G3 Genes | Genomes | Genetics

Drosophila Muller F elements maintain a distinct set of genomic properties over 40 million years of evolution

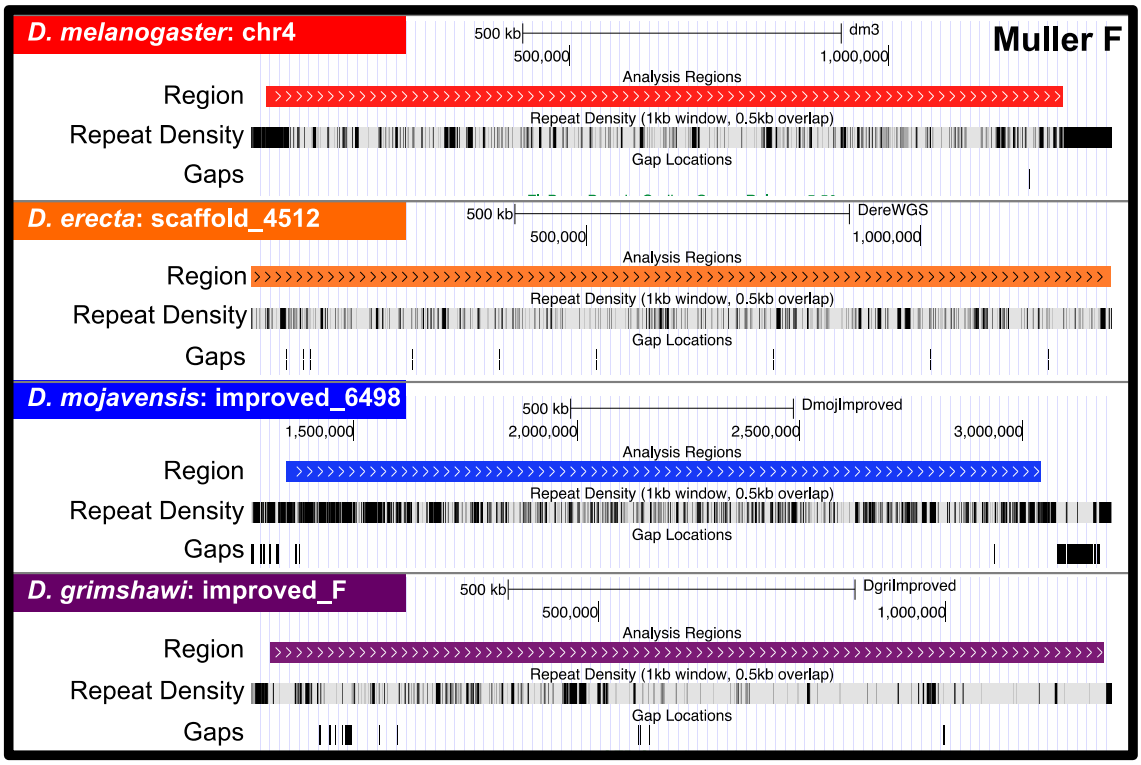
Wilson Leung[a], … Sarah C. R. Elgin[a,*]

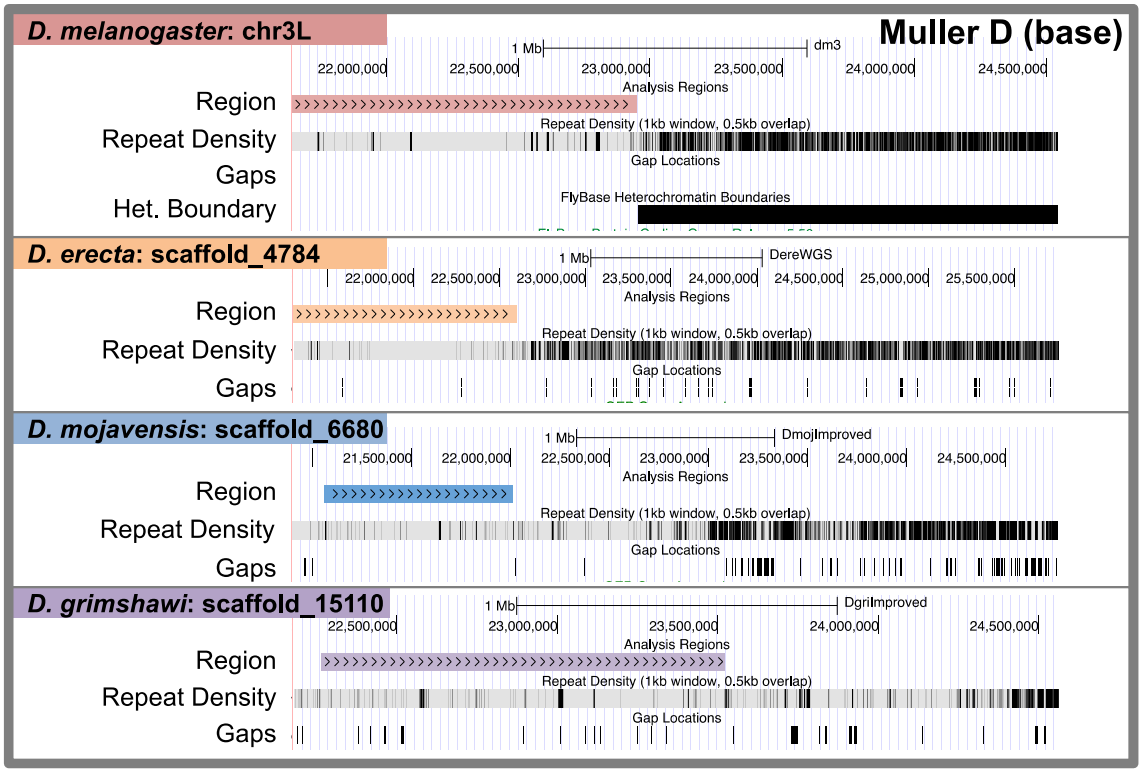
[a]Department of Biology, Washington University in St. Louis, St. Louis, MO 63130

[*]Corresponding author: Washington University in St. Louis, Campus Box 1037, One Brookings Drive, St. Louis, MO 63130-4899. Phone: (314) 935-5348. Email: selgin@wustl.edu

