Supplemental Material

This document contains Supplemental Tables S1-2 and Supplemental Figures S1-9 for the article "Genomic regulatory blocks underlie extensive microsynteny conservation in insects" by P. G. Engström et al.

Supplemental Table S3 is in a separate document (Excel workbook).

Table S1. Overlap with HCNE-dense regions for selected gene categories

Gene category		**Number of genes overlapping HCNE-dense regions** [a]	**Total amount of gene sequence within HCNE-dense regions** [b]
Genes annotated with GO term(s)	generation of precursor metabolites and energy (GO:0006091)	14/491 (2.85%)	29k/2056k (1.40%)
	cellular protein metabolic process (GO:0044267)	77/2030 (3.79%)	334k/11549k (2.90%)
	cell organization and biogenesis (GO:0016043)	76/1680 (4.52%)	775k/15021k (5.16%)
	transport (GO:0006810)	69/1517 (4.55%)	185k/10115k (1.83%)
	signal transduction (GO:0007165)	83/1338 (6.20%)	876k/15800k (5.54%)
	multicellular organismal development (GO:0007275)	133/1380 (9.64%)	1829k/18544k (9.86%)
	regulation of transcription, DNA-dependent (GO:0006355)	103/768 (13.41%)	1385k/7736k (17.90%)
	multicellular organismal development (GO:0007275) AND regulation of transcription, DNA-dependent (GO:0006355)	81/334 (24.25%)	1234k/5313k (23.22%)
All genes		684/13733 (4.98%)	4001k/67464k (5.93%)

[a] Ratio between number of genes overlapping HCNE-dense regions and the total number of genes in the category. [b] Ratio between amount of gene sequence covered by HCNE-dense regions and the total amount of sequence spanned by genes in the category. HCNE-dense regions were identified by sliding a 40 kb window across the genome in steps of 1 kb and reporting windows for which at least 1% (400 bp) of the sequence was covered by HCNEs. Overlapping HCNE-dense windows were merged, resulting in 421 HCNE-dense regions covering a total of 13.5 Mb (11.41% of the investigated sequence).

Table S2. Properties of pairwise and 5-way synteny blocks among *Dmel* and four other *Drosophila* species

	Pairwise synteny blocks between *Dmel* and indicated query species				Five-way synteny blocks
	Dana	*Dpse*	*Dvir*	*Dmoj*	
Number of synteny blocks	627	814	920	923	899
Number of query sequences included	30 / 13,772 scaffolds	16 / 16 ultra-scaffolds 26 / 2649 scaffolds+contigs	24 / 13,562 scaffolds	11 / 6843 scaffolds	n.c.
Number of query sequence fusions [a]	11	5	7	3	n.c.
Amount of query sequence spanned	114 Mb (49%)	120 Mb (85%)	131 Mb (63%)	134 Mb (69%)	n.c.
Amount of *Dmel* sequence spanned	111 Mb (94%)	110 Mb (93%)	107 Mb (90%)	105 Mb (89%)	90 Mb (76%)
Number of *Dmel* genes spanned [b]	12863 (94%)	12434 (91%)	11690 (85%)	11451 (83%)	8958 (65%)
Number of *Dmel* genes covered [c]	11631 (85%)	10685 (78%)	9531 (69%)	9279 (68%)	6308 (46%)

n.c., not calculated

[a] To avoid artificial synteny breaks due to scaffold breaks in the assembly, our algorithm for synteny block construction includes a step for fusing scaffolds that, when combined, show colinearity with the *Dmel* sequence (see Methods).

[b] We considered a gene to be spanned by a synteny block if at least 90% of its total CDS were within the extreme borders of the block on the *Dmel* genome.

[c] We considered a gene to be covered by a synteny block if the block contained aligned sequence for at least 60% the gene's total CDS.

None of the four species that we compared to *Dmel* has a finished genome assembly. *Dana* appears to have the most fragmented assembly, consisting of 13,772 scaffolds. Nevertheless, our results indicate that reliable synteny blocks can be constructed because most of the sequence is in very large scaffolds. Although the pairwise synteny blocks included few scaffolds (e.g., 30 *Dana* scaffolds were included), they spanned more than 89% of the *Dmel* euchromatic sequence. As expected, the number of synteny blocks increased with evolutionary distance between *Dmel* and the compared species (consistent with an accumulation of rearrangements over evolutionary time). Notably, for *Dvir* and *Dmoj*, which are equidistant from *Dmel*, we obtained nearly identical synteny block counts. Not unexpectedly, the span of synteny blocks over the *Dmel* sequence and over gene annotations on that sequence decreased with increasing evolutionary distance to the compared species.

