higher, indicating that consensus error is unlikely to be a major cause of mismatches observed between the two strains of *D. virilis*.

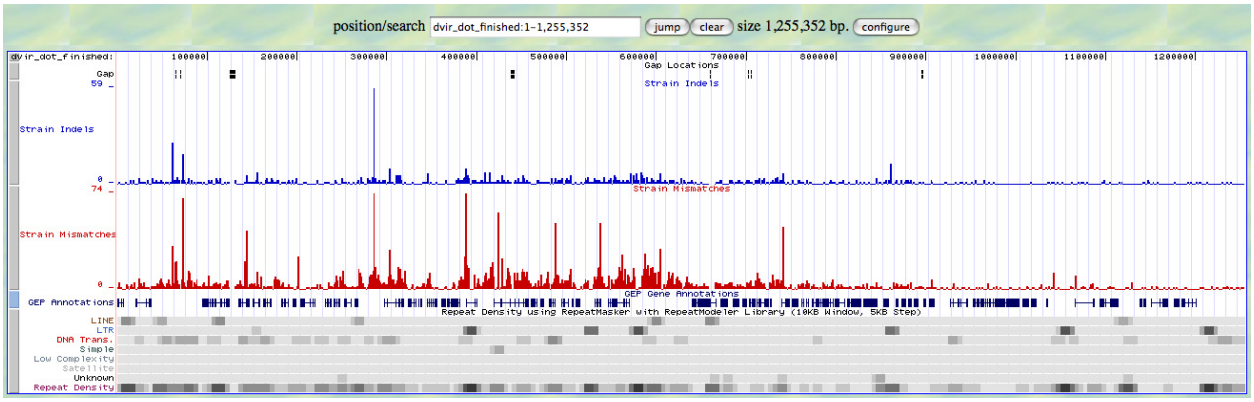


FIGURE S7.—Non-uniform distribution of indels and mismatches across the *D. virilis* dot chromosome. Analysis of indels and mismatches (per 10 kb window) shows that regions that are closer to the telomere (right) have a lower frequency of indels and mismatches compared to regions that are closer to the centromere (left). This difference may be attributed to the lower rate of recombination near the centromere.

FILE S1

Additional Materials, Methods, and References

***D. virilis* finishing quality standards:**
The fosmid assemblies were generated using the *phred/phrap/consed* package (EWING *et al.* 1998; GORDON *et al.* 1998). All fosmids were improved to the quality standards used for the mouse genome: single stranded regions have a minimum *phred* score of 30 (1 error in 1000 bases) and double stranded regions have a minimum *phred* score of 25 (http://genome.wustl.edu/platforms/sequence_improvement/mouse_finishing_rules). Regions with high quality discrepancies were resolved by manual inspection and manipulation of the assembly. The integrity of each fosmid assembly was verified by comparing the *in silico* restriction digests of the assembly with real restriction digests of the fosmid generated using four different enzymes, with at least two matches required. Additional PCR reactions were used to close gaps not covered by the fosmid library. These PCR-only regions were also finished to the mouse genome standard.

Use of GLEAN-R models:
In addition to providing the genomic assembly, the 12 Genomes Consortium also released a set of GLEAN-R gene predictions for the *D. virilis* CAF1 assembly (DROSOPHILA 12 GENOMES CONSORTIUM *et al.* 2007; ELSIK *et al.* 2007). To determine if the GLEAN-R gene models are adequate for this type of genome-level analysis, we compared the GEP gene models with the corresponding GLEAN-R models in the CAF1 assembly of the dot chromosome. Based on the FlyBase ortholog assignment, 69 of the 81 genes annotated in the GEP strain have corresponding GLEAN-R models. Comparison of each GEP model with the corresponding GLEAN-R model shows that they have a mean percent identity of 90% and a median percent identity of 98%, which suggests that the GLEAN-R models on the dot chromosome were mostly congruent with our manually annotated models (data not shown). Hence the GLEAN-R gene predictions from the dot chromosome and the euchromatic reference regions in the CAF1 strain of *D. virilis* were included in the analyses reported here.

Curation of three novel genes found on the D. virilis dot chromosome:
A. *CG16719-alpha*
CG16719 is a gene nested within *CG6767* in *D. melanogaster* 3L (Muller D element) It contains a conserved domain DUF1042. There is only one other annotated gene in the *D. melanogaster* genome (*CG12395*) that also contains this conserved domain and its putative ortholog can be mapped to scaffold_13042 (Muller A element) in *D. virilis*. The putative ortholog for *CG16719* is mapped to scaffold_13049 (Muller D element) in the *D. virilis* CAF1 assembly, but also has significant alignment with the *D. virilis* dot chromosome. Given that this gene on the *D. virilis* dot chromosome aligns significantly better to *CG16719* than *CG12395*, we classified this gene on the *D. virilis* dot chromosome as a putative paralog of *CG16719*.

