Drosophila muller f elements maintain a distinct set of genomic properties over 40 million years of evolution

Sarah C.R. Elgin
*Washington University in St. Louis*, selgin@wustl.edu

Wilson Leung
*Washington University in St. Louis*, wleung@wustl.edu

Follow this and additional works at: [https://openscholarship.wustl.edu/bio_facpubs](https://openscholarship.wustl.edu/bio_facpubs)

Recommended Citation

This Article is brought to you for free and open access by the Biology at Washington University Open Scholarship. It has been accepted for inclusion in Biology Faculty Publications & Presentations by an authorized administrator of Washington University Open Scholarship. For more information, please contact digital@wumail.wustl.edu.
Drosophila Muller F Elements Maintain a Distinct Set of Genomic Properties Over 40 Million Years of Evolution

Wilson Leung and Participating Students and Faculty of the Genomics Education Partnership

ABSTRACT The Muller F element (4.2 Mb, ~80 protein-coding genes) is an unusual autosome of Drosophila melanogaster; it is mostly heterochromatic with a low recombination rate. To investigate how these properties impact the evolution of repeats and genes, we manually improved the sequence and annotated the genes on the D. erecta, D. mojavensis, and D. grimshawi F elements and euchromatic domains from the Muller D element. We find that F elements have greater transposon density (25–50%) than euchromatic reference regions (3–11%). Among the F elements, D. grimshawi has the lowest transposon density (particularly DINE-1: 2% vs. 11–27%). F element genes have larger coding spans, more coding exons, larger introns, and lower codon bias. Comparison of the Effective Number of Codons with the Codon Adaptation Index shows that, in contrast to the other species, codon bias in D. grimshawi F element genes can be attributed primarily to selection instead of mutational biases, suggesting that density and types of transposons affect the degree of local heterochromatin formation. F element genes have lower estimated DNA melting temperatures than D element genes, potentially facilitating transcription through heterochromatin. Most F element genes (~90%) have remained on that element, but the F element has smaller syntenic blocks than genome averages (3.4–3.6 vs. 8.4–8.8 genes per block), indicating greater rates of inversion despite lower rates of recombination. Overall, the F element has maintained characteristics that are distinct from other autosomes in the Drosophila lineage, illuminating the constraints imposed by a heterochromatic milieu.

With an estimated size of 4.2 Mb overall, the Drosophila melanogaster Muller F element, (also known as the dot chromosome, or the fourth chromosome in that species) is unusual in that it appears entirely heterochromatic by most criteria, but the distal 1.3 Mb has a gene density and fraction of active genes (~50% in S2 cells) that are similar to the euchromatic regions of the D. melanogaster genome (Riddle et al. 2009, 2012). Insertion of a PEV reporter (hsp70-driven white) in most cases results in a variegating phenotype (partial silencing; see Supplemental Text in File S1), indicating that even this distal region of the F element is packaged as heterochromatin (Sun et al. 2004; Riddle et al. 2008). Subsequent high-resolution mapping of the chromatin landscape of the F element supports this conclusion (Riddle et al. 2012). These characteristics of the F element have made it an ideal platform for elucidating factors that are involved in heterochromatin formation and for exploring their impact on genes that are embedded in a heterochromatic domain (Elgin and Reuter 2013).

Immunofluorescent staining of polytene chromosomes with antibodies directed against H3K9me2 shows that, similar to D. melanogaster, the F elements of D. erecta, D. mojavensis, and D. grimshawi also are enriched in H3K9me2 (Figure 1, left). These enrichment patterns indicate that the F element has maintained its heterochromatic properties in species (i.e., D. mojavensis and D. grimshawi) that last shared a common...
ancestor with *D. melanogaster* about 40 million years ago (Powell 1997; Figure 1, right).

To investigate the evolution of this unusual domain, we performed comparative analyses of the repeat and gene characteristics of the F element in four *Drosophila* species. The *Drosophila* 12 Genomes Consortium (*Drosophila* 12 Genomes Consortium *et al.* 2007) and the modENCODE project (Kharchenko *et al.* 2011) have produced a large collection of genomic datasets for *D. melanogaster* and 11 other *Drosophila* species. Previous analyses of the evolution of these *Drosophila* species have relied primarily on the Comparative Analysis Freeze 1 (CAF1) draft assembly and computational (GLEAN-R) gene predictions (*Drosophila* 12 Genomes Consortium *et al.* 2007). Most of these analyses only focused on the Muller elements A–E and the properties of the F element generally have not been examined carefully.

In this study, we have built on these genomic resources by performing manual sequence improvement and gene annotation of the *D. erecta*, *D. mojavensis*, and *D. grimshawi* F elements and euchromatic reference regions derived from the Muller D elements. The D element analysis regions (referred to as “base”) are located proximal to the pericentric heterochromatin so that they have a similar topological position in the nucleus as the F element. To identify characteristics that are associated with the proximity to pericentric or telomeric heterochromatin, we also analyzed two additional euchromatic regions from the *D. erecta* D element: a 1.4-Mb region that extends further from the base of the D element (referred to as “extended”) and a 1.3-Mb region adjacent to the telomeric region of the D element (referred to as “telomeric”). [See the exact coordinates of all the analysis regions in Table S1, Genome Browser views (showing repeat density and gaps) in Figure S1, and a detailed description of how these regions were selected in File S1.]

The high-quality assemblies and gene annotations generated in this study enable us to address several questions about the evolution of the F element: What are the differences in the types and distributions of repeats among the F elements? Do F element genes exhibit different characteristics (e.g., coding spans, intron sizes) compared with genes on the other autosomes? How does the low recombination rate affect codon bias, the selective pressure experienced by F element genes, and the frequency of gene movement?

Our analyses show that F element genes in both the Sophophora and *Drosophila* clades have maintained a set of distinct characteristics (larger gene size, lower codon bias, lower melting temperature) compared with genes on other autosomes. Most of the *D. melanogaster* F element genes (~90%) have remained on the same Muller element in all four *Drosophila* species, but there have been a large number of inversions. F elements of the species in the *Drosophila* clade (i.e., *D. mojavensis* and *D. grimshawi*) exhibit different repeat distributions and gene characteristics compared to the species in the melanogaster subgroup (i.e., *D. melanogaster* and *D. erecta*). F element genes generally exhibit lower codon bias and weaker positive selection compared to genes in the euchromatic reference regions; these characteristics are less pronounced in *D. grimshawi*, which also has a much lower density of the *Drosophila* INterspersed Element 1 (DINE-1) transposon. Despite these differences, our analyses show that F element genes in all four species generally share a common set of characteristics that presumably reflect the local environment and could contribute to their ability to function in a heterochromatic domain.

**MATERIALS AND METHODS**

**General overview**

Sequence improvement and gene annotation of the three *Drosophila* species studied here were organized using the framework provided by Genomics Education Partnership (Shaffer *et al.* 2010). Additional details for some of the analysis protocols are available in File S1. We have set up an instance of the University of California, Santa Cruz (UCSC) Genome Browser (Kent *et al.* 2002) to facilitate the visualization and access to the improved sequences and gene annotations produced in this study (available at http://gander.wustl.edu). The improved sequences and annotations are also available in File S9.

Most of the data conversions were performed with the use of tools in the Kent source utilities (part of the UCSC Genome Browser source
Immunofluorescent staining of polytene chromosomes
The *D. erecta* (14021—0224.01), *D. mojavensis* (15081—1352.22), and *D. grimshawi* (15287—2541.00) stocks were obtained from the *Drosophila* Species Stock Center at the University of California, San Diego. The protocol for the immunofluorescent staining of polytene chromosomes from *Drosophila* third instar larval salivary glands has been described previously (Stephens et al. 2004). An anti-H3K9me2 rabbit polyclonal antibody (Upstate 07-441) was used at a dilution of 1:250. Secondary antibody labeled with Alexa-Fluor 594 (red) was used at a 1:750 dilution (Invitrogen, catalog number A-11012). Formaldehyde fixation times were 12 min, with the exception of *D. grimshawi* salivary glands, which were fixed for 10 min before squashing and staining.