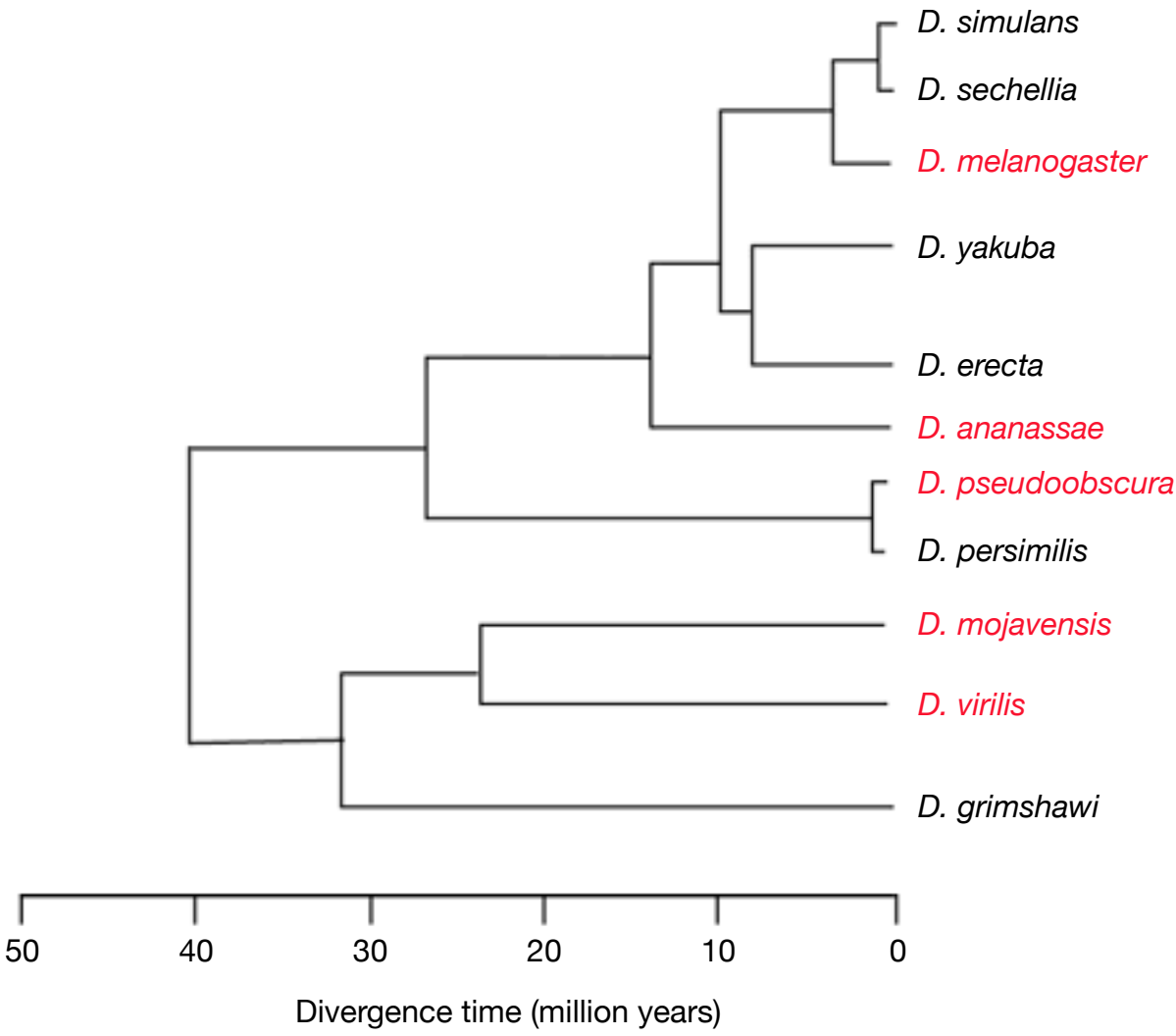

Figure S1. Phylogenetic tree of *Drosophila* species.
Shown are *Drosophila* species for which whole-genome alignments to *D. melanogaster* were available in the UCSC Genome Browser Database (http://genome.ucsc.edu) at the time of this study. The species used in this study are shown in red. Although the remote phylogenetic position of *D. grimshawi* suggests that including that organism would have been informative, we chose not to because its genome assembly appeared to be in an early state (25,052 scaffolds compared to at most 13,772 scaffolds for any of the included species; see Supplemental Table S2).