B. *eIF-5A-beta*
There are two regions (scaffold_13049 and scaffold_13052) in the *D. virilis* CAF1 assembly that shows significant homology to the *D. melanogaster eIF-5A* and both regions contained the conserved IF5A domain. Limited EST evidence from *D. virilis* suggests that both copies of the genes are expressed. However, the gene on the *D. virilis* dot chromosome (scaffold_13052) has weaker sequence homology to the *D. melanogaster eIF-5A*. Hence we conclude that the gene on the *D. virilis* dot is likely a paralog of the *D. melanogaster eIF-5A*.

C. *GEP001*
The novel *GEP001* gene contains a conserved Deme6 domain (pfam10300); conserved predicted orthologs are present in *D. grimshawi*, *D. willistoni*, *Anopheles gambiae* (ref|XP_311530.4) and *Culex quinquefasciatus* (ref|XP_001851338.1) but a presumptive ortholog could not be found by *TBLASTN* searches against the *D. melanogaster* genome assembly (data not shown; see browser at http://gander.wustl.edu [*D. virilis* Manuscript assembly]).

Genes on the D. melanogaster dot chromosome that cannot be definitively mapped onto the D. virilis CAF1 assembly:
A. *JYalpha*
The gene model for *JYalpha* is incomplete on the *D. melanogaster* dot chromosome. A more comprehensive gene model (*CG40625*) is available in the unassembled region (arm U) of the *D. melanogaster* assembly. Using this model, we found that *CG40625* maps to a 100kb scaffold (scaffold_12949) in the *D. virilis* CAF1 assembly that has not been assigned to a Muller element. Hence, we cannot conclusively determine if this gene is on the same or different Muller elements in *D. melanogaster* and *D. virilis*.

B. *CG11231* and *CG11260*
CG11231 and *CG11260* can be mapped to multiple scaffolds in the *D. virilis* CAF1 assembly. *BLASTP* searches against the non-redundant protein database (nr) showed that *CG11231* had weak sequence similarity to reverse transcriptase in *D. melanogaster* and *A. gambiae*. A *BLASTP* search of *CG11260* against the nr protein database showed that it contained a conserved integrase core

domain. Hence our analysis suggests that these two gene annotations in the *D. melanogaster* genome may in fact be remnants of repetitive elements that have been annotated as genes.

C. *CG32021*
For the gene *CG32021*, a *BLASTP* search against the nr protein database showed only a single high quality (e-value < 1e-5) hit, to the annotation in *D. melanogaster*. Examination of the region surrounding *CG32021* in the *D melanogaster* genome identified flanking Transib and *1360* transposable element sequences. Given the close proximity of repetitive elements and lack of support from any other species, *CG32021* is unlikely to be a real gene and may instead be a repetitive element.

D. *CG33797*
CG33797 cannot be mapped definitively to the *D. virilis* CAF1 assembly using *TBLASTN*. *CG33797* is a short (255nt) gene that is nested within *CG11152* on the *D. melanogaster* dot chromosome and contains a conserved Arl6 domain. *TBLASTN* searches of this protein in *D. melanogaster* against the entire *D. virilis* CAF1 assembly detected a single significant hit to scaffold_12875. However, additional investigation revealed that the aligned region is limited to the conserved Arl6 domain and that this region of the CAF1 assembly in *D. virilis* most likely contains a putative ortholog to *CG7735* (another protein in *D. melanogaster* that contains the Arl6 domain). We also extracted the region that encompassed the putative ortholog to *CG11152* (Dvir\GJ15974) and searched this region against the *D. melanogaster* protein *CG33797* using *TBLASTN* with more sensitive parameters. This *TBLASTN* search did not reveal any significant alignments. There are three possible explanations for why this gene is missing from the *D. virilis* CAF1 assembly: it could be a species-specific gene, found only in *D. melanogaster*; it could be present in other species but in regions that are not part of the CAF1 assembly (e.g. in gaps or heterochromatic regions); or it could be an error in the *D. melanogaster* annotation. Given the available evidence, we favor the latter explanation.

Reconstruction of discrepant regions in the CAF1 assembly:
A 2kb region encompassing the discrepant coordinates was extracted from the CAF1 assembly. A *megablast* search was performed against the *D. virilis* WGS database in the NCBI Trace Archive using this extracted region as the query with default parameters and the e-value cutoff at 1e-10 (ZHANG *et al.* 2000). All the traces that showed sequence similarity to the region of interest were downloaded from the NCBI Trace Archive. These traces were then assembled using the *phredPhrap* script and the resulting assembly was examined using *consed* (EWING *et al.* 1998; GORDON *et al.* 1998).