Sequence improvement
The *D. mojavensis* and *D. grimshawi* CAF1 assemblies produced by the *Drosophila* 12 Genomes Consortium were retrieved from the AAA: 12 *Drosophila* Genomes web site (http://rana.lbl.gov/drosophila/). The placements of the fosmid end reads were specified in the *reads_placed* file in each CAF1 assembly. The F and D element scaffolds were partitioned into a list of overlapping fosmids based on the *reads_placed* file for each species. This list of fosmids was obtained from the *Drosophila* Genomics Resource Center at Indiana University and used as templates for sequencing reactions. However, because many of the fosmid clones used to construct the original *D. grimshawi* CAF1 assemblies were unavailable from the *Drosophila* Genomics Resource Center, we could only improve approximately 90% of the *D. grimshawi* F element. Hence the analysis of this region was performed on a mosaic of the original CAF1 assembly and improved regions.

The overall sequence improvement protocol has previously been described (Slawson et al. 2006; Leung et al. 2010). Reads placed in each fosmid region were retrieved from the National Center for Biotechnology Information Trace Archive (http://www.ncbi.nlm.nih.gov/Traces/home/) and assembled using the Phred, Phrap, and Consed software package (Ewing and Green 1998; Gordon et al. 1998). In collaboration with the Genome Institute at Washington University, we improved each fosmid project by identifying and resolving misassemblies as well as designing additional sequencing reactions to resolve gaps and low quality regions. These fosmid projects were improved to a sequence improvement standard similar to the one used by the mouse genome project (Mouse Genome Sequencing Consortium et al. 2002). To ensure the correctness of the final assembly, inconsistent mate pairs within each fosmid project were resolved and restriction digests were used to confirm the final assembly. Each fosmid was digested with four restriction enzymes (*i.e.*, EcoRI, EcoRV, HindIII, and SacI). The fragment sizes of the *in silico* digests of the final consensus sequence must be in congruence with the fragment sizes of at least two of the actual restriction digests to meet the standard. Each fosmid project was completed by at least two students independently; experienced undergraduates worked with the Genomics Education Partnership (GEP) staff to reconcile the results and produce the final consensus sequence.

To identify differences between the CAF1 and improved sequences, the CAF1 sequence was soft-masked using WindowMasker with default parameters. The improved sequences were compared against the original CAF1 sequence using MegaBLAST (Morgulis et al. 2008) with an E-value threshold of 1e-5. The UCSC Chain and Net protocol (Kent et al. 2003) was then applied to the MegaBLAST alignments. The Net alignments were converted into PSL and BED formats to facilitate analysis of the differences between the two assemblies.

Repeat analysis
WindowMasker (Morgulis et al. 2006) was run on the different analysis regions using default parameters and the results were converted into BED format using custom Perl scripts. Tallymer (Kurtz et al. 2008) was used to estimate k-mer frequencies in the different analysis regions. Each genome assembly was indexed using mkindex and the occtratio program was used to determine the distributions of unique k-mer. The count of each 13-mer was generated using the search program in Tallymer. Tandem repeats were identified using Tandem Repeats Finder (Benson 1999) with the following parameters: Match = 2, Mismatch = 7, Delta = 7, Match Probability = 80, Mismatch Probability = 10, Minscore = 50, and MaxPeriod = 2000. Simple repeats and low complexity regions were identified using tantan (Frith 2011) with default parameters (r = 0.005), and the results were reported in BED format (-f 3). The distribution of dinucleotide repeats was determined using a Perl script that iterates from a dinucleotide repeat size of 2—100. Each dinucleotide repeat was searched against the analysis regions and the (potentially overlapping) matches were tabulated and plotted using Microsoft Excel.

Transposon analysis
The protocols used to construct and classify the species-specific transposon libraries are described in File S1. The *Drosophila* RepBase repeat library (release 17.07) was obtained from RepBase (Jurka et al. 2005). The ReAS repeat library (version 2) was obtained from the FlyBase FTP site at ftp://ftp.flybase.net/genomes/aaa/transposable_elements/ReAS/v2/consensus_fasta/.

RepeatMasker (Smit et al. 1996) (version open-3.4.0) was run on the analysis regions using the *cross_match* search engine at the most sensitive (-s) setting, without masking low complexity or simple repeats (-nolow). Transposon fragments identified by RepeatMasker were converted into BED format using custom scripts for subsequent analysis. Overlapping transposon fragments identified by RepeatMasker were merged together using BEDTools only if the overlapping repeats had the same repeat class. Repeat density was calculated using a sliding window of 1 kb with a step size of 500 bp.

Gene annotations
This comparative analysis used the high-quality *D. melanogaster* gene annotations (release 5.50) produced by FlyBase as reference (Marygold et al. 2013). The annotation protocol has been described previously (Shaffer et al. 2010). GEP students annotated each fosmid by using computational evidence organized on an instance of the UCSC Genome Browser (Kent et al. 2002) set up by the GEP staff. The computational evidence included sequence similarity to *D. melanogaster* proteins as well as predictions from multiple *ab initio* and evidence-based gene predictors. For species with RNA-Seq data, additional evidence tracks such as RNA-Seq read coverage, splice junction predictions from TopHat (Trapnell et al. 2009) and assembled transcripts from Cufflinks (Trapnell et al. 2010) were also made available. See File S1 for additional details on the protocol used to construct the RNA-Seq transcriptome and predicted protein libraries for each species.

The GEP has developed a set of annotation guidelines (Annotation Instruction Sheet) to standardize the treatment of annotations that are ambiguous because of insufficient evidence. These annotation guidelines
and additional resources supporting the GEP annotation protocol are available on the GEP web site (http://gep.wustl.edu).

Each annotation project was completed independently by at least two GEP students. The GEP staff supervised students who reconciled the submitted annotations using the Apollo Genome Annotation Curation Tool (Lewis et al. 2002). These reconciled gene annotations were mapped back to the improved genomic scaffolds and were incorporated into the GEP UCSC Genome Browser (available through the "GEP Genes" track, http://gander.wustl.edu). The GEP staff reviewed these gene models in the context of all the available evidence tracks to resolve any remaining annotation issues.

The D. erecta, D. mojavensis, and D. grimshawi GLEAN-R gene annotations (Release 1.3) produced by the Drosophila 12 Genomes Consortium were compared to the annotations produced here. The GLEAN-R annotations were obtained from FlyBase (available at http://flybase.org/static_pages/downloads/bulkdata7.html) and converted into BED format using custom scripts. We used BLAT (Kent 2002) with default parameters to map the D. mojavensis and D. grimshawi GLEAN-R gene predictions against the improved assemblies because the underlying genomic sequences for these two species have changed due to the sequence improvements reported here. Utilities in BEDTools (Quinlan and Hall 2010) and custom scripts were then used to compare the GLEAN-R predictions with our gene annotations.

Analysis of gene characteristics

The gene annotations are in BED format, and most of the gene characteristics (e.g., gene size, coding exon size) were determined using BEDTools (Quinlan and Hall 2010) and custom scripts. When calculating the coding exon sizes for the first and last coding exons, only the translated portion of the exon was included even though the transcribed exon may be larger because of untranslated regions. The gene characteristics of the most comprehensive isoform for each gene were imported into R (version 3.0.2) for subsequent analysis and visualization of the results.

Violin plots of the different gene characteristics were generated by the vioplot function in the vioplot package. The Kruskal-Wallis Rank Sum Test was performed using the kruskal.test function in R (R Core Team 2013). The kruskalmc function in the pgirmess package was used to perform the multiple comparison tests after Kruskal-Wallis.

Codon bias analysis

The Effective Number of Codons (Nc) and the Codon Adaptation Index (CAI) for each gene in the analysis regions were determined using the chips and the cai programs in the EMBOSS package (Rice et al. 2000), respectively. Typically, highly expressed genes are used as the reference set when calculating CAI because they are under the strongest translational selection and would typically show a strong preference for a subset of transfer RNAs (Rocha 2004). Because expression data were unavailable for some of the species used in this study, we used the program scnRCA (O’Neill et al. 2013) to analyze all of the GLEAN-R predictions to construct the species-specific reference gene set that exhibits the dominant codon bias for each species. The scnRCA parameters used to construct the reference gene sets were as follows: -i r -g true -d -1 -2.0 -p 1.0 -m -1.

The codon frequency table for each species was created by analyzing the species-specific reference gene set with the cusp program in the EMBOSS package. The species-specific codon usage tables were then used in the cai program (via the -c parameter) to calculate the CAI value for each gene. The violin plots and Kruskal-Wallis Tests were created using the same procedure as described in the "Analysis of gene characteristics" section.

Heat maps of codon bias for each gene in the analysis regions were created using the heatmap.2 function in the R package gplots. The dendrograms next to the heat maps were created using Ward hierarchical clustering with Euclidean distance.