DOI: 10.1534/g3.114.015966

SUPPORTING FIGURES





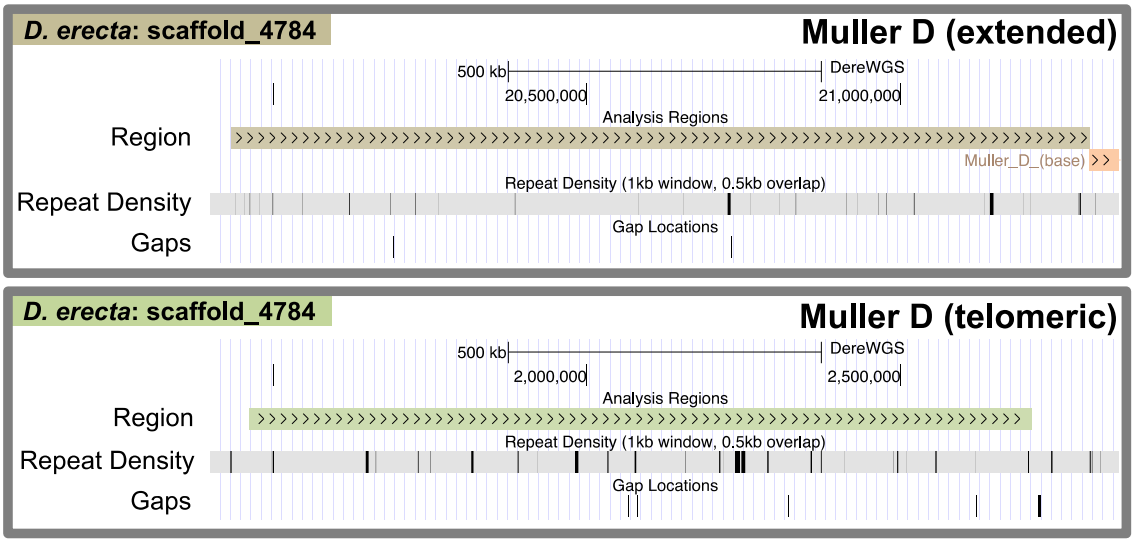


Figure S1 Defining the analysis regions for *D. melanogaster*, *D. erecta*, *D. mojavensis*, and *D. grimshawi* based on changes in repeat density. The bar in the "Region" track corresponds to the region analyzed in this study. The grayscale in the "Repeat Density" track corresponds to the results of sliding window analysis (1 kb window and 0.5 kb step size) of total transposon density using a species-specific repeat library. Darker regions in this track correspond to regions with higher repeat density. One of the characteristics of heterochromatin is its high repeat density and the selections of the euchromatic reference regions at the base of the D element correspond to regions with mostly uniform low repeat density juxtaposed with regions with high repeat density. The Genome Browser screenshots of the base of the D elements show the region that spans from the start of the analysis region to the end of the assembled scaffold.

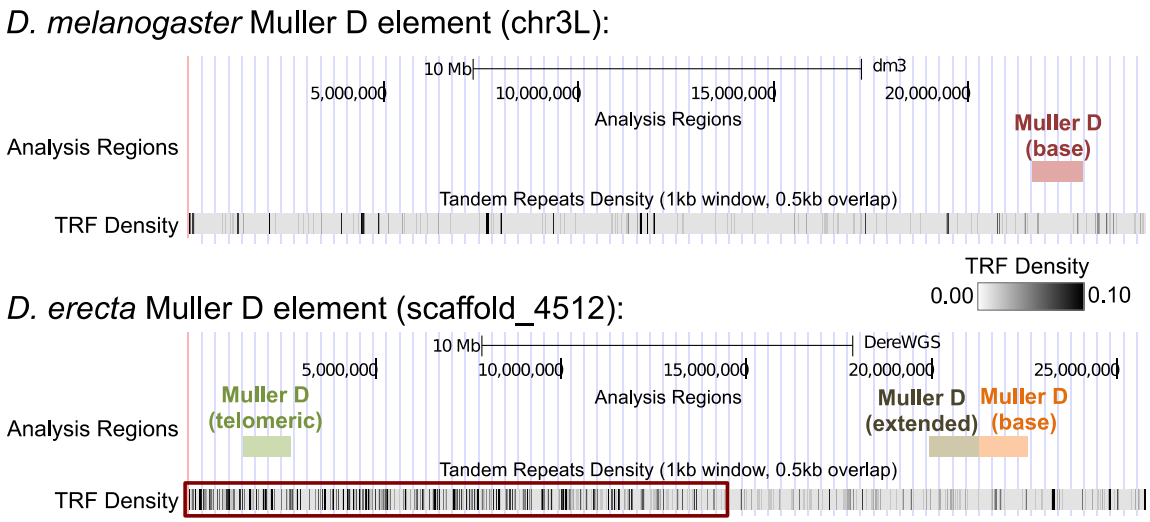


Figure S2 Sliding window analysis of the distribution of tandem repeats shows that the portion of the *D. erecta* D element closer to the telomere has a higher tandem repeat density than the portion of the D element closer to the centromere. Tandem repeats on the D element are identified by Tandem Repeats Finder (TRF) and the density of tandem repeats (TRF density) is calculated using a 1 kb sliding window with 500 bp step size. (Grayscale: darker regions have a higher density of tandem repeats (range from 0 to 10%).) The sliding window analysis shows that the distal half of the *D. erecta* D element (red box) has higher density of tandem repeats than the proximal half. The *D. melanogaster* D element has a lower density of tandem repeats than the *D. erecta* D element, and it does not show the same skew in the density of tandem repeats.