Adapted from the figure at http://rana.lbl.gov/drosophila/.

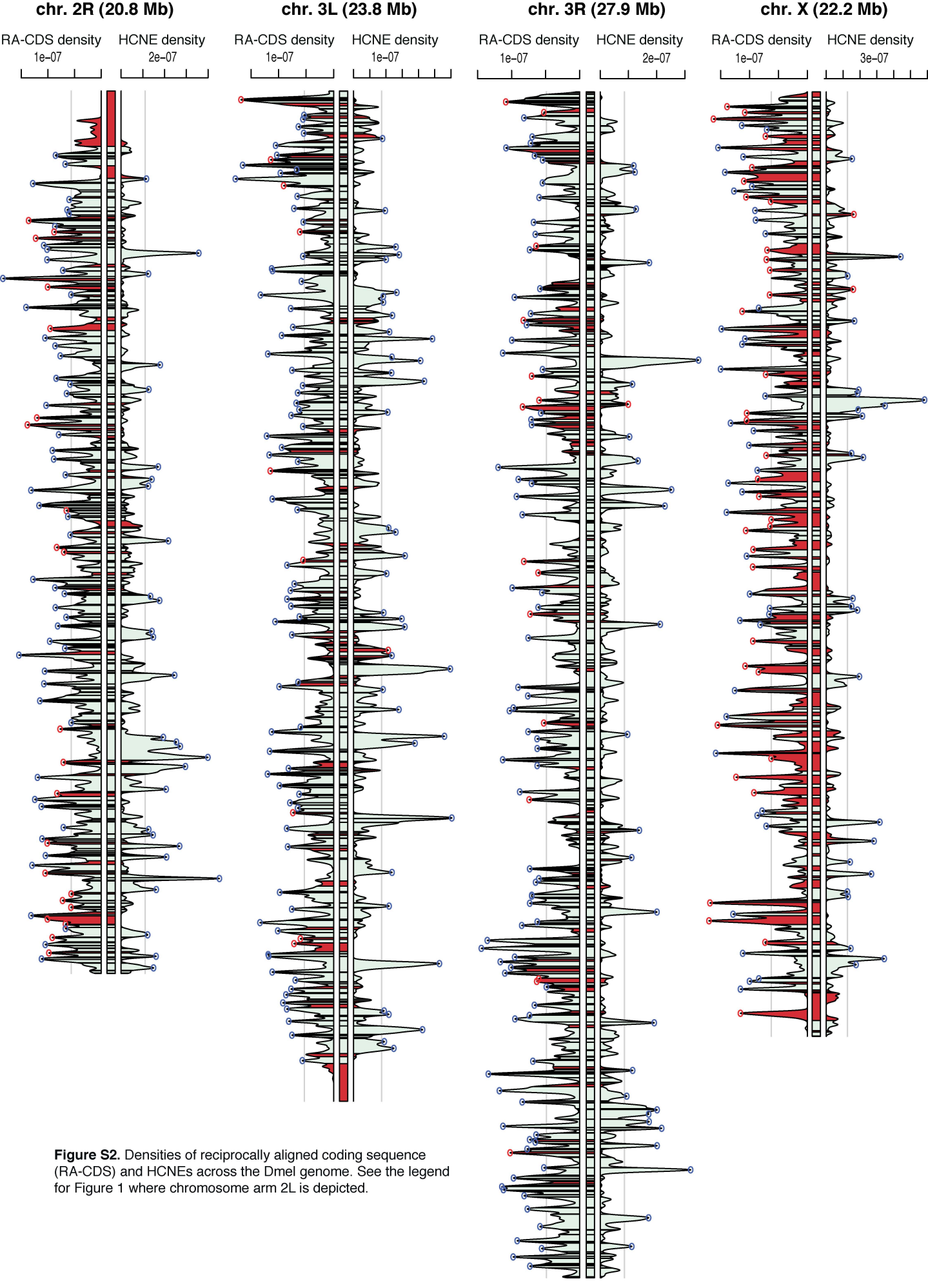

Figure S2. Densities of reciprocally aligned coding sequence (RA-CDS) and HCNEs across the Dmel genome. See the legend for Figure 1 where chromosome arm 2L is depicted.