Nc vs. CAI scatterplots

The codon bias statistics for each gene were calculated as described above and the results were imported into R to produce the Nc vs. CAI scatterplots. We then applied locally estimated scatterplot smoothing (LOESS) to identify the major trends in the scatterplots (Cleveland and Devlin 1988). The span parameter for the LOESS regression line was determined by generalized cross-validation (criterion = gcv, family = symmetric) using the loess.as function in the R package IANCOVA.

Melting temperature metagene profile: Because the transcription start sites have not been identified in D. erecta, D. mojavensis, and D. grimshawi gene annotations, we used the coding span (i.e., from start codon to stop codon, including introns) and the 2 kb upstream and downstream of the coding spans as a first approximation for this analysis. The melting temperatures were determined by the dan tool in the EMBASS package using a sliding window of 9 bp (windowsize = 9) and a step size of 1 (shiftincrement = 1) with the following parameters: dnaconc = 50, saltconc = 50, minterm = 55. The results were converted into BigWig format (Kent et al. 2010) for subsequent analysis.

Melting temperatures for the coding spans were normalized to 3 kb using bigWigSummary (part of the Kent source utilities). Melting temperatures for the normalized 3 kb region and the 2 kb flanking regions were imported into R and the standard graphics plot function in R was used to produce the metagene profiles.

Distance–Distance plots of gene characteristics

To determine whether any subset of F element genes has characteristics that differ from those of the group of genes as a whole, we constructed Distance–Distance plots for each F element separately using the rcorv package in R. Eight characteristics of the most comprehensive isoform of each gene were used in this analysis: coding span (bp from start to stop codon, including introns); andron repeat size (total size of all transposon fragments within introns); size of coding regions (sum of all coding exons in bp); number of coding exons; median size (in bp) of coding exons; median size (in bp) of introns; and Nc and CAI (calculated as described previously).

Using these eight gene characteristics, we calculated the classical Mahalanobis distance (MD) for each gene. MD measures the difference between the characteristics of each gene and the centroid (which is derived from the multivariate distribution of the characteristics of all F element genes). Unlike Euclidean distances, MD accounts for the variance of each gene characteristic and the covariance among the eight gene characteristics. The magnitude of MD corresponds to the dissimilarity of the characteristics of each gene compared to the centroid (i.e., large MD indicates that the gene has very different characteristics compared to the rest of the genes in the dataset).

However, because MD is sensitive to extreme outliers, we also calculated the robust Mahalanobis distance (RD) using the Stahel-Donoho estimator (sde). This robust estimator mitigates the impact of outliers on MD by assigning a weight to each gene based on its outlyingness (calculated using projection pursuit; Van Aelst et al. 2012). Hence a scatterplot of MD vs. RD (i.e., Distance–Distance plot) can be used to identify additional outliers that were masked by classical MD.

To create the Distance–Distance plots, the gene characteristics were normalized using the scale function in R because the different variables have values that differ by orders of magnitude (e.g., gene span vs. CAI).
The CovRobust function in the rrcov package was used to calculate the robust distances (with the parameter "sde"). Plots of the RD vs. the MD were produced using the generic plot command in R (with the parameter “which=’dd’”). Points were considered to be outliers if their values were greater than the square root of the 97.5% quantile of the $\chi^2$ distribution with 8 degrees of freedom (i.e., 4.19).

### Whole-genome alignments

To facilitate analysis of the wanderer genes (genes present on the F element in one species and on another Muller element in a different species), we produced a set of whole-genome alignments for D. melanogaster, D. yakuba, D. erecta, D. mojavensis, D. virilis, and D. grimshawi. (The Chain and Net alignments are available on the GEP UCSC Genome Browser, http://gander.wustl.edu.) Repeats in each genome were soft masked and the genome assemblies were aligned against each other using LAST (Kidbasa et al. 2011) with default parameters followed by the UCSC Chaining and Netting protocol (Kent et al. 2003).

### RESULTS

**Improved F and D element assemblies and gene annotations**

**Sequence improvement:** Previous studies have shown that the Drosophila F elements have a greater repeat density than the other autosomes (Leung et al. 2010), which could lead to a greater frequency of gaps and misassemblies. These assembly issues could introduce substantial bias into the analysis of genome characteristics (Salzberg and Yorke 2005). Quality assessments (see File S1) of the CAF1 assemblies (Drosophila 12 Genomes Consortium et al. 2007) led us to improve the D. mojavensis F element, the D. grimshawi F element, and the D. mojavensis euchromatic reference region from the D element to a quality standard that is similar to those used for the mouse genome project. As part of this sequence improvement standard, we resolved inconsistent mate pairs within each assembly and confirmed each assembly using restriction digests (see the section Materials and Methods for details). These experimental data provided additional confirmation of the accuracy of the F element assemblies, and enabled us to perform genomic analysis of the F elements with high confidence, ensuring accuracy (in particular) in the repeat and gene movement analyses.

Collectively, sequence improvement of the D. mojavensis and D. grimshawi analysis regions covered a total of approximately 3.8 Mb (1.7 Mb from the D. mojavensis F element, 1.1 Mb from the D. grimshawi F element, and 1.0 Mb from the D. mojavensis D element), closing 72 of 86 gaps and adding a total of 44,468 bases (Table S2A). Alignments between the CAF1 and the improved regions identified a total of 309 changes; 127 (41.1%) of these changes are single base substitutions, insertions, or deletions, while the remaining changes are more substantial (Table S2B). Detailed alignments between the CAF1 and the improved regions are available through the “D. mojavensis CAF1 Difference” and “D. grimshawi CAF1 Difference” tracks on the GEP UCSC Genome Browser (http://gander.wustl.edu).

An example of the improvement achieved is shown for the region surrounding the GLEAN-R annotation GI14058-PA (a putative ortholog of the D. melanogaster unc-13 gene) in D. mojavensis; this illustrates how the improved assemblies enabled us to produce more accurate gene models for the D. mojavensis F element (Figure 2).

**Manual gene annotations:** We also constructed manually curated gene models, including all isoforms, for each of the analysis regions. Because of the large evolutionary distance among D. melanogaster, D. mojavensis, and D. grimshawi and the limited expression data available, this analysis only focuses on the coding regions of genes. (See the section Materials and Methods and File S1 for detailed description of the annotation protocol.) The manual annotation process also allows us to identify potential annotation errors in D. melanogaster (e.g., rdgC as described in File S1).

Collectively, we annotated a total of 878 genes (1619 isoforms). A summary of the changes in the number of isoforms and coding exons, as well as descriptions of other noncanonical features (e.g., novel GC donor sites) compared with D. melanogaster (release 5.50) is available in File S2. Overall, 58% (552/947) of the GLEAN-R gene predictions match our annotation of the most comprehensive isoform (i.e., the isoform with the largest coding region, Table S3A), and 85% (3648/4287) of the coding exons predicted by GLEAN-R match the coding exons in the most comprehensive isoform (Table S3B).

Although a similar percentage of the coding exons predicted by GLEAN-R match our annotations in both the F and D elements (80.7–82.8%), a substantially lower percentage of the GLEAN-R gene models match our annotations on the D. mojavensis and D. grimshawi F elements (32.1% and 39.1%, respectively) than on the D elements (57.6% and 58.0%, respectively). Many of the differences between the GLEAN-R predictions and our annotations on the D. mojavensis and D. grimshawi F elements can be traced to improvement of the underlying sequence (e.g., unc-13 in Figure 2). Hence, the lower percentage of GLEAN-R gene models that match our annotations can primarily be attributed to the higher rate of assembly problems in the CAF1 assemblies for the D. mojavensis and D. grimshawi F elements. Our results show that manual sequence improvement and gene annotation can improve over half of the gene models in regions with high repeat density.

**F elements consistently show high repeat density but vary in repeat composition**

The most striking difference between the D. melanogaster F element and the other autosomes is its high density of repeats, primarily remnants of transposable elements (Bergman et al. 2006; Riddle et al. 2009). To obtain an overview of the repetitive element landscape of F elements in the four Drosophila species, we analyzed the types and distribution of repeats using four different approaches: WindowMasker, tandem, Tandem Repeats Finder, and RepeatMasker with species-specific transposon libraries (Figure 3). (Detailed repeat statistics are available in File S3 and File S4.)