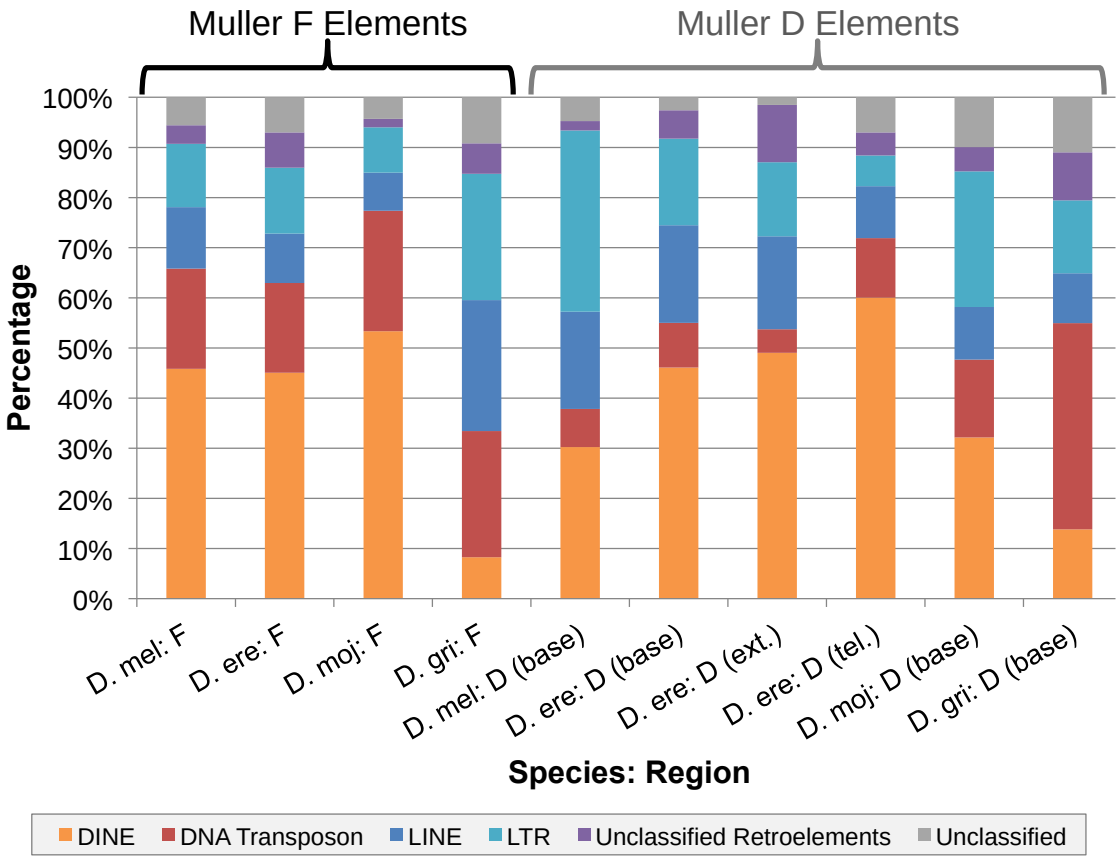


Figure S3 Composition of repeat fragments identified by RepeatMasker using species-specific repeat libraries. Approximately half of the transposon fragments identified by RepeatMasker on the *D. melanogaster* (45.9%), *D. erecta* (45.1%), and *D. mojavensis* (53.3%) F elements have sequence similarity to DINE-1 elements. Only 8.3% of the transposon fragments identified on the *D. grimshawi* F have similarity to DINE-1 elements. The *D. erecta* euchromatic reference regions are enriched in DINE-1 elements compared to the euchromatic regions in the other species.

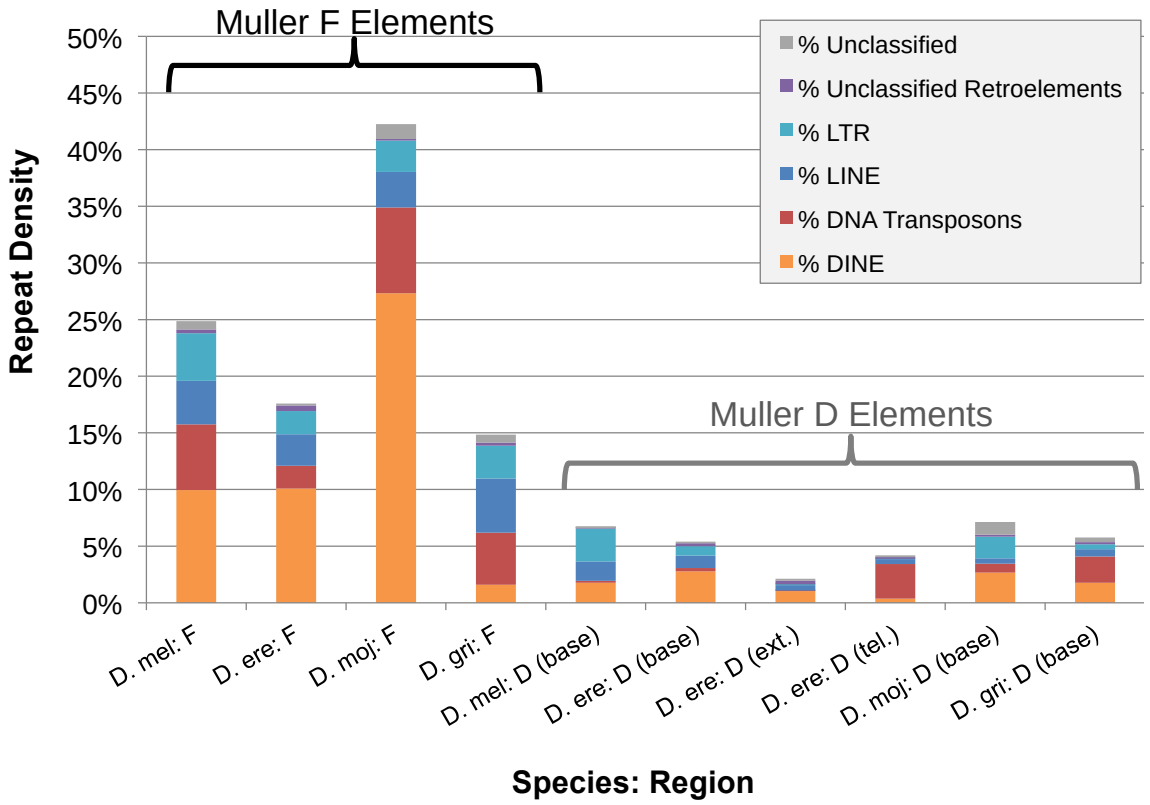


Figure S4 The F element shows higher repeat density than the D euchromatic reference region irrespective of the repeat library used with RepeatMasker. In congruence with the results obtained using the combined repeat library, repeat density estimates using species-specific ReAS repeat libraries show that the F elements have a higher transposon density than the D elements in all four species. Unlike other *de novo* repeat algorithms, ReAS identifies repeats by finding k-mers that occur at a high frequency within genomic reads. Hence the quality of the ReAS library is less likely to be affected by misassemblies.