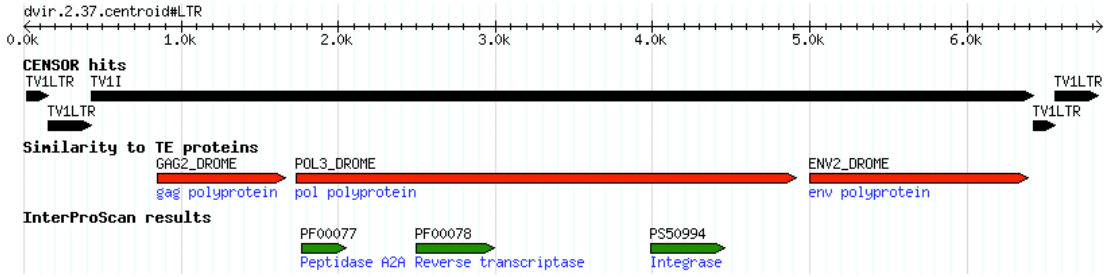
Possible errors in the coding regions of the CAF1 dot chromosome sequence
A. *CG33521*
CG33521 shows an incomplete alignment (with 130/140 bases aligned) between the GEP and CAF1 strains. Reconstruction of this discrepant region of the CAF1 assembly suggests that the differences in the last 10 bases of this region are genuine (see *above* for details on the reconstruction process). Interestingly, there is an alternative canonical splice donor site near the end of the alignment that would keep the rest of the model in the same reading frame (data not shown).

B. *Thd1*
The last coding exon of *Thd1* had the least sequence identity between the two strains among all the dot chromosome exons we have analyzed. Direct mapping of the exon from the GEP model to the CAF1 assembly shows that these differences introduce two premature stop codons in the peptide translation for the CAF1 strain. Attempts to reconstruct this region of the CAF1 assembly using *phred/phrap/Consed* created a new consensus sequence that matched our GEP consensus sequence (data not shown). Hence the large number of differences in this *Thd1* exon is likely caused by errors in the CAF1 consensus sequence. However, we should note that the CAF1 consensus is generated using a different technique (e.g. bases were called using the *KB* base caller and the final assembly was generated by reconciling the *Arachne* and the *Celera* assemblies). Because each assembler has unique strengths and weaknesses, we cannot rule out the possibility that the differences between the two strains of *D. virilis* are genuine.

Curation of four novel LTR retroelement that are recently active in the *D. virilis* dot chromosome

Curation strategy: The consensus sequences for each repetitive element were searched against the Drosophila Repbase repeat library (version 15.04) using the *CENSOR* program (KOHANY *et al.* 2006) to identify regions with sequence similarity to known repetitive elements. Each consensus sequence is also searched against a library of transposable element proteins (from the RepeatRunner package) using *BLASTX*. Finally, InterProScan (ZDOBNOV and APWEILER. 2001) was used to identify conserved protein domains in the consensus sequence for each repetitive element. These results are filtered, collected, and rendered using a custom *Perl* script that utilizes the Bio::Graphics CPAN package. The range of each feature is also summarized in a table. The Sim column in the table corresponds to the similarity between two aligned fragments as defined by *CENSOR*.

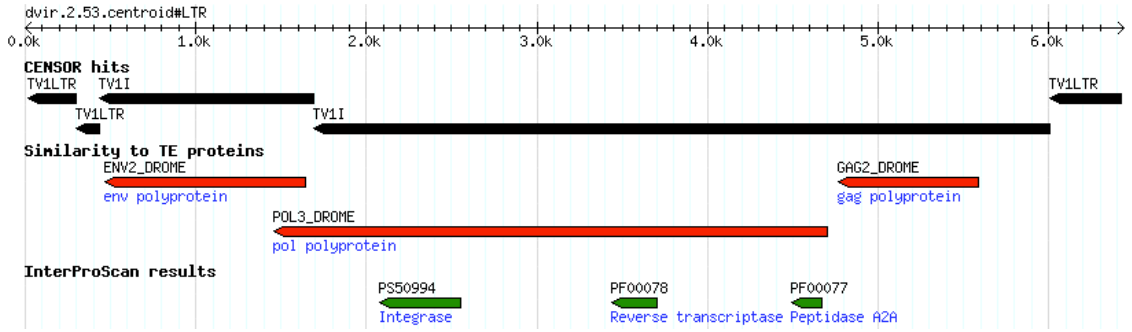
A. *dvir.2.37.centroid*



The dvir.2.37.centroid consensus sequence has a high degree of sequence similarity with the repetitive element TV1. TV1 is an LTR retroelement that is a member of the Gypsy family. The consensus sequence contains the gag, pol, and env polyproteins. The order of peptidase, reverse transcriptase, integrase protein domains within the pol polyprotein is consistent with the assignment of this consensus repetitive sequence as a member of the Gypsy family.

Name	Sim	Start	End	Strand	Source	Description
TV1LTR	0.9925	18	151	+	CENSOR	LTR from TV1
TV1LTR	0.9928	152	430	+	CENSOR	LTR from TV1
TV1I	0.9804	431	6423	+	CENSOR	Internal portion of TV1
TV1LTR	1.0000	6424	6556	+	CENSOR	LTR from TV1
TV1LTR	0.9928	6557	6835	+	CENSOR	LTR from TV1
GAG2_DROME	NA	846	1667	+	BLASTX_TE	gag polyprotein
POL3_DROME	NA	1733	4916	+	BLASTX_TE	pol polyprotein
ENV2_DROME	NA	4997	6386	+	BLASTX_TE	env polyprotein
PF00077	NA	1768	2053	+	InterProScan	Peptidase A2A
PF00078	NA	2497	2992	+	InterProScan	Reverse transcriptase
PS50994	NA	3989	4460	+	InterProScan	Integrase