**WindowMasker analysis shows the F elements have high repeat density:** To obtain an overview of the total repeat content, we tabulated the total number of bases masked by WindowMasker for each of the analysis regions. Unlike other repeat finding tools, WindowMasker relies only on the genomic sequence to identify over-represented sequences that correspond to low complexity sequences, simple repeats, or transposable elements, which makes it an ideal tool for analyzing the repeat contents of genomes without comprehensive repeat libraries (Morgulis et al. 2006). The results show that F elements consistently exhibit higher repeat densities than their corresponding euchromatic reference regions (D elements) in all four species (Figure 3A). D. mojavensis and D. grimshawi have higher repeat densities than D. melanogaster and D. erecta in both the F elements and the D elements. In fact, the D. mojavensis and D. grimshawi D elements have repeat densities that are similar to those of the D. melanogaster and D. erecta F elements.

To better understand the composition of the repeats identified by WindowMasker, we used Tallymer (Kurtz et al. 2008) to analyze the frequency of short sequences (i.e., words) in each analysis region. A more repetitive region requires a larger word size in order to achieve...
the same percentage of words that are unique compared with a less repetitive region (Chor et al. 2009). Tallymer analysis shows that approximately 95% of the 13-mers (i.e., sequences with a length of 13) are unique in the euchromatic reference regions (Table S4). In congruence with the WindowMasker results, which show that the D. mojavensis F element has the highest repeat density, we find that more 13-mers appear at a greater frequency on the D. mojavensis F element than in the other analysis regions. In contrast, most of the 13-mers at the base of the D. melanogaster and D. erecta D elements occur at low frequencies. The Tallymer analysis also shows that the D. grimshawi F and D elements have the most similar distributions of 13-mers (i.e., the most similar repeat density) among the four species (Figure 4A).

Examination of the 13-mers identified by Tallymer shows that many of the 13-mers that appear at a high frequency in D. mojavensis and D. grimshawi contain AT and CA dinucleotide repeats. Analyses of the distribution of dinucleotide repeats show that CA dinucleotide repeats are shorter on the D. melanogaster and D. erecta F elements, but longer on the D. mojavensis and D. grimshawi F elements, than in the euchromatic reference regions (Figure 4B). Thus, while low density of CA repeats was previously associated with the F element in D. melanogaster (Pardue et al. 1987), this does not seem to hold in general. The D. mojavensis and D. grimshawi F elements are also enriched in AT dinucleotide repeats compared with those of D. melanogaster and D. erecta. The lack of CG repeats in both the F and D elements is also striking (see the Discussion section).

Simple and low complexity repeats are particularly abundant on the D. grimshawi F element: The Tantan analysis (Frith 2011) shows that D. mojavensis and D. grimshawi have a greater density of simple and low complexity sequences in both the F element and the euchromatic reference regions compared with the corresponding regions in D. melanogaster and D. erecta (Figure 3B). The analysis also reveals some species-specific differences: simple and low complexity repeats appear to contribute the most to the repeat density of the D. grimshawi genome. The D. grimshawi F element has a substantially greater density of simple and low complexity repeats (18%) compared with the F elements of the other species examined (7–11%). In contrast to the other species, the D. mojavensis F element shows a lower density of simple and low complexity repeats compared to its euchromatic reference region (11% vs. 14%).

Tandem repeats show a skewed distribution on the D. erecta D element: Tandem repeats may play a particular role in genome rearrangement and regulation of gene expression (Sinha and Siggia 2005; Farré et al. 2011). For this analysis, tandem repeats are defined as regions with a minimum size of 25 bases and a maximum period of 2000 (see the section Materials and Methods for the complete list of search parameters). Results from Tandem Repeats Finder (Benson 1999) show that the D. mojavensis and D. grimshawi F elements and their euchromatic reference regions have a higher density of tandem repeats than the corresponding regions in D. melanogaster and D. erecta (Figure 3C). Although the base and the extended regions of the D. erecta D element both show a low density of tandem repeats, the analysis region near the telomere shows a high density, as do the euchromatic reference regions in D. mojavensis and D. grimshawi. A skew to a greater density of tandem repeats toward the telomere is apparent in a sliding window analysis of the D. erecta D element as a whole. In contrast, the D. melanogaster D element does not show the same skew in the density of tandem repeats (Figure S2).

Recent expansion of DINE-1 transposons leads to high transposon density on the D. mojavensis F element: Transposons may play an important role in targeting heterochromatin formation (Grewal and Elgin 2007). Because many transposons are species-specific, we constructed transposon libraries for each species and then used RepeatMasker (Smit et al. 1996) to identify transposon remnants in each analysis region. (See File S1 for the protocols used to construct and classify the species-specific transposon libraries, and File S4 for transposon density estimates using different species-specific transposon libraries.) Among the F elements, D. mojavensis has the highest transposon density (~50%) whereas D. grimshawi has the lowest (~20%). Strikingly, ~53% of the transposon fragments on the D. mojavensis F element show sequence similarity to DINE-1 elements.

The RepeatMasker results are generally in concordance with the WindowMasker results (Figure 3D): F elements have a greater transposon density compared with the euchromatic reference regions (D...
In some cases the transposon density estimate is higher than the total repeat density estimate by WindowMasker (e.g., \textit{D. mojavensis} \textit{F} element). This discrepancy is primarily caused by the difficulty associated with precisely defining the boundaries of each repeat copy (Bao and Eddy 2002).

Although the WindowMasker analysis (Figure 3A) shows that the \textit{D. grimshawi} and \textit{D. mojavensis} \textit{F} elements have a similar repeat density (38% and 44%, respectively), the RepeatMasker analysis (Figure 3D) shows that the \textit{D. grimshawi} \textit{F} element has a much lower density of transposons than the \textit{D. mojavensis} \textit{F} element (20% and 50%, respectively). This difference can primarily be attributed to the density of DINE-1 elements (2% in \textit{D. grimshawi} vs. 27% in \textit{D. mojavensis}) and DNA transposons (5% vs. 12%). In particular, DINE-1 (a helitron) accounts for 53% of the \textit{D. mojavensis} \textit{F} element transposon fragments but only 8% of the transposon fragments on the \textit{D. grimshawi} \textit{F} element (Figure S3). DINE-1 elements account for approximately half of all transposon fragments on the \textit{D. melanogaster} and \textit{D. erecta} \textit{F} elements (46% and 45%, respectively). The high level of DINE-1 in \textit{D. mojavensis} suggests a recent expansion.

To ensure that the low transposon density found on the \textit{D. grimshawi} \textit{F} element is not an artifact of misassemblies in the CAF1 genome assembly (see File S1), we performed an additional repeat analysis using the species-specific ReAS libraries previously produced by the \textit{Drosophila} 12 Genomes Consortium (\textit{Drosophila} 12 Genomes Consortium et al. 2007). ReAS is less susceptible to the effects of misassemblies compared with alignment-based \textit{de novo} repeat finders because it identifies repeats by finding overrepresented sequences within genomic reads (Li et al. 2005). This analysis did not alter the conclusion that the \textit{D. grimshawi} \textit{F} element has the lowest transposon density among the species analyzed here (Figure S4).

**Multiple subfamilies of the DINE-1 element are observed:** The RepeatMasker results show that most of the differences in the transposon density of the \textit{F} elements can be attributed to the DINE-1 element (Figure 3D). Comparison of the DINE-1 fragments identified by RepeatMasker using the species-specific libraries vs. the RepBase \textit{Drosophila} library (Jurka et al. 2005) shows that there are additional DINE-1 elements in the \textit{D. grimshawi}, \textit{D. mojavensis}, and \textit{D. erecta} species-specific transposon libraries that are not in the \textit{Drosophila} RepBase library. Analysis of the distribution of the DINE-1 elements shows that 40% of the DINE-1 fragments (based on total size) on the \textit{D. grimshawi} \textit{F} and \textit{D} elements, and 29% on the \textit{D. mojavensis} \textit{D} element found by the species-specific repeat libraries do not overlap with repeats in the \textit{Drosophila} RepBase library. In contrast, although the \textit{D. mojavensis} \textit{F} element appears to have an expanded number of DINE-1 elements, only 9% do not overlap with repeats in the \textit{Drosophila} RepBase library (Table S5 and File S5).

Analysis of the scaffolds assembled from unmapped \textit{D. mojavensis} mod\textsc{ENCODE} RNA-Seq reads suggests that some of these helitrons are being transcribed in the \textit{D. mojavensis} genome; a potential
candidate is shown in Figure S5. (See File S1 for a more detailed description of this analysis.)