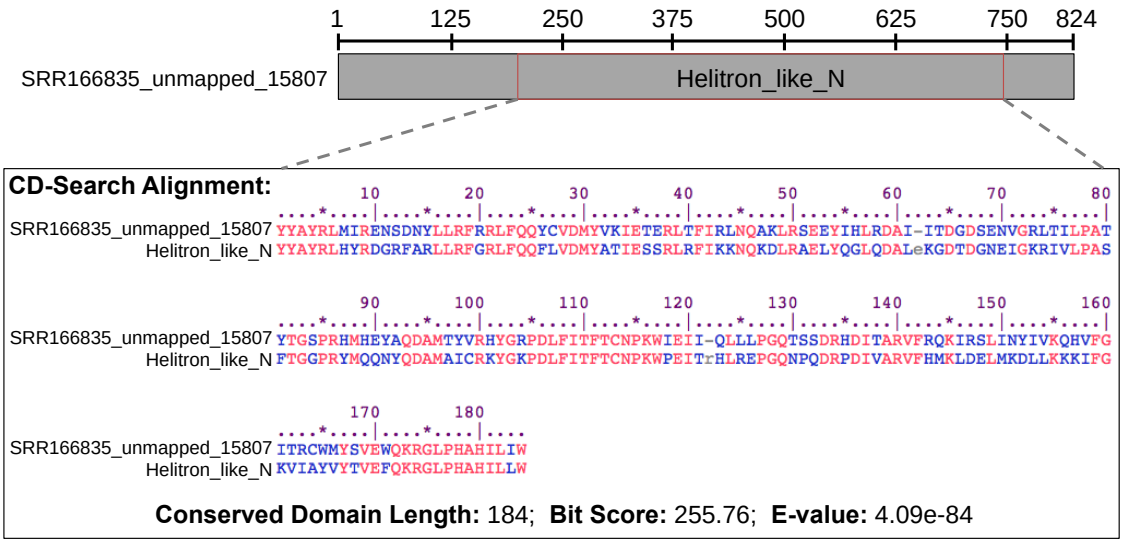
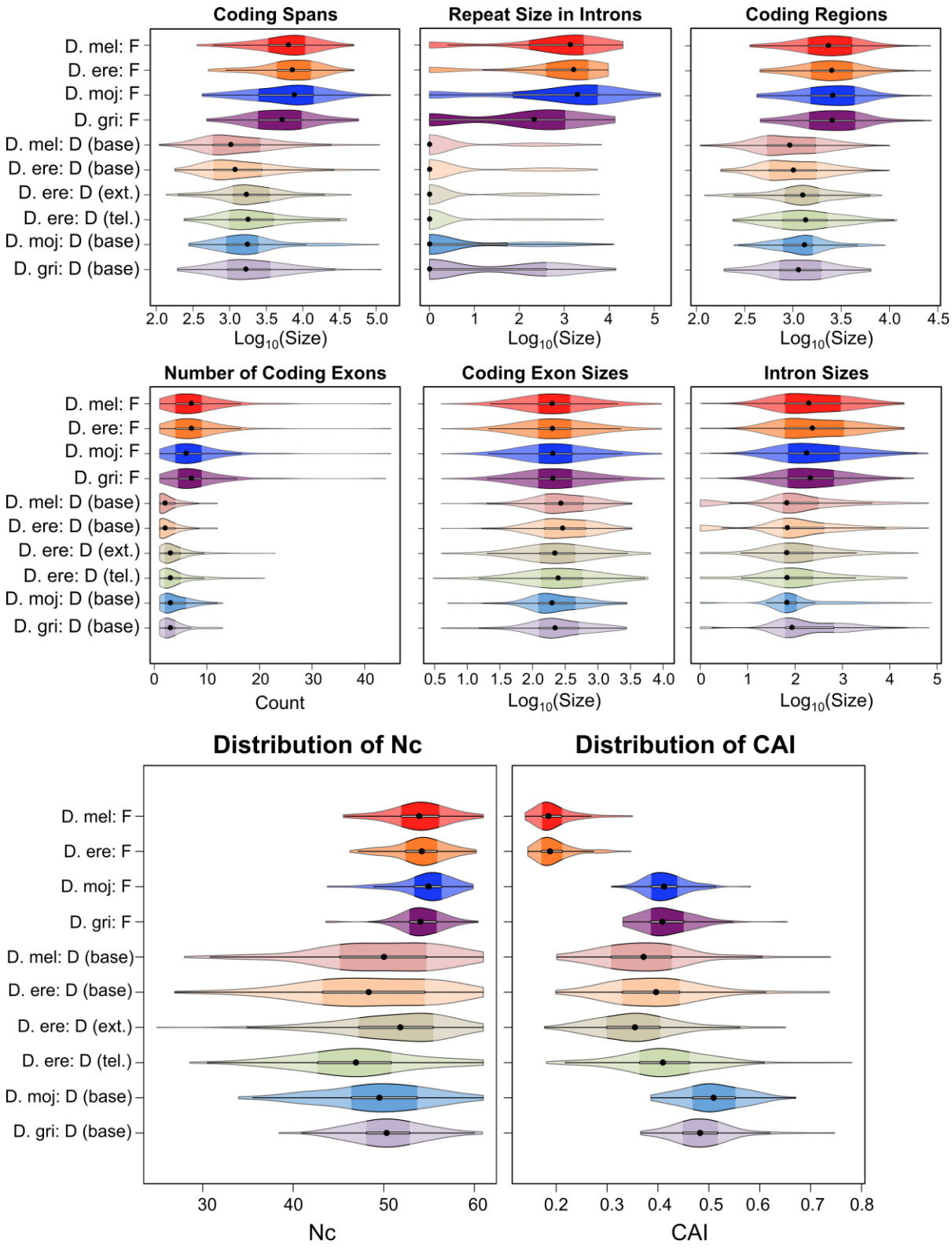


Figure S5 Scaffolds assembled from unmapped RNA-Seq reads show some of the helitrons are being actively transcribed in the *D. mojavensis* genome. CD-Search (Marchler-Bauer *et al.* 2011) of a scaffold assembled from unmapped RNA-Seq reads (SRR166835_unmapped_15807) against the NCBI Conserved Domains database shows a full-length match to the Helitron_like_N domain (pfam 14214). This domain contains a helicase and is commonly found at the N-terminus of helitrons.


Coding Spans
Repeat Size in Introns
Coding Regions
D. mel: F
D. ere: F
D. moj: F
D. gri: F
D. mel: D (base)
D. ere: D (base)
D. ere: D (ext.)
D. ere: D (tel.)
D. moj: D (base)
D. gri: D (base)
2.0 2.5 3.0 3.5 4.0 4.5 5.0
0 1 2 3 4 5
2.0 2.5 3.0 3.5 4.0 4.5
Log10(Size)
Number of Coding Exons
Coding Exon Sizes
Intron Sizes
0 10 20 30 40
Count
0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0
0 1 2 3 4 5
Distribution of Nc
Distribution of CAI
30 40 50 60
Nc
0.2 0.3 0.4 0.5 0.6 0.7 0.8
CAI

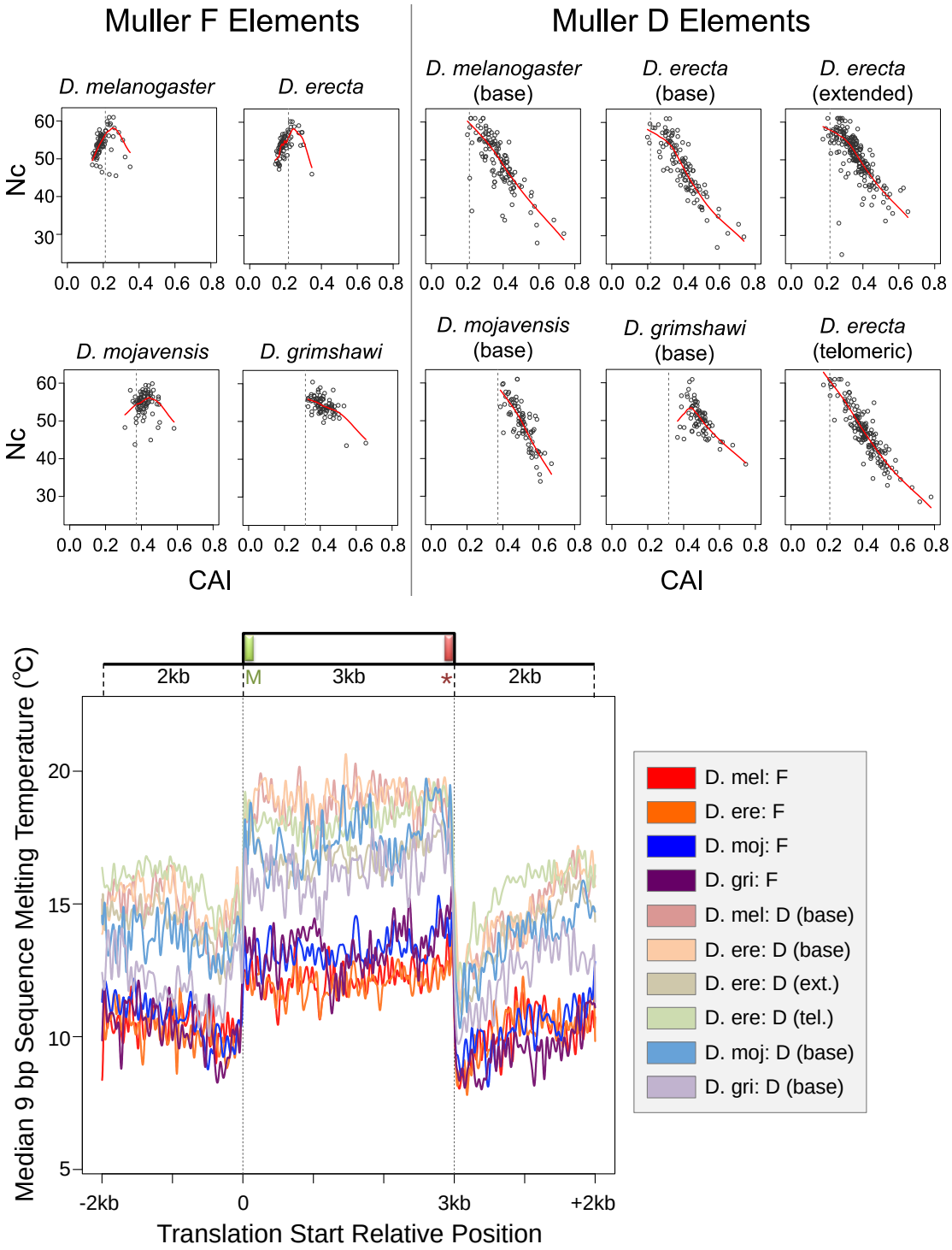


Figure S6 Violin plots, Nc versus CAI scatterplots, and melting temperature profiles for all of the analysis regions. These plots are the same as the figures in the manuscript but they also include the extended and telomeric regions from the *D. erecta* D element (D. ere: D (ext.) and D. ere: D (tel.), respectively). These plots show that F element genes have different characteristics from genes found in the D element euchromatic reference regions. The violin plots also show that the genes in the three regions (base, extended, and telomeric) of the *D. erecta* D element have similar characteristics. These gene characteristics are distinct from those observed on the *D. erecta* F element.

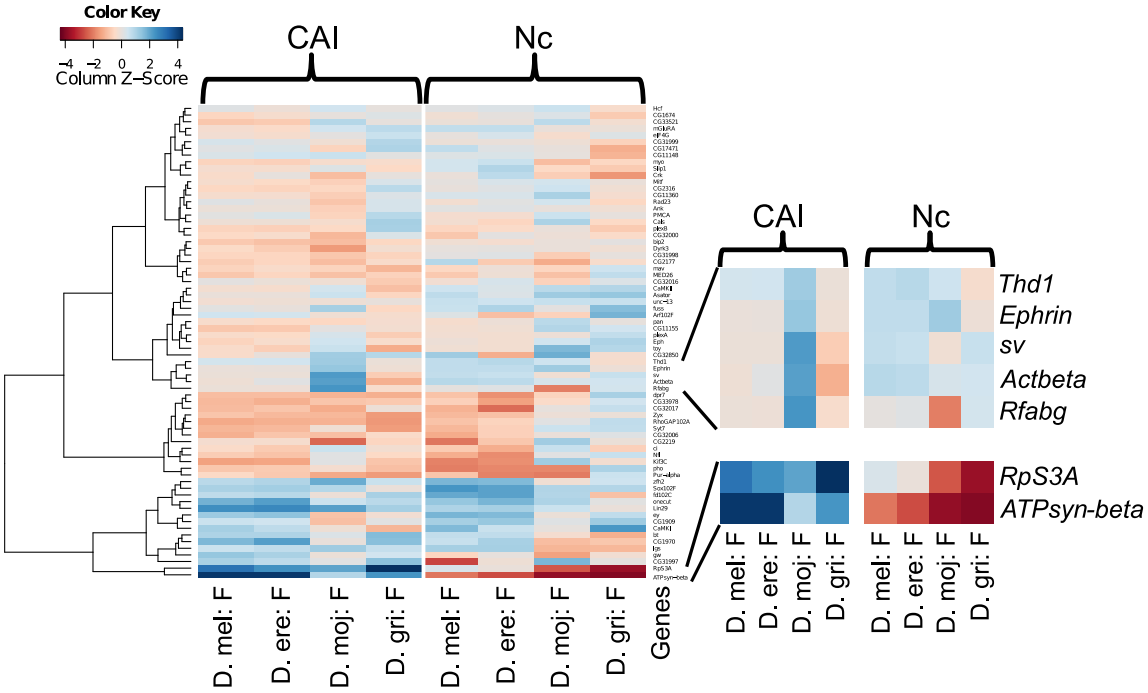


Figure S7 Heat map analysis of codon bias on the F elements shows that genes on the *D. mojavensis* F element exhibit a different pattern of codon bias (measured using CAI and Nc) than *D. melanogaster*, *D. erecta*, and *D. grimshawi*. The *D. mojavensis* orthologs of *Thd1*, *Ephrin*, *sv*, *Actbeta*, and *Rfabg* show higher CAI (blue) than those genes in other Drosophila species (red), indicating more optimal codon usage. *RpS3A* and *ATPsyn-beta* exhibit the highest CAI and lowest Nc, indicating that they are under the strongest selective pressure among all the F element genes in the four Drosophila species. The order of the genes in the heat map is determined by Ward hierarchical clustering.

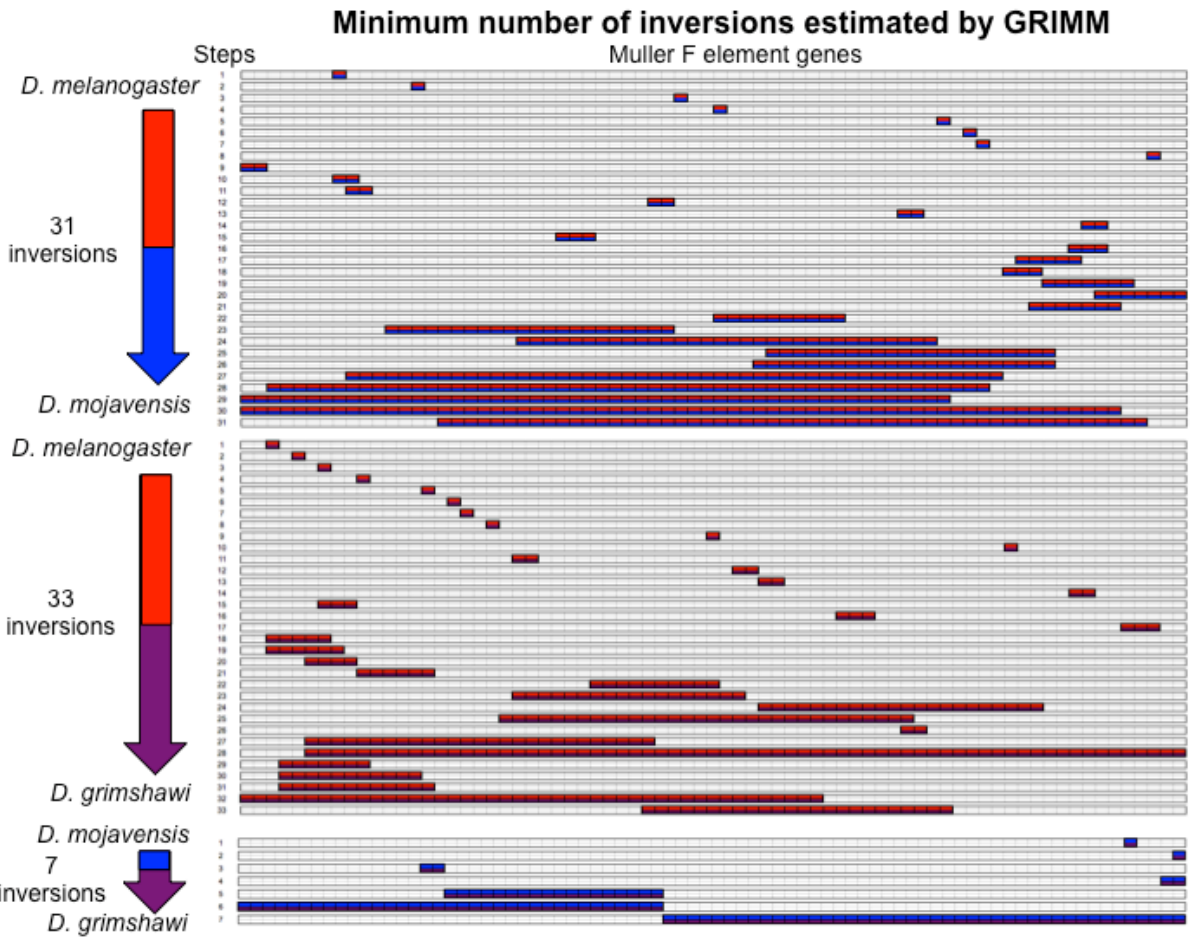


Figure S8 Possible gene inversion scenarios for the *D. melanogaster*, *D. mojavensis,* and *D. grimshawi* F elements as determined by GRIMM. Each box in the diagram corresponds to a gene that is shared between the source and target genomes while each row represents an inversion step. The color boxes in each row correspond to the size of the inverted region at that step. (Top) GRIMM estimates that it requires a minimum of 31 inversions to transform the gene order and orientation seen in the *D. melanogaster* F element to that seen in the *D. mojavensis* F element (72 genes in common). (Middle) A minimum of 33 inversions is required to transform the *D. melanogaster* F element gene order and orientation to that seen in the *D. grimshawi* F element (73 genes in common). (Bottom) A minimum of seven inversions is required to transform the *D. mojavensis* F element gene order and orientation to that seen in the *D. grimshawi* F element (78 genes in common).

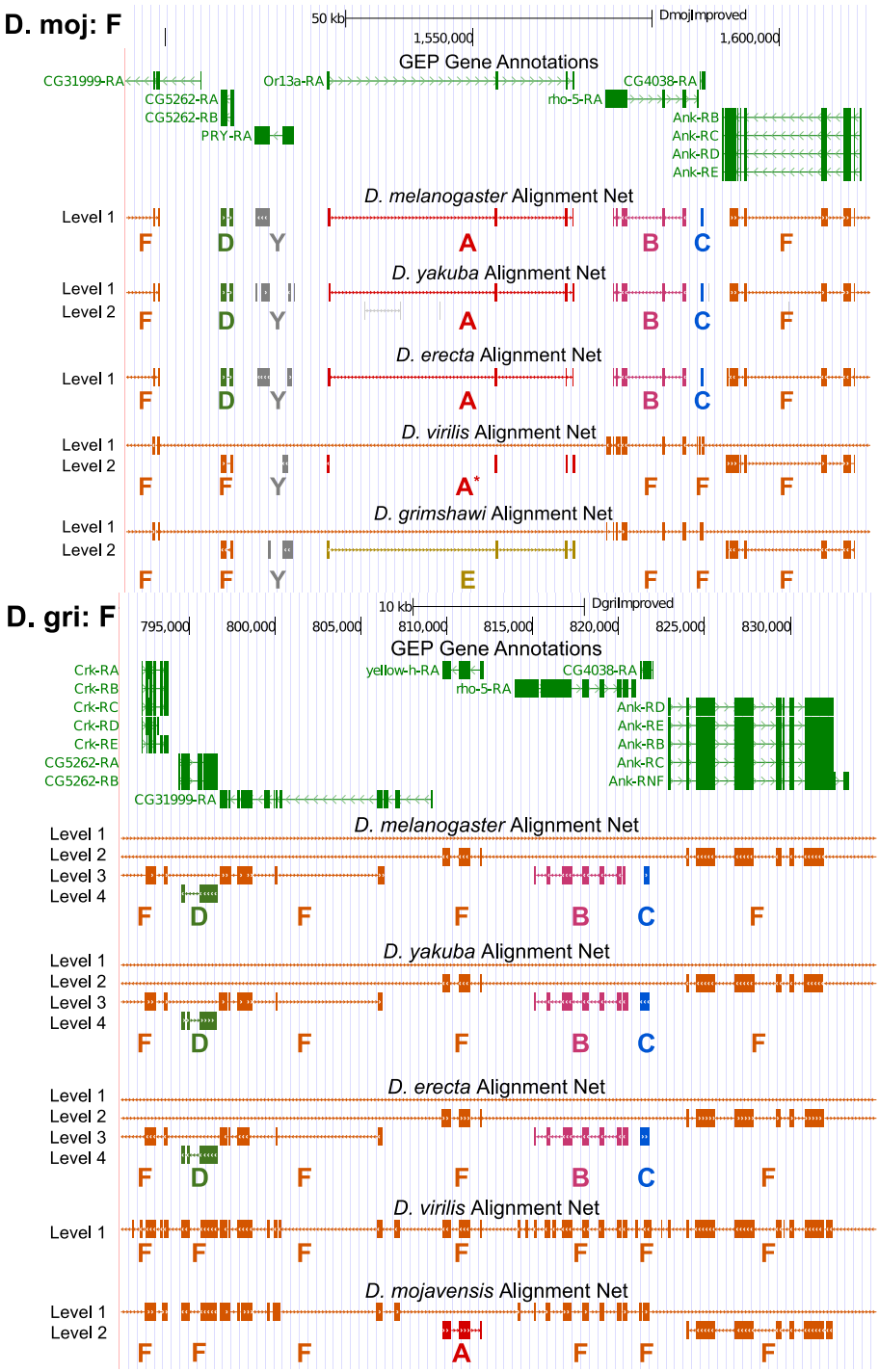


Figure S9 Net alignment of the whole genome assemblies from five Drosophila species against the *D. mojavensis* F element shows a single region (i.e. hotspot) where five of the six *D. melanogaster* wanderer genes are found. A similar hotspot on the *D. grimshawi* F element contains three of the four wanderer genes relative to *D. melanogaster* and one wanderer gene (*yellow-h*) relative to *D. mojavensis*. We have changed the color of the Net alignments to reflect the assignment of genomic scaffolds to Muller elements determined by previous studies (Schaeffer *et al.* 2008; Koerich *et al.* 2008).

SUPPORTING TABLES

Table S1 Regions used in the comparative analyses

Label	Database (Assembly)	Region in assembly	First gene in region	Last gene in region
D. mel: F	dm3 (Release 5)	chr4:24053-1274855	*JYalpha*	*Caps*
D. mel: D (base)	dm3 (Release 5)	chr3L:21645371–22948385	*CG43980*	*CG32461*
D. ere: F	DereWGS (CAF1)	scaffold_4512:1–1283782	*Caps*	*CG40625*
D. ere: D (base)	DereWGS (CAF1)	scaffold_4784:21301945–22602847	*CG43980*	*CG32461*
D. ere: D (ext.)	DereWGS (CAF1)	scaffold_4784:19933404–21301944	*CG7017*	*Atox1*
D. ere: D (tel.)	DereWGS (CAF1)	scaffold_4784:1461828–2709938	*stet*	*Fife*
D. moj: F	DmojImproved (GEP)	improved_6498:1348734–3038944	*CG31999*	*Cyp1_alpha*
D. moj: D (base)	DmojImproved (GEP)	improved_6680:21061438–22011287	*CG14826*	*CG11370*
D. gri: F	DgriImproved (GEP)	improved_F:28556–1228705	*CG32850*	*sv*
D. gri: D (base)	DgriImproved (CAF1)	scaffold_15110:22264226–23522110	*W*	*fal*

Table S2 Sequence improvement statistics

A. Number of gaps closed by sequence improvement

Region	# Gaps (Improved)	Total gap size (Improved)	# Gaps (CAF1)	Total gap size (CAF1)	Difference # gaps	Difference gap size
Dgri: F	13	13,670	56	42,648	-43	-28,978
Dmoj: D (base)	0	0	7	1,765	-7	-1,765
Dmoj: F	1	1,166	23	14,891	-22	-13,725
Total	**14**	**14,836**	**86**	**59,304**	**-72**	**-44,468**

B. Number of changes between the improved and CAF1 assemblies

Region	# Changes	Total size of changes	# Single base changes	% Single base changes
Dgri: F	151	56,000*	56	37.1%
Dmoj: D (base)	42	2,791	18	42.9%
Dmoj: F	116	15,500	53	45.7%
Total	**309**	**74,291**	**127**	**41.1%**

* The first 40 kb of the *D. grimshawi* F element scaffold in the CAF1 assembly contains a major misassembly. The discrepancies in this region account for 25/151 changes and cover a total of 24,303 bases.