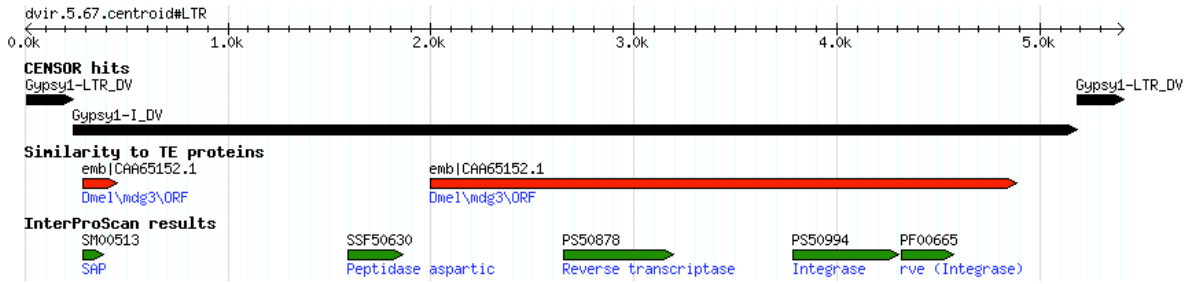
B. *dvir.2.53.centroid*



Similar to dvir.2.37.centroid, the dvir.2.53.centroid has high degree of sequence similarity with the repetitive element TV1. The gag, pol, env polyproteins can be identified in the consensus sequence using BLASTX. The order of peptidase, reverse transcriptase, and integrase protein domains within the pol polyprotein is consistent with the assignment of this consensus repetitive sequence as a member of the Gypsy family.

Name	Sim	Start	End	Strand	Source	Description
TV1LTR	0.9928	21	299	-	CENSOR	LTR from TV1
TV1LTR	1.0000	300	432	-	CENSOR	LTR from TV1
TV1I	0.9833	433	1688	-	CENSOR	Internal portion of TV1
TV1I	0.9765	1689	6005	-	CENSOR	Internal portion of TV1
TV1LTR	0.9904	6006	6419	-	CENSOR	LTR from TV1
ENV2_DROME	NA	470	1643	-	BLASTX_TE	env polyprotein
POL3_DROME	NA	1456	4701	-	BLASTX_TE	pol polyprotein
GAG2_DROME	NA	4767	5588	-	BLASTX_TE	gag polyprotein
PS50994	NA	2077	2548	-	InterProScan	Integrase
PF00078	NA	3440	3704	-	InterProScan	Reverse transcriptase
PF00077	NA	4495	4666	-	InterProScan	Peptidase A2A

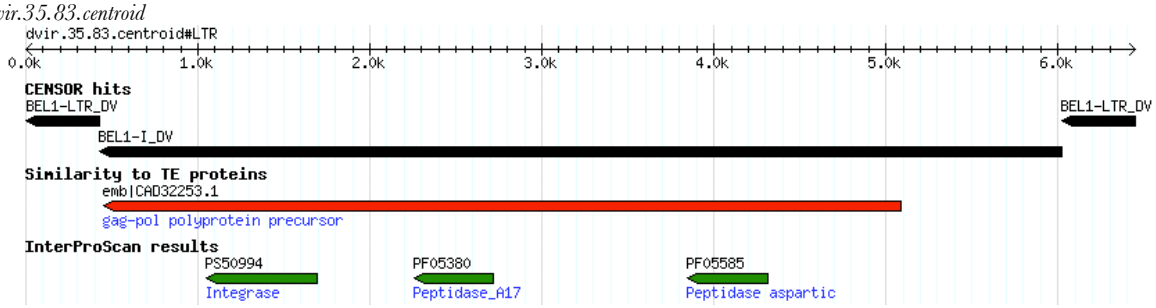
C. *dvir.5.67.centroid*



The dvir.5.67.centroid consensus sequence has high degree of sequence similarity with the repetitive element Gypsy1. Gypsy1 is an LTR retroelement that is a member of the Gypsy family. BLASTX detected sequence homology with the open reading frame within the mdg3 retroelement in *D. melanogaster*. In addition to the peptidase, reverse transcriptase, integrase protein domains, InterProScan also detected a conserved SAP domain (which is a DNA binding domain) within the consensus sequence.

Name	Sim	Start	End	Strand	Source	Description
Gypsy1-LTR_DV	0.9912	9	235	+	CENSOR	LTR from Gypsy1
Gypsy1-I_DV	0.9988	236	5185	+	CENSOR	Internal portion of Gypsy1
Gypsy1-LTR_DV	0.9912	5186	5412	+	CENSOR	LTR from Gypsy1
emb|CAA65152.1	NA	288	452	+	BLASTX_TE	Dmel\mdg3\ORF
emb|CAA65152.1	NA	1998	4884	+	BLASTX_TE	Dmel\mdg3\ORF
SM00513	NA	287	389	+	InterProScan	SAP
SSF50630	NA	1589	1862	+	InterProScan	Peptidase aspartic
PS50878	NA	2654	3194	+	InterProScan	Reverse transcriptase
PS50994	NA	3783	4305	+	InterProScan	Integrase
PF00665	NA	4317	4575	+	InterProScan	rve (Integrase)

D. *dvir.35.83.centroid*



The dvir.35.83.centroid consensus sequence has high degree of sequence similarity with the repetitive element BEL1. BEL1 is a LTR retroelement that is a member of the BEL family. BLASTX searches against the TE database revealed a gag-pol polyprotein precursor within the consensus sequence. InterProScan detected two peptidase domains and a single integrase domain within the consensus sequence.

Name	Sim	Start	End	Strand	Source	Description
BEL1-LTR_DV	0.9953	2	428	-	CENSOR	LTR from BEL1
BEL1-I_DV	0.9964	429	6020	-	CENSOR	Internal portion of BEL1
BEL1-LTR_DV	0.9953	6021	6447	-	CENSOR	LTR from BEL1
emb|CAD32253.1	1	453	5087	-	BLASTX_TE	gag-pol polyprotein precursor
PS50994	1	1048	1693	-	InterProScan	Integrase
PF05380	1	2255	2720	-	InterProScan	Peptidase_A17
PF05585	1	3848	4310	-	InterProScan	Peptidase aspartic

LITERATURE CITED

DROSOPHILA 12 GENOMES CONSORTIUM, A. G. CLARK, M. B. EISEN, D. R. SMITH, C. M. BERGMAN *et al*, 2007 Evolution of genes and genomes on the Drosophila phylogeny. Nature **450:** 203-218.

ELSIK, C. G., A. J. MACKEY, J. T. REESE, N. V. MILSHINA, D. S. ROOS *et al*, 2007 Creating a honey bee consensus gene set. Genome Biol. **8:** R13.

EWING, B., L. D. HILLIER, M. C. WENDL and P. GREEN, 1998 Base-calling of automated sequencer traces using PHRED. I. accuracy assessment. Genome Res. **8:** 175-185.

GORDON, D., C. ABAJIAN and P. GREEN, 1998 CONSED: A graphical tool for sequence finishing. Genome Res. **8:** 195-202.

KOHANY, O., A. J. GENTLES, L. HANKUS AND J. JURKA, 2006 Annotation, submission and screening of repetitive elements in repbase: RepbaseSubmitter and censor. BMC Bioinformatics **7:** 474.

ZHANG, Z., S. SCHWARTZ, L. WAGNER and W. MILLER, 2000 A greedy algorithm for aligning DNA sequences. Journal of Computational Biology **7:** 203-214.

ZDOBNOV, E. M., AND R. APWEILER, 2001 InterProScan-an integration platform for the signature-recognition methods in InterPro. Bioinformatics **17:** 847-848.

FILE S2

Species-specific RepeatModeler libraries for *D. melanogaster*

File S2 is available for download as a text file at http://www.genetics.org/cgi/content/full/genetics.110.116129/DC1.

File S3

Species-specific RepeatModeler libraries for *D. virilis*

File S3 is available for download as a text file at http://www.genetics.org/cgi/content/full/genetics.110.116129/DC1.

TABLE S1

Total repeat density for different regions of *D. melanogaster* and *D. virilis*

Species: Muller Element	RepeatMasker + RepBase (Drosophila)	RepeatRunner + Superlibrary	RepeatMasker + RepeatModeler (Species-specific)
D. melanogaster:			
Muller D Het. Region	65.7%	73.2%	68.1%
D. virilis (CAF1):			
Muller D Het. Region	37.4%	55.3%	65.5%
D. melanogaster:			
Muller F (Dot)	27.2%	30.8%	28.4%
D. virilis (GEP):			
Muller F (Dot)	17.9%	24.2%	27.1%
D. virilis (CAF1):			
Muller F (Dot)	17.5%	23.1%	25.9%
D. melanogaster:			
Muller D Euch. Region	8.7%	10.2%	8.9%
D. virilis (CAF1):			
Muller D Euch. Region	9.3%	12.0%	12.6%
D. virilis (GEP):			
Random Fosmids	6.6%	9.7%	9.1%
D. virilis (CAF1):			
Random Fosmids	8.8%	11.4%	10.5%

Overall repeat densities for different regions of the *D. melanogaster* and *D. virilis* genomes; values used in Figure 1.

TABLE S2

Repeat class distributions using *RepeatMasker* with the RepeatModeler library

Species: Muller Element	LINEs	LTR elements	DNA Transposons	Unclassified	Other (Simple + Low Complexity)
D. melanogaster:					
Muller D Het. Region	20.2%	32.5%	11.1%	3.1%	1.2%
D. virilis (CAF1):					
Muller D Het. Region	11.6%	17.4%	18.9%	16.7%	1.0%
D. melanogaster:					
Muller F (Dot)	3.4%	4.3%	16.5%	1.1%	3.1%
D. virilis (GEP):					
Muller F (Dot)	2.7%	3.5%	11.2%	4.7%	5.1%
D. virilis (CAF1):					
Muller F (Dot)	2.5%	2.1%	11.4%	4.7%	5.3%
D. melanogaster:					
Muller D Euch. Region	1.7%	3.0%	2.5%	0.1%	1.6%
D. virilis (CAF1):					
Muller D Euch. Region	1.0%	0.2%	3.6%	2.6%	5.3%
D. virilis (GEP):					
Random Fosmids	0.1%	0.1%	1.0%	2.7%	5.1%
D. virilis (CAF1):					
Random Fosmids	2.2%	0.1%	0.8%	2.2%	5.2%

Presence of different repeat classes; values used for Figure 2.

TABLE S3

Tandem repeats density for different regions of *D. melanogaster* and *D. virilis*

Species: Muller Element	*D. melanogaster*	*D. virilis*
Muller D Het. Region	1.1%	3.8%
Muller F (Dot)	2.2%	3.3% (GEP) / 3.3% (CAF1)
Muller D Euch. Region	0.8%	3.7%

Tandem repeats densities; values used for Figure 3.

TABLE S4

Five-number summary statistics for the distribution of gene sizes

Region	Minimum	1st Quartile	Median	3rd Quartile	Maximum
D. melanogaster: Muller D Het. Region	228.00	1251.00	2429.00	8648.00	146400.00
D. melanogaster: Muller F (Dot)	258.00	2542.00	6193.00	10740.00	49830.00
D. virilis (GEP): Muller F (Dot)	453.00	2409.00	5718.00	9928.00	51420.00
D. virilis (CAF1): Muller F (Dot)	168.00	1934.00	4999.00	8595.00	61000.00
D. melanogaster: Muller D Euch. Region	108.00	591.00	1062.00	2490.00	111600.00
D. virilis (CAF1): Muller D Euch. Region	165.00	960.00	1792.00	3518.00	62850.00

TABLE S5

Two-sided Kolmogorov-Smirnov tests on gene sizes (raw p-values)

	D. melanogaster Dot	*D. virilis* Dot (GEP)	*D. virilis* Dot (CAF1)	*D. melanogaster* Euch. Reference	*D. virilis* Euch. Reference
D. melanogaster					
Het. Reference	3.84E-02	6.54E-02	3.33E-01	4.85E-02	2.13E-01
D. melanogaster					
Dot	-	9.13E-01	2.40E-01	1.74E-12*	1.26E-10*
D. virilis					
Dot (GEP)	-	-	6.02E-01	1.58E-11*	5.45E-09*
D. virilis					
Dot (CAF1)	-	-	-	5.95E-09*	8.07E-07*
D. melanogaster					
Euch. Reference	-	-	-	-	3.65E-03

p-values from the two-sided Kolmogorov-Smirnov (KS) tests indicate that the difference in gene sizes between the dot chromosomes of *D. melanogaster* and *D. virilis* are not statistically significant. The differences between the gene sizes on the dot chromosomes are systematically different from the gene sizes in the euchromatic reference regions. Asterisks (*) indicate cells with raw p-values below 3.33E-03 (0.05/15), considered to be statistically significant.

TABLE S6

Five-number summary statistics for the distribution of coding exon sizes

Region	Minimum	1st Quartile	Median	3rd Quartile	Maximum
D. melanogaster: Muller D Het. Region	23.00	125.00	172.00	271.00	2493.00
D. melanogaster: Muller F (Dot)	3.00	123.00	199.00	380.50	9471.00
D. virilis (GEP): Muller F (Dot)	3.00	123.00	201.00	433.00	9474.00
D. virilis (CAF1): Muller F (Dot)	9.00	120.20	196.50	403.80	9474.00
D. melanogaster: Muller D Euch. Region	3.00	150.50	284.00	582.00	3328.00
D. virilis (CAF1): Muller D Euch. Region	3.00	130.50	238.50	568.50	6779.00

TABLE S7

Two-sided Kolmogorov-Smirnov tests on coding exon sizes (raw p-values)

	D. melanogaster Dot	*D. virilis* Dot (GEP)	*D. virilis* Dot (CAF1)	*D. melanogaster* Euch. Reference	*D. virilis* Euch. Reference
D. melanogaster Het. Reference	2.48E-02	1.50E-02	3.51E-02	8.82E-06*	1.03E-04*
D. melanogaster Dot	-	7.23E-01	9.96E-01	7.02E-05*	6.17E-05*
D. virilis Dot (GEP)	-	-	9.93E-01	1.38E-03*	5.72E-03
D. virilis Dot (CAF1)	-	-	-	3.05E-04*	6.36E-04*
D. melanogaster Euch. Reference	-	-	-	-	3.08E-01

p-values from the two-sided KS tests indicate that the coding exon sizes for the heterochromatic reference region are distinct from the coding exon sizes on the euchromatic reference regions. The coding exon sizes in the dot chromosomes of *D. melanogaster* and *D. virilis* are statistically different from the euchromatic reference regions. The coding exon sizes on the *D. melanogaster* dot chromosome are not statistically different from the coding exon sizes on the *D. virilis* dot chromosome. Asterisks (*) indicate cells with raw p-values below 3.33E-03, considered to be statistically significant.

TABLE S8

Five-number summary statistics for the distribution of intron sizes

Region	Minimum	1st Quartile	Median	3rd Quartile	Maximum
D. melanogaster: Muller D Het. Region	50.00	65.25	752.50	4898.00	40110.00
D. melanogaster: Muller F (Dot)	47.00	61.00	201.50	940.20	20170.00
D. virilis (GEP): Muller F (Dot)	48.00	68.00	144.50	674.00	19550.00
D. virilis (CAF1): Muller F (Dot)	33.00	67.00	131.00	735.00	22960.00
D. melanogaster: Muller D Euch. Region	41.00	60.00	74.00	465.50	66250.00
D. virilis (CAF1): Muller D Euch. Region	41.00	61.00	69.00	276.00	39000.00

TABLE S9

One-sided Wilcoxon Rank Sum tests on intron sizes (raw p-values)

	D. melanogaster Dot	*D. virilis* Dot (GEP)	*D. virilis* Dot (CAF1)	*D. melanogaster* Euch. Reference	*D. virilis* Euch. Reference
D. melanogaster					
Het. Reference	2.95E-05*	3.33E-04*	3.85E-04*	1.87E-06*	9.32E-08*
D. melanogaster					
Dot	-	8.96E-01	9.05E-01	2.18E-03*	3.32E-05*
D. virilis					
Dot (GEP)	-	-	4.93E-01	3.66E-06*	1.21E-12*
D. virilis					
Dot (CAF1)	-	-	-	5.35E-06*	2.74E-12*
D. melanogaster					
Euch. Reference	-	-	-	-	3.98E-01

p-values from the Wilcoxon Rank Sum tests (one-sided test with alternative="greater") indicate that the intron sizes for the heterochromatic reference region are systematically larger than the intron sizes on the dot chromosomes and the euchromatic reference regions. The intron sizes in the dot chromosomes of *D. melanogaster* and *D. virilis* are systematically larger than the euchromatic reference regions. However, the intron sizes on the *D. melanogaster* dot chromosome are not statistically different from the intron sizes on the *D. virilis* dot chromosome. Asterisks (*) indicate cells with raw p-values below 3.33E-03, considered to be statistically significant.

TABLE S10

Five-number summary statistics for the distribution of intron sizes after removing repetitive elements

Region	Minimum	1st Quartile	Median	3rd Quartile	Maximum
D. melanogaster: Muller D Het. Region	2.00	61.50	561.00	2017.00	152900.00
D. melanogaster: Muller F (Dot)	2.00	59.00	176.00	621.00	267900.00
D. virilis (GEP): Muller F (Dot)	18.00	66.00	117.00	436.00	258400.00
D. virilis (CAF1): Muller F (Dot)	0.00	66.00	115.00	454.50	285600.00
D. melanogaster: Muller D Euch. Region	35.00	59.50	74.00	438.50	267400.00
D. virilis (CAF1): Muller D Euch. Region	27.00	61.00	68.00	269.50	273100.00

TABLE S11

One-sided Wilcoxon Rank Sum tests on intron sizes after removing repetitive elements (raw p-values)

	D. melanogaster Dot	*D. virilis* Dot (GEP)	*D. virilis* Dot (CAF1)	*D. melanogaster* Euch. Reference	*D. virilis* Euch. Reference
D. melanogaster					
Het. Reference	2.97E-05*	3.07E-04*	3.20E-04*	1.07E-04*	5.08E-06*
D. melanogaster					
Dot	-	7.79E-01	7.54E-01	6.90E-02	3.61E-03
D. virilis					
Dot (GEP)	-	-	4.67E-01	1.22E-03*	1.24E-08*
D. virilis					
Dot (CAF1)	-	-	-	2.13E-03*	5.07E-08*
D. melanogaster					
Euch. Reference	-	-	-	-	2.84E-01

After removing repetitive elements from the introns, p-values from the Wilcoxon Rank Sum tests (one-sided test with alternative="greater") indicate that the intron sizes for the heterochromatic reference region are still systematically larger than the intron sizes on the dot chromosomes and the euchromatic reference regions. However, the difference in intron sizes between the dot chromosome and the euchromatic reference region in *D. melanogaster* is no longer statistically significant. In most cases where the differences in intron sizes were statistically significant, the p-value increased when repeats are removed from the introns. Our results suggest that the larger intron sizes on the dot chromosomes and the heterochromatic reference regions can partly be attributed to repeats within introns. Asterisks (*) indicate cells with raw p-values below 3.33E-03, considered to be statistically significant.

TABLE S12

Five-number summary statistics for the distribution of codon bias

Region	Minimum	1st Quartile	Median	3rd Quartile	Maximum
D. melanogaster: Muller D Het. Region	41.40	49.52	52.28	54.52	61.00
D. melanogaster: Muller F (Dot)	45.52	51.65	53.92	56.15	61.00
D. virilis (GEP): Muller F (Dot)	44.10	54.90	56.55	57.70	61.00
D. virilis (CAF1): Muller F (Dot)	34.42	54.42	56.30	57.48	61.00
D. melanogaster: Muller D Euch. Region	27.99	45.10	49.39	54.53	61.00
D. virilis (CAF1): Muller D Euch. Region	32.23	45.44	49.14	52.86	61.00

TABLE S13

Two-sided Kolmogorov-Smirnov tests on codon bias (raw p-values)

	D. melanogaster Dot	*D. virilis* Dot (GEP)	*D. virilis* Dot (CAF1)	*D. melanogaster* Euch. Reference	*D. virilis* Euch. Reference
D. melanogaster					
Het. Reference	5.28E-02	3.42E-05*	2.01E-04*	3.61E-02	5.37E-02
D. melanogaster					
Dot	-	1.53E-06*	6.28E-06*	7.89E-08*	9.32E-10*
D. virilis					
Dot (GEP)	-	-	7.50E-01	5.55E-16*	< 2.2E-16*
D. virilis					
Dot (CAF1)	-	-	-	2.22E-14*	< 2.2E-16*
D. melanogaster					
Euch. Reference	-	-	-	-	6.01E-01

p-values from the two-sided KS tests indicate that the difference in codon bias between the dot chromosomes from *D. virilis* and *D. melanogaster* and the differences between the dot chromosomes and the euchromatic reference regions are statistically significant. Differences between the euchromatic regions of *D. melanogaster* and *D. virilis* are not statistically significant. Asterisks (*) indicate cells with raw p-values below 3.33E-03, considered to be statistically significant.

TABLE S14

Ka/Ks ratios for 22 genes with both synonymous and non-synonymous differences between the two strains of the *D. virilis* dot chromosome

Putative *D. melanogaster ortholog*	Gene Length	Seq. Identity	Seq. Similarity	Gaps	Protein Percent ID	cDNA Percent ID	Ka	Ks	Ka/Ks
CG32016-RB	3324	3318	3318	3	99.82	99.91	0.0009	0.0009	1.0943
Ephrin-RA	2055	2044	2044	6	99.41	99.76	0.0025	0.0023	1.0732
ci-RA	4338	4332	4332	3	99.86	99.93	0.0007	0.0007	1.0410
GEP001-RA	1746	1742	1742	0	99.48	99.77	0.0023	0.0023	0.9814
onecut-RA	3231	3217	3217	6	99.53	99.75	0.0022	0.0034	0.6468
pho-RA	1557	1555	1555	0	99.81	99.87	0.0010	0.0019	0.5402
Ank-RA	4899	4897	4897	0	99.94	99.96	0.0003	0.0008	0.3518
gw-RA	4185	4177	4177	3	99.78	99.86	0.0010	0.0028	0.3453
bip2-RA	4731	4724	4724	0	99.81	99.85	0.0009	0.0027	0.3432
Rfabg-RA	10035	10024	10024	0	99.85	99.89	0.0007	0.0021	0.3409
CG5262-RA	1518	1516	1516	0	99.80	99.87	0.0009	0.0027	0.3273
CG32000-RH	4383	4380	4380	0	99.93	99.93	0.0004	0.0013	0.2771
CG31999-RA	2769	2765	2765	0	99.78	99.86	0.0009	0.0035	0.2669
lgs-RA	4503	4500	4500	0	99.93	99.93	0.0003	0.0013	0.2513
CG33978-RA	13746	13724	13724	0	99.83	99.84	0.0008	0.0035	0.2441
CG32017-RB	1515	1512	1512	0	99.80	99.80	0.0009	0.0052	0.1747
zfh2-RA	9621	9586	9586	12	99.78	99.76	0.0010	0.0066	0.1498
Actbeta-RA	2787	2771	2771	12	99.89	99.82	0.0005	0.0065	0.0724
Asator-RD	4248	4243	4243	0	99.93	99.88	0.0003	0.0042	0.0720
RhoGAP102A-RE	3231	3225	3225	0	99.91	99.81	0.0004	0.0065	0.0629
bt-RF	26898	26894	26894	0	99.98	99.99	0.0009	0.0178	0.0530
CG11093-RB	2325	2318	2318	0	99.87	99.70	0.0006	0.0118	0.0473

TABLE S15

Ka/Ks ratios for 9 genes with both synonymous and non-synonymous differences between random fosmids in two strains of *D. virilis*

Putative *D. melanogaster* ortholog	Gene Length	Seq. Identity	Seq. Similarity	Gaps	Protein Percent ID	cDNA Percent ID	Ka	Ks	Ka/Ks
Oamb	1025	1004	1004	18	99.40	99.70	0.0029	0.0035	0.8179
CG33260	366	360	360	0	96.69	98.35	0.0150	0.0230	0.6507
CG14130	597	595	595	0	99.49	99.66	0.0020	0.0111	0.1823
CG1732	1821	1818	1818	0	99.83	99.83	0.0008	0.0043	0.1797
Best1	2214	2199	2199	0	99.59	99.32	0.0025	0.0185	0.1368
CG10440	1059	1055	1055	0	99.72	99.62	0.0013	0.0111	0.1166
CG32521	1884	1878	1878	0	99.84	99.68	0.0007	0.0105	0.0682
CG9384	1770	1758	1758	0	99.66	99.32	0.0015	0.0239	0.0632
Egfr	4062	4053	4053	0	99.93	99.78	0.0003	0.0102	0.0305