**Overall repeat distribution on the F element:** Collectively, the repeat analysis shows the F elements have a higher repeat density than the euchromatic reference regions in all four Drosophila species. It also shows that although the D. mojavensis and D. grimshawi F elements have similar total repeat densities, they have strikingly different repeat compositions. A total of 75% of the repeats that overlap with a repeat identified by WindowMasker on the D. mojavensis F element are transposons (particularly DINE-1 elements) compared to only 27% on the D. grimshawi F element, whereas the D. grimshawi F element shows a greater density of simple and low complexity repeats than the D. mojavensis F element (39% vs. 20%). These differences in repeat composition could impact the local chromatin structure and thus the evolution of the resident genes.

**Evolution of F element genes**
Despite its high repeat density, the distal arm of the D. melanogaster F element contains 79 genes, many of which have important developmental and housekeeping functions (Riddle et al. 2012). Our manual gene annotations (described previously) show that the D. melanogaster, D. erecta, D. mojavensis, and D. grimshawi F elements all have approximately 80 genes. The gene density of the F element is lower than that of the euchromatic reference regions from the D element (~60 genes/Mb vs. ~80 genes/Mb) for these four species (Table S6). Among the four species, the D. mojavensis F element has the lowest gene density (48 genes/Mb compared with 60–66 genes/Mb in the other F elements).

This reflects the increased size of the D. mojavensis F element due to the expansion of repetitive elements (1.7 Mb vs. 1.2–1.3 Mb in the other F elements) (Table S6 and Figure 3).

Although we have produced annotations for all isoforms, our analysis below is based only on the isoform with the largest coding region (i.e., the most comprehensive isoform) for each gene. Restricting our analysis to the most comprehensive isoform allows us to avoid counting the same region multiple times because of alternative splicing. We initially examined genes at the base, extended, and telomeric regions (described previously) of the D. erecta D element. Since the genes in these three euchromatic regions exhibit similar characteristics, the primary focus of the following analysis is on the comparison of genes between the F element and the base of the D element (results for all of the analysis regions are available in Figure S6). Summary statistics for all of the gene characteristics, and results of multiple comparison tests after the Kruskal-Wallis (KW) rank sum tests (Kruskal and Wallis 1952), are available in File S6.

**F element genes are larger because they have larger introns and more coding exons:** Comparisons of the distribution of gene characteristics using violin plots (Hintze and Nelson 1998) show that the coding span (i.e., the region that spans from the start codon to the stop codon, including introns) for F element genes is much larger (median 5156–7569 bp) than for genes at the base of the D elements (median 1028–1736 bp) (Figure 5, top center). The KW test shows that this difference is statistically significant (p-value: 2.12E-48).

Part of this difference in the coding span can be attributed to the significantly higher transposon density (KW test p-value: 2.40E-82) within the introns of F element genes (Figure 5, top center; “repeat size” is the total size of the transposon fragments within the introns of a gene, in bp). Among the four species analyzed in this study, 71–83% of the F element genes contain at least one transposon fragment in an intron. In contrast, only 20–46% of the D element genes contain at least one transposon fragment. Consistent with the results of the
transposon density analysis, we find that the D. mojavensis F element has the highest intron transposon density (median 1930 bp) whereas D. grimshawi has the lowest (median 210 bp).

In addition to differences in the repeat sizes within introns, the violin plots also show that the coding regions (i.e., the region that spans from the start codon to the stop codon, excluding introns) of F element genes are significantly larger (median 2313–2565 bp) than the coding regions for D element genes (median 918–1305 bp) (Figure 5, top right). The KW test shows that this difference in the size of the coding regions is statistically significant (p-value = 7.03E-33). Furthermore, although the actual genes found at the base of the D elements of D. mojavensis and D. grimshawi differ from those found at the base of the D. melanogaster and D. erecta D elements (due to various rearrangements), a multiple comparison test after KW shows no significant difference in the size of the coding regions.

To further analyze the difference in the distribution of coding spans and the coding regions between the genes on the F and D elements, we examined the distributions of the number of exons, the coding exon sizes, and intron sizes. Previous analysis has shown that D. melanogaster F element genes have more transcribed exons than genes in other domains (Riddle et al. 2012). In congruence with this observation in D. melanogaster, our analysis shows that F element genes in the four Drosophila species have significantly more coding exons (median 6–7) than D element genes (median 2–3) (KW test p-value = 5.59E-50) (Figure 5, bottom left). In contrast, the distributions of coding exon sizes are similar between F element genes (median 196–201.5 bp) and D element genes (median 195–284.5 bp). A KW test indicates that there is a significant difference in the distribution of coding exon sizes (p-value = 2.12E-07). However, multiple comparison tests show that only the differences between the coding exons of all four F elements and the coding exons from the base of the D. melanogaster and D. erecta D elements (due to various rearrangements), a multiple comparison test after KW shows no significant difference in the size of coding regions.

Consistent with the greater transposon density on the F element, we find that F element genes generally have significantly larger introns (median 172.5–228 bp) than D element genes (median 65–84 bp) (Figure 5, bottom right; KW test p-value = 6.14E-62). Multiple comparison tests show that D. grimshawi is the exception, as the difference in intron sizes between the D. grimshawi F and D element genes is not statistically significant. The intron size distribution for the D. grimshawi D element is significantly different from that of the other D elements, but is not significantly different from that of the D. melanogaster and D. erecta F elements. These observations are in concordance with the results of the transposon density analysis, which shows that the D. grimshawi F and D elements have more similar transposon densities compared to those of other species (see Figure 3D).

Hence the larger coding spans observed for F element genes (Figure 5, top left) can primarily be attributed to a combination of significantly larger repeat sizes within introns (Figure 5, top center) and larger coding regions (Figure 5, top right). The larger coding regions of F element genes can be attributed to a significantly higher number of coding exons (Figure 5, bottom left) but not to the size of the individual coding exons (Figure 5, bottom center). Introns of F element genes are significantly larger than introns of genes in the euchromatic reference regions for D. melanogaster, D. erecta, and D. mojavensis but not for D. grimshawi (Figure 5, bottom right).

**F element genes show lower codon bias than D element genes:** Previous analysis of codon usage bias in 12 Drosophila species (using 33 D. melanogaster F element genes and their corresponding GLEAN-R annotations) showed that F element genes exhibit lower codon bias compared with genes on the other Muller elements (Vicario et al. 2007). Here we expand the codon bias analysis to all of the manually annotated F element genes in four Drosophila species using two metrics: the Effective Number of Codons (Nc), which measures deviations from uniform codon usage (Wright 1990), and the CAI, which measures deviations from the species-specific optimal codon usage (Sharp

**Figure 5** Violin plots of gene characteristics for each analysis region. A violin plot is composed of a boxplot and a kernel density plot: the black dot denotes the median; the darker regions and the thin white box denote the range between the first (Q1) and third (Q3) quartiles i.e., the interquartile range (IQR); Whiskers extending from the white box span from Q1–1.5×IQR to Q3+1.5×IQR; the data points beyond the whiskers are outliers. For violin plots using a log scale, a pseudocount of one was added to all data points. The larger coding spans of F element genes can be attributed not only to larger introns (often containing repeats), but also to larger coding regions. The larger coding regions reflect the higher number of coding exons.
Codon bias in D. grimshawi F element genes can primarily be attributed to selection: To infer the selective pressure experienced by genes in the different analysis regions, we compared the Nc and CAI values of each gene using a scatterplot (Vicario et al. 2007). This analysis posits that Nc measures deviations from uniform codon usage that could either be attributed to mutational bias or selection, while CAI measures deviations from optimal codon usage and primarily reflects selection. Hence, genes that exhibit both large deviations from uniform codon usage (i.e., low Nc) and small deviations from optimal codon usage (i.e., high CAI) are thought to be under stronger selective pressure, while genes with low Nc and low CAI are under stronger influence from mutational biases (Vicario et al. 2007). After constructing the Nc vs. CAI scatterplots for each analysis region, we applied locally estimated scatterplot smoothing (LOESS, Cleveland and Devlin 1988)) to capture the overall trends seen in each scatterplot (Figure 6C). Regression lines that show a positive slope indicate that the codon bias can primarily be attributed to mutational biases, while a negative slope indicates that the codon bias can primarily be attributed to selection on codon usage.

Consistent with previous reports using a smaller gene set (Vicario et al. 2007), our analysis shows that codon bias for most of the genes on the D. melanogaster and D. erecta F elements can be attributed to mutational biases rather than selection (i.e., most of the genes are in the part of the LOESS regression line that shows a positive slope), indicating low selective pressure relative to what is seen for the D element genes. In contrast, we find that codon bias for most of the genes on the D. grimshawi F element, along with genes on the D elements, can primarily be attributed to selection (i.e., most of the genes are in the part of the LOESS regression line with negative slope). Thus we observe that the F element with the lowest transposon density (D. grimshawi) differs from the other F elements in this regard, with more of the genes showing evidence of response to selective pressure. We also find that most of the D. mojavensis F element genes have CAI values that are higher than those for a gene with equal codon usage (dotted line in Figure 6C), indicating a more optimal pattern of codon usage compared to F element genes in D. melanogaster and D. erecta. Although most of the F element genes within each Nc vs. CAI scatterplot follow a similar trend, there are a few outliers (Figure 6C). For example, the Muller F element genes ATPsyn-beta and Rps3A exhibit low Nc and high CAI in all four Drosophila species (Figure S7, see Discussion) (See File S1 and the heat maps in File S7 for the detailed analysis on the changes in codon usage preferences for each amino acid).

A subset of F element genes exhibits distinct characteristics in all four species: Our analyses show that the overall characteristics of F element genes are distinct from genes at the base of the D element. However, previous studies have shown that some regions on the D. melanogaster F element differ from the general case in being enriched in H3K27me3, rather than H3K9me2/3, in a tissue-specific fashion; genes that reside in these regions are associated with Polycomb (PcG)
(Kharchenko et al. 2011; Riddle et al. 2012). PcG proteins regulate the expression of many genes involved in development (such as homeotic genes) by altering the chromatin structure (reviewed in (Lanzuolo and Orlando 2012)). Hence it is of particular interest to ask whether the six F element genes associated with PcG exhibit characteristics that differ from the rest of the F element genes.

Because there are only six genes on the D. melanogaster F element that are associated with PcG, there is insufficient statistical power to analyze each gene characteristic separately to ascertain if PcG genes exhibit significantly different properties compared to the other F element genes. Consequently, we performed a multivariate analysis of the gene characteristics described above (see File S1 for details). For each F element, we constructed a Distance–Distance (DD) plot (Rousseeuw and Van Zomeren 1991) of gene characteristics to identify outliers (Figure 7). Detection of outliers using MDs (Mahalanobis 1936) show that there are three F element genes (bt, fd102C, and Sox102F) that consistently exhibit characteristics that are distinct from other F element genes in all four species. The bt gene, for example, is an outlier because it has a substantially larger coding span, larger coding region, and more coding exons compared to the other F element genes in all four species. The DD plot also identifies some species-specific outliers: CG31999 is an outlier in the D. mojavensis F element because it has a gene size of 157 kb (compared to 10 kb in D. melanogaster).

Detection of outliers using robust distances identifies additional outliers (triangles in the top left quadrant, Figure 7) that are not detected by the MD because of the masking effect (Ben-Gal 2005). Robust distance in the DD plots identifies 25–29 F element genes as outliers and 14 of these outliers are found in all four species. Analysis of these 14 genes using modMINE (Contrino et al. 2012) shows that they are significantly enriched in “RNA polymerase II distal enhancer sequence specific DNA binding transcription factor activity” (GO:0003705, Holm-Bonferroni adjusted p-value = 8.36E-4).

Of the 14 outliers that are found in all four species, five of them (ey, fd102C, Sox102F, sv, and toy) are associated with PcG domains. The only exception is zfh2, which is an outlier in three of the four species (D. melanogaster, D. mojavensis, and D. grimshawi). Hence the DD plot analysis suggests that F element genes that reside in domains enriched in H3K27me3 might have different characteristics than F element genes that reside in domains enriched in H3K9me2/3.

**F element genes show lower melting temperature metagene profiles**

Despite residing in a domain with heterochromatic properties, D. melanogaster F element genes exhibit expression levels that are similar to those of other euchromatic genes (Riddle et al. 2012). One of the mechanisms for regulating gene expression is the pausing of RNA Polymerase II during early elongation (reviewed in (Adelman and Lis 2012)). Previous analysis has shown that the efficacy of elongation depends on the stability of the 9-bp RNA-DNA hybrid in the elongation complex (Tadigotla et al. 2006). Genes that exhibit polymerase pausing have a distinct 9 bp melting temperature profile (i.e., greatest melting temperature at 25–30 bp downstream of the transcription start site, where pausing occurs) (Nechaev et al. 2010).

Previous studies have shown that D. melanogaster F element genes exhibit lower melting temperatures than genes that reside in other domains (Riddle et al. 2012). To ascertain whether this difference is conserved in other Drosophila species, we performed a metagene analysis of the melting temperature profile. (See the section Materials and Methods for details on the definition of the metagene.)

The metagene profiles show that F element genes in all four Drosophila species have lower melting temperatures (T_m) than genes at the base of the D element. In all cases, the coding spans (i.e., from start codon to stop codon, including introns) show substantially higher melting temperatures than the 2-kb flanking regions (Figure 8). Coding spans of the D. mojavensis and D. grimshawi F elements show greater T_m than those of D. melanogaster and D. erecta. Comparing the F element and D element genes within a given species, we find that those of D. grimshawi show the smallest difference in the melting temperature profiles.

**Figure 7** Distance–Distance Plots of robust distance (RD) vs. Mahalanobis distance (MD) show both common and species-specific outliers. The horizontal and vertical lines correspond to the cutoff values for outliers (97.5% quantile of the x² distribution, see File S1). Values greater than the cutoff values identify outliers. Triangles in the upper right quadrant are outliers based on both RD and MD. Triangles in the upper left quadrant are outliers only based on RD. The dashed line corresponds to points with equal RD and MD values. F element genes that reside in a Polycomb domain in D. melanogaster are highlighted in gray.
F element gene rearrangements and gene movements

Changes in F element gene order: Previous studies have estimated that approximately 95% of the genes in D. melanogaster remain on the same Muller element across the 12 Drosophila species (Bhutkar et al. 2008). To ascertain whether the low rate of recombination would affect the rate of rearrangements and gene movements on the F element, we analyzed the placement of D. melanogaster F element genes in the other Drosophila species.

Of the 79 D. melanogaster F element genes annotated by FlyBase, two of the genes were omitted from the gene movement analysis because they are either a partial gene (JYalpha) or a possible misannotation (CG11231). (See File S1 for details.) Of the remaining 77 D. melanogaster F element genes, all 77 genes (100.0%) are found on the D. erecta F element, 72 (93.5%) are found on the D. mojavensis F element and 73 (94.8%) are found on the D. grimshawi F element.

Except for CG11231, the D. erecta F element is completely syntenic with respect to the D. melanogaster F element. GRIMM (Tesler 2002) estimates that a minimum of 31 inversions are required to transform the D. melanogaster F element gene order and orientation to that observed in the D. mojavensis F element (72 genes in common). Similarly, at least 33 inversions are required to transform the D. melanogaster F element gene order to that observed in D. grimshawi (73 genes in common). There are 78 genes that are found on both the D. mojavensis and D. grimshawi F elements, and GRIMM estimates a minimum of seven inversions are required to transform the gene order in D. mojavensis to that observed in D. grimshawi. (See possible rearrangement scenarios estimated by GRIMM in Figure S8.)

Analysis of the number of genes per syntenic block (i.e., syntenic block sizes) shows that the F elements have smaller syntenic blocks than the previously reported genome averages (Bhutkar et al. 2008). The D. mojavensis F element has an average syntenic block size of 3.4 genes compared to an average of 8.8 genes per syntenic block for the whole genome. The corresponding numbers for D. grimshawi are 3.6 and 8.4 genes per syntenic block for the F and D elements, respectively. Thus inversions are common on the F element despite its low rate of recombination.

Identifying a wanderer gene hotspot: Movement of genes between different chromosomes typically results from gene duplications (via ectopic recombination or retrotransposition) followed by the loss of the original copy of the gene (Meisel et al. 2009). There are 12 genes that are found on the F element in one Drosophila species, but on another Muller element in a different Drosophila species (“wanderer genes”: Figure 9A). One of these wanderer genes is a putative paralog of Cyp1 (Cyp1_alpha) that is found on the D. mojavensis F element and the D. grimshawi B element but is not found in either D. melanogaster or D. erecta.

To further analyze the distribution of wanderer genes on the F elements, we compared the genome assemblies of six Drosophila species (D. melanogaster, D. yakuba, D. erecta, D. virilis, D. mojavensis, and D. grimshawi) using the UCSC Chain and Net protocol (Kent et al. 2003). Examination of the Net alignment tracks shows there is a single region (i.e., hotspot) in both the D. mojavensis and D. grimshawi F elements where most of the wanderer genes are found (Figure S9). The D. mojavensis F element hotspot contains five of the six wanderer genes relative to D. melanogaster (Figure 9B, top). The hotspot on the D. grimshawi F element contains three of the four wanderer genes relative to D. melanogaster and one of the wanderer genes (yellow-h) relative to D. mojavensis (Figure 9B, bottom).

Because three of the wanderer genes (CG5262, rho-5, and CG4038) are found in the wanderer gene hotspots of both the D. mojavensis and D. grimshawi F elements (relative to D. melanogaster), we can use them to infer the direction of gene movements of the rest of the wanderer genes in the hotspot. The yellow-h gene likely moved from the F element to the A element in D. mojavensis. In contrast, both the PRY and Or13a genes likely moved from other chromosomes (the Y chromosome and the A element, respectively) to the D. mojavensis F element. Hence our analysis indicates that gene movement occurs in both directions on the F element and that the cumulative effect of these gene movements is that there are a similar number of genes (~80) on the F element in all four species.

DISCUSSION

F elements exhibit distinct characteristics in Drosophila

The D. melanogaster F element is unusual in that it appears to be predominantly heterochromatic, but in the distal 1.3 Mb has a gene density similar to the euchromatic chromosome arms (Sun et al. 2004; Riddle et al. 2011, 2012). Immunofluorescent staining of the polytene...
chromosomes shows that the D. melanogaster, D. erecta, D. mojavensis, and D. grimshawi F elements are enriched in H3K9me2 (Figure 1), which suggests that the F element is generally packaged as heterochromatin in these four species. In order to elucidate the impact of these unusual characteristics on the evolution of the F element and its genes, we performed a comparative analysis of the F elements and euchromatic regions near the base of the D elements (coordinates listed in Table S1).

To increase the accuracy of our analysis, we improved the assemblies of the D. mojavensis and D. grimshawi F elements and the base of the D. mojavensis D element, closing 72 of 86 gaps and adding 44,468 bases to these assemblies (Figure 2 and Table S2). Restriction digests and consistent mate pairs provide strong experimental support for the final assemblies. We also produced gene annotations for the regions under study in D. erecta, D. mojavensis, and D. grimshawi (878 genes, 1619 isoforms). Each gene was annotated at least twice independently and reconciled by a third investigator, giving increased confidence in the results. We find substantial differences between our manually curated gene models and the GLEAN-R gene predictions, with only 32–58% of the GLEAN-R gene models showing complete congruence in the cases of D. mojavensis and D. grimshawi (Table S3). These results illustrate the benefits of manual sequence improvement and gene annotations for regions with moderate repeat density.

Our analysis shows that the F element has generally maintained its distinct characteristics compared with the autosomes in species that diverged from D. melanogaster 40–60 million years ago. Compared with the euchromatic reference regions within each species, we find that F elements have higher repeat density (Figure 3 and Figure 4), and the genes are larger, have larger introns, more coding exons (Figure 5), lower codon bias (Figure 6), and lower melting temperatures (Figure 8). Most F element genes exhibit similar characteristics within each species but there are also species-specific and common outliers among the four Drosophila species (Figure 7). Analysis of gene movements shows that the F elements have smaller syntenic blocks than the genome average and that there is a single hotspot in both the D. mojavensis and D. grimshawi F elements where most of the wanderer genes are found (Figure 9). We also identified genes that have moved both on and off of the F element, maintaining approximately the same number of genes in the four species. It is striking that these gene movements (presumably due to transposition) occur at a rate similar to that seen for the other autosomes, and inversions are more frequent, while recombination is reduced. This suggests that the frequency of such events is not dictated solely by DNA accessibility, as such a simple model of the consequences of heterochromatin packaging might have been thought to impact all three types of events equally.

Although the F elements generally show similar characteristics, we also find some differences among the four Drosophila species (particularly between the species in the Drosophila clade vs. the species in the melanogaster subgroup of the Sophophora clade). These differences could provide insights into the impact of low recombination rate on the evolution of the genomic landscape (e.g., repeats and gene characteristics) of the F element.

### F elements have different repeat compositions

One of the prominent characteristics of heterochromatin is its high repeat density. Previous studies have shown that the difference in total repeat density is one of the major contributors to the changes in genome size among the different Drosophila species (Bosco et al. 2007). A critical consideration here is that some classes of transposons and tandem repeats have been implicated in gene silencing and heterochromatin formation (Martienssen 2003; Riddle et al. 2008; Sentmanat and Elgin 2012).

In concordance with previous reports for many eukaryotes (Tóth et al. 2000), our dinucleotide repeat analysis shows a lack of CG dinucleotide repeats on both the F and D elements in all four Drosophila species (Figure 4B). Previous studies have shown that there is a strong mutational bias in Drosophila toward A/T, whereas codon bias tends to favor G/C at synonymous sites (Moriyama and Powell 1997; Vicario et al. 2007). Hence the lack of CG dinucleotide repeats on
the F element could be explained by its low recombination rate. However, this mutational bias does not explain the lack of CG dinucleotide repeats on the D elements. Previous studies have shown that methylated CpG sequences have a greater rate of mutation because they are susceptible to spontaneous deamination, and the low frequency of CG repeats has been attributed to this (Duncan and Miller 1980). Hence, the lack of CG dinucleotide repeats on the D element is striking given the low levels (if any) of DNA methylation in Drosophila (Raddatz et al. 2013; Takayama et al. 2014). Another explanation for the lack of CG repeats is clearly needed.

Previous in situ hybridization analyses by Pardue et al. (1987) have shown that CA/GT dinucleotide repeats are highly enriched on the D. melanogaster X chromosome but are depleted in the F element and \( \beta \)-heterochromatin (i.e., heterochromatin that is replicated during polytenization). In contrast, the D. viridis F element is enriched in CA/GT dinucleotide repeats (Pardue et al. 1987). Our analysis shows that, similar to D. viridis, the D. mojavensis and D. grimshawi F elements have long CA and AG dinucleotide repeats, whereas the D. melanogaster F element is notably depleted in these dinucleotide repeats (Figure 4B). However, the significance of these differences in the distribution of dinucleotide repeats is unclear.

Our analysis also shows that the D. mojavensis and D. grimshawi F elements contain longer AT dinucleotide repeats than D. melanogaster and D. erecta (Figure 4B). Previous analyses have shown that long AT dinucleotide repeats inhibit the formation of nucleosomes (reviewed in (Kapitonov and Jurka 2007)). Previous analysis of 12 Drosophila species shows that DINE-1 fragments often are found in introns or within 1 kb of the coding regions (Yang and Barbash 2008). Hence the DINE-1 element may play an important role in shaping the genomic landscape of the F elements and their genes.

The high repeat density of the F element has a direct impact on gene characteristics. One of the factors that contributes to the significantly larger coding span of F element genes compared to D element genes is that F element genes have significantly larger introns in all of the species examined here except for D. grimshawi (Figure 5, lower right). This difference in intron size can partly be attributed to the differences in intron repeat density (Figure 5, top center). However, this does not a priori explain the other factor contributing to the larger coding span of F element genes — the larger number of coding exons.

### The D. grimshawi F element genes exhibit different patterns of codon bias

A salient characteristic of the F element is its low rate of recombination (Arguello et al. 2010; Wang et al. 2002). Codon bias is correlated with the recombination rate because of the Hill-Robertson effect (Hill and Robertson 1966; Klisman and Hey 1993). In agreement with this effect, we find that F element genes exhibit lower codon bias than D element genes based on both the Nc and the CAI metric (Figure 6).

Although F element genes for all four species exhibit smaller deviations from uniform codon usage (i.e., low Nc) than D element genes, we find that D. mojavensis and D. grimshawi genes show a more optimal pattern of codon usage (i.e., greater CAI) than D. melanogaster and D. erecta genes in both the F and D elements. The greater CAIs in the D. mojavensis and D. grimshawi analysis regions are in congruence with the results from previous whole genome analysis of CAIs in 12 Drosophila species, which shows that the distribution of CAIs for species in the Drosophila subgroup are shifted to the right (i.e., greater CAI) compared with the melanogaster subgroup (Heger and Ponting 2007).

In concordance with the hypothesis that greater CAI reflects stronger selection because of greater transfer RNA abundance (Moriyama and Powell 1997) and greater expression levels (Duret and Mouchiroud 1999), we find that the F element genes \( \text{AtPSm-beta} \) and Rps3A exhibit strong codon bias in all four Drosophila species (Figure S7). \( \text{AtPSm-beta} \) is an ATPase (Peña and Garesse 1993) and Rps3A is a ribosomal protein (Van Beest et al. 1998). Both genes are very highly expressed in all developmental stages in D. melanogaster (Graveley et al. 2011).