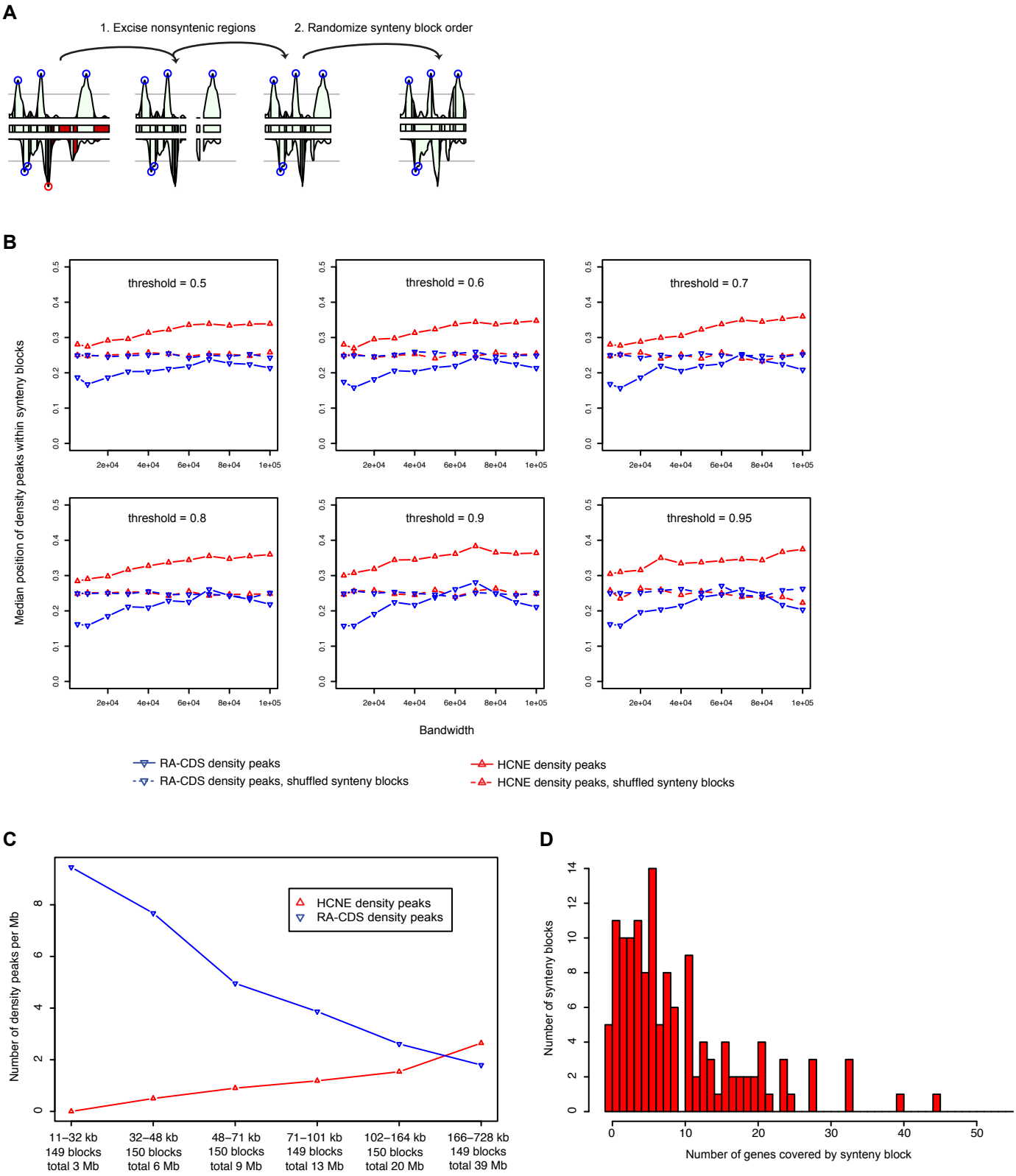

Figure S3. HCNE density peaks tend to be centrally located in large synteny blocks that contain multiple genes.
See also Figure 1.
(A) Randomization strategy. Nonsyntenic regions were excised and the order of synteny blocks was randomