FILE S1

Additional Materials, Methods, and References

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

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

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

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

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

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

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

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

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

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

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

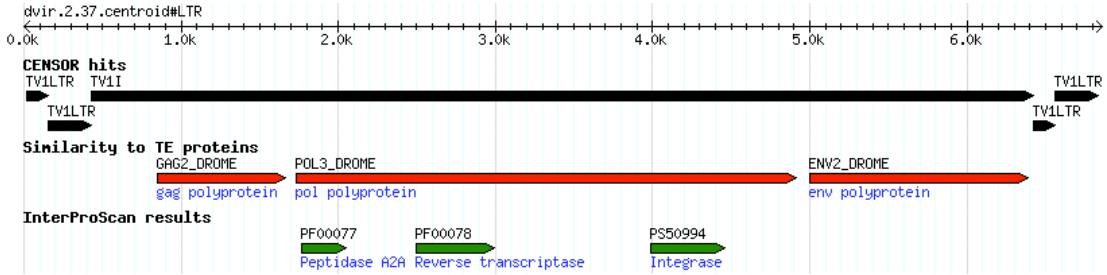
Possible errors in the coding regions of the CAF1 dot chromosome sequence
A. *CG33521*
CG33521 shows an incomplete alignment (with 130/140 bases aligned) between the GEP and CAF1 strains. Reconstruction of this discrepant region of the CAF1 assembly suggests that the differences in the last 10 bases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

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

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

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

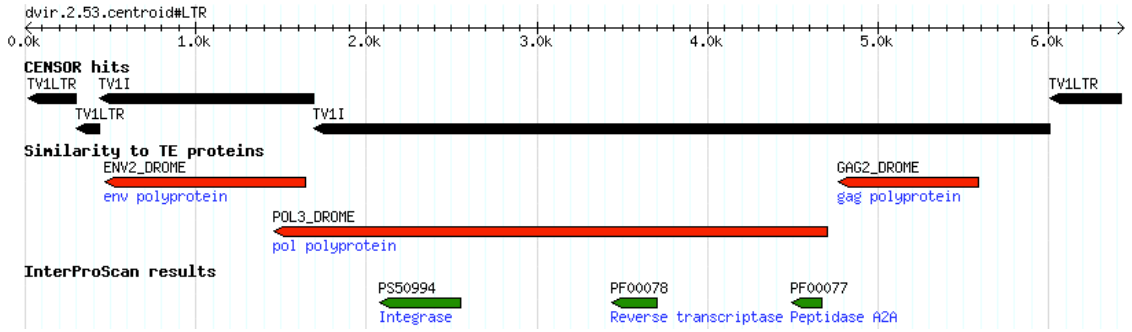
A. *dvir.2.37.centroid*



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

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

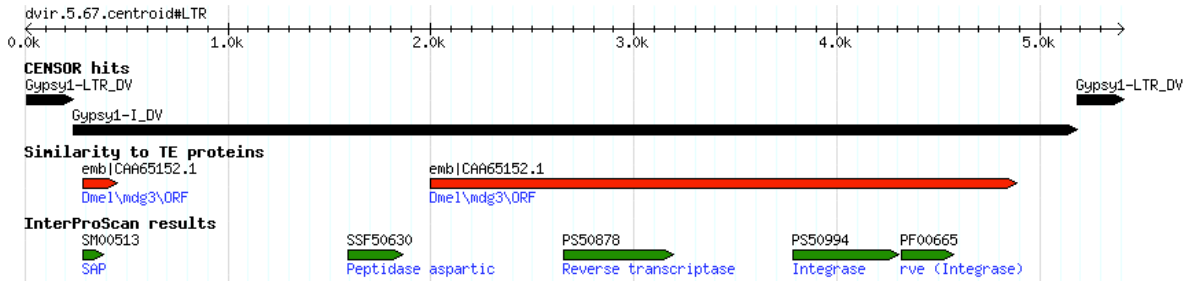
B. *dvir.2.53.centroid*



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

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

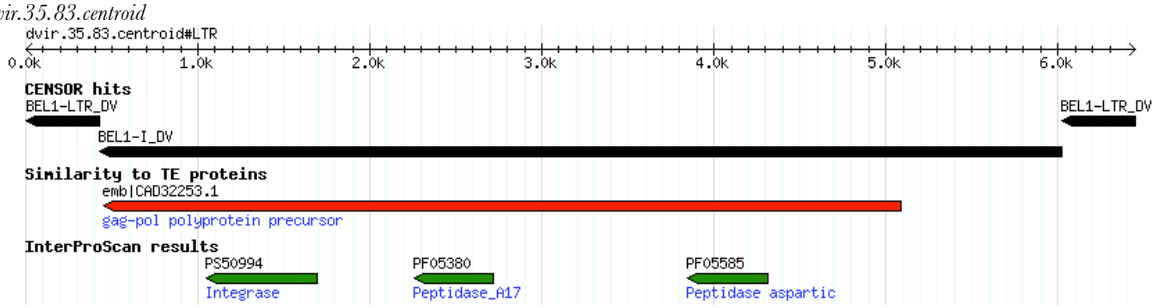
C. *dvir.5.67.centroid*



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

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

D. *dvir.35.83.centroid*



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

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

LITERATURE CITED

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

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

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

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

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

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

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