In concordance with previous reports (Kuhn and Heslop-Harrison 2011), comparison of the overlap between the DINE-1 fragments identified by the species-specific transposon library and the Drosophila RepBase library indicates that there are at least two major subfamilies of DINE-1 elements in D. mojavensis (Table S5). We found that 67% of the DINE-1 fragments in the species-specific library overlap with the Hom06 transposon whereas 22% overlap with the Helitron1_Dmoj transposon (File S5). Analysis of the D. mojavensis RNA-Seq data (Graveley et al. 2011) identified a scaffold that contains a conserved Helitron_like_N (Pfam accession: PF14214) domain (Figure S5), indicating that some of the DINE-1 elements may still be active. A transposable element present at a high density, in a genome that expresses that transposable element, could well be a target for silencing, promoting heterochromatin formation.

The horizontal transfer and subsequent amplification of helitrons occur in many organisms, including mammals, reptiles, and insects (Thomas et al. 2010). Helitrons can capture adjacent gene fragments during transposition and can affect the evolution of the host species [reviewed in (Kapitonov and Jurka 2007)]. Previous analysis of 12 Drosophila species shows that DINE-1 fragments often are found in introns or within 1 kb of the coding regions (Yang and Barbash 2008). Hence the DINE-1 element may play an important role in shaping the genomic landscape of the F elements and their genes.
The results of the repeat density and codon bias analyses suggest that the *D. grimshawi* F element has a greater rate of recombination. This might be a consequence of the lower transposon density, given that transposons can be targets for heterochromatin formation (Lippman and Martienssen 2004; Sentmanat et al. 2013). Furthermore, the low density of DINE-1 elements on the *D. grimshawi* F element compared with the other species suggests that this transposon might play an important role in promoting heterochromatin assembly. However, the transposon families present vary in the different species, and there may well be other transposable elements, present in other species but absent from *D. grimshawi* [e.g., 1360 and *Galileo*, (Marzo et al. 2008)] that could contribute substantially to silencing.

**Lower melting temperatures may facilitate transcription of F element genes**

Previous analysis has shown that a much smaller fraction of the *D. melanogaster* F element genes exhibit polymerase pausing (1.6%) compared with genes found in pericentric heterochromatin (12.5%) or euchromatin (15.0%). F element genes also show a lower melting temperature near the transcription start site than genes in the other *D. melanogaster* Muller elements, irrespective of whether the genes exhibit polymerase pausing (Riddle et al. 2012). Our metagene analysis of melting temperature profiles shows that F element genes in all four *Drosophila* species exhibit lower melting temperatures across the entire span of the metagene than D element genes (Figure 8). The lower melting temperature suggests that, similar to *D. melanogaster*, only a small fraction of the *D. erecta*, *D. mojavensis*, and *D. grimshawi* F element genes will exhibit polymerase pausing.

The elongation rate of RNA Polymerase II can affect the total mRNA level (Danko et al. 2013) and previous studies have found that the rate of elongation is negatively correlated with GC content within the gene body, and with exon density (Jonkers et al. 2014). Although F element genes are larger and have more coding exons than euchromatic genes (Figure 5), the metagene has a substantially lower melting temperature (Figure 8), presumably because of the high AT content within introns and the low codon bias. The high AT content in the genes could be a consequence of the less effective selection for codon bias (because of the low rate of recombination on the F elements) coupled with the A/T mutational bias in *Drosophila*. The lower GC content within the gene body could facilitate transcription and hence help explain how F element genes can have expression levels that are similar to genes in euchromatic regions, despite residing in a domain with heterochromatic properties.

This study provides an initial survey of the evolution of the F element and its genes in four *Drosophila* species. Our results show that the F element has maintained its distinct characteristics in both the *Sophophora* and *Drosophila* subgenera. The unusual mixture of a heterochromatic domain with a euchromatin-like gene density on the F element enabled us to investigate a number of interesting questions relating genome organization to gene function. The genomic resources (e.g., improved assemblies, gene annotations, genome browsers) produced in this study provide a foundation for future investigations into the factors that impact chromatin packaging and gene expression in a heterochromatin domain.

**ACKNOWLEDGMENTS**

In several cases a class worked together to produce a high-quality annotation of a given project; we thank the following classes for their contributions: Amherst College, Biology 19: Molecules, Genes & Cells (Fall 2009, Fall 2010, and Fall 2011); California State University, Monterey Bay, Biology 361: Eukaryotic Molecular Biology (Fall 2011); College of William & Mary, BIOL404: Genomics & Functional Proteomics (Spring 2009); Georgetown University, Cell Biology 363 (Fall 2007, Fall 2008, Fall 2009, and Fall 2010); Johnson C. Smith University, BIO 433: Explorations in Genomics (Fall 2010); Lindenwood University, BIO 422: Biochemistry: Metabolism (Spring 2011); Loyola Marymount University, Molecular Biology 439 (Fall 2007, Fall 2008, Fall 2009, and Fall 2010); New Mexico Highlands University, Biol 499: Independent Study (Fall 2008, Spring 2009, and Fall 2009); New Mexico Highlands University, Biol 620: Advanced Topic in Biology (Summer 2010); Prairie View A&M University, Biol 4061-Research (Spring 2008, Spring 2009, Spring 2010, Spring 2011, and Spring 2012); San Francisco State University, Biol 638/738: Bioinformatics & Genome Annotation (Fall 2007, Fall 2008, Fall 2009, and Fall 2010); The City College of New York, Grove School of Engineering, Sci280: Bioinformatics & Biomolecular Systems (Spring 2009, Fall 2009, Spring 2010, Fall 2010, Spring 2011, Fall 2011, and Spring 2012); The City College of New York, Grove School of Engineering, Sci316: Genomics and Structural Bioinformatics (Summer 2009); The City College of New York, Biology 3132: Genomics and Bioinformatics (Spring 2010); University of Nebraska-Lincoln, BIOC498: Eukaryotic Bioinformatics Research (Spring 2008, Spring 2009, Spring 2010, and Spring 2011); University of the Incarnate Word, BIOL 3461: Genetics & Lab (Fall 2011, Spring 2012, and Summer 2012); University of West Florida, PCB4905: Genome Data Analysis Class (Spring 2008); Wilkes University, Biology 345: Genetics (Fall 2009 and Fall 2010); Worcester State University, BI 371: Molecular Biology (Fall 2009). The authors also thank additional students who contributed data analysis to this project, but for various reasons did not participate in reviewing the manuscript. These students were enrolled in one of the classes listed in File S8, which lists all participating classes. We thank The Genome Institute at the Washington University School of Medicine for generating the raw sequences reported here and for providing training and support in sequence improvement for many of the coauthors. We also thank the two anonymous reviewers for valuable comments and suggestions. This work was supported by grant #52007051 from the Howard Hughes Medical Institute (HHMI) Precollege and Undergraduate Science Education Professors Program to Washington University (for 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 support provided by the Genome Institute at the Washington University School of Medicine was funded by grant 2U54 HG00307910 from the National Human Genome Research Institute (Richard K. Wilson, principal investigator). The content of this article is solely the responsibility of the authors and does not necessarily represent the official views of HHMI, the National Human Genome Research Institute, the National Institute of General Medical Sciences, or the NIH.

**LITERATURE CITED**


Li, R., J. Ye, S. Li, J. Wang, Y. Han et al., 2005 ReAs: Recovery of ancestral sequences for transposable elements from the unassembled reads of a whole genome shotgun. PLOS Comput. Biol. 1: e43.


Present addresses

1Department of Biology, University of the Fraser Valley, Abbotsford, BC V2S 7M8, Canada
2Biology Department, Massasoit Community College, Brockton, MA 02302
3Department of Biology, Washington University in St. Louis, St. Louis, MO 63130
4College of Engineering & Science, University of Detroit Mercy, Detroit, MI 48221
5Department of Biology, University of San Diego, San Diego, CA 92110
6Department of Biology, University of the Cumberlands, Williamsburg, KY 40769
71 Cranberry Hill, Suite 403, Lexington, MA 02421
8School of Natural and Health Sciences, Mount Mary University, Milwaukee, WI 53222
9Department of Biology, The University of Alabama at Birmingham, Birmingham, AL 35294