Table S3 Comparison of GLEAN-R gene predictions and the most comprehensive isoform (i.e. isoform with the largest coding region) annotated by the GEP

A. GLEAN-R gene models that match the comprehensive isoform annotated by the GEP

Region	# Match	# Mismatch	# GEP annotations	# GLEAN-R annotations	% Match
Dere: F	53	28	78	81	65.4%
Dmoj: F	26	55	81	81	32.1%
Dgri: F	36	56	79	92	39.1%
Dere: D (base)	84	37	110	121	69.4%
Dere: D (ext.)	141	81	207	222	63.5%
Dere: D (tel.)	108	62	162	170	63.5%
Dmoj: D (base)	57	42	84	99	57.6%
Dgri: D (base)	47	34	77	81	58.0%
Total	**552**	**395**	**878**	**947**	**58.3%**

B. GLEAN-R coding exons that match the coding exons in the comprehensive isoform annotated by the GEP

Region	# Match	# Mismatch	# GEP Annotations	# GLEAN-R annotations	% Match
Dere: F	510	56	597	566	90.1%
Dmoj: F	451	108	624	559	80.7%
Dgri: F	513	122	629	635	80.8%
Dere: D (base)	288	49	324	337	85.5%
Dere: D (ext.)	737	99	827	836	88.2%
Dere: D (tel.)	613	84	711	697	87.9%
Dmoj: D (base)	305	73	353	378	80.7%
Dgri: D (base)	231	48	277	279	82.8%
Total	**3648**	**639**	**4342**	**4287**	**85.1%**

Table S4 Results of the Tallymer to determine the fraction of unique k-mers in each analysis region

A. F elements

K-mer	Dmel: F	Dere: F	Dmoj: F	Dgri: F
8	0.015	0.015	0.006	0.018
9	0.203	0.196	0.148	0.221
10	0.525	0.515	0.472	0.540
11	0.751	0.742	0.717	0.751
12	0.870	0.862	0.843	0.858
13	0.927	0.921	0.900	0.911
14	0.953	0.949	0.925	0.938
15	0.965	0.961	0.937	0.953
16	0.970	0.967	0.943	0.962
17	0.973	0.970	0.947	0.967
18	0.975	0.972	0.949	0.971
19	0.976	0.974	0.952	0.973
20	0.977	0.975	0.954	0.975

B. Euchromatic reference regions from the D elements

K-mer	Dmel: D (base)	Dere: D (base)	Dere: D (ext.)	Dere: D (tel.)	Dmoj: D (base)	Dgri: D (base)
8	0.002	0.002	0.002	0.003	0.022	0.016
9	0.121	0.124	0.117	0.136	0.237	0.199
10	0.501	0.501	0.488	0.511	0.578	0.517
11	0.786	0.787	0.779	0.795	0.804	0.754
12	0.916	0.918	0.915	0.924	0.911	0.882
13	0.965	0.968	0.968	0.971	0.957	0.943
14	0.982	0.985	0.987	0.986	0.976	0.969
15	0.988	0.991	0.993	0.992	0.984	0.981
16	0.990	0.993	0.995	0.993	0.987	0.986
17	0.991	0.994	0.996	0.994	0.989	0.989
18	0.992	0.995	0.997	0.994	0.990	0.991
19	0.992	0.995	0.997	0.995	0.991	0.992
20	0.992	0.996	0.997	0.995	0.992	0.992

Table S5 Statistics on the DINE-1 fragments identified by the species-specific library that do not overlap with known repeats in the RepBase library.

Region	DINE-1 elements that do not overlap with RepBase repeats (count)	DINE-1 elements that do not overlap with RepBase repeats (bp)	DINE-1 elements identified by the species-specific library (bp)	Percentage of DINE-1 elements that do not overlap with RepBase repeats (bp)
Dmel: F	278	12,479	161,138	7.7%
Dere: F	209	8,225	144,399	5.7%
Dmoj: F	602	40,587	449,309	9.0%
Dgri: F	120	8,524	19,533	43.6%
Dmel: D (base)	62	2,999	31,903	9.4%
Dere: D (base)	64	2,265	41,760	5.4%
Dere: D (ext.)	33	1,589	21,657	7.3%
Dere: D (tel.)	25	15,470	34,476	44.9%
Dmoj: D (base)	161	9,782	33,262	29.4%
Dgri: D (base)	123	8,256	19,751	41.8%

Table S6 Gene density for the different analysis regions

Region	# Genes	Region size (bp)	# Genes / Mb
Dmel: F	79	1250803	63
Dere: F	77	1283782	60
Dmoj: F	81	1690211	48
Dgri: F	79	1200150	66
Dmel: D (base)	121	1303015	93
Dere: D (base)	110	1300903	85
Dere: D (ext.)	207	1368541	151
Dere: D (tel.)	161	1248111	129
Dmoj: D (base)	84	949850	88
Dgri: D (base)	76	1257885	60

SUPPORTING FILES

Available for download at http://www.g3journal.org/lookup/suppl/doi:10.1534/g3.114.015966/-/DC1

File S1

Supplemental text and methods

File S2

Summary of gene structure changes compared to *D. melanogaster*

File S3

WindowMasker, tandem repeats, and simple repeats statistics

File S4

Transposon density estimates using different transposon libraries

File S5

Statistics on RepBase transposons that overlap with DINE-1 fragments

File S6

Kruskal-Wallis rank sum test results for different gene characteristics

File S7

Codon usage heat map

File S8

List of participating courses

File S9

Improved sequences and annotations in GenBank ASN.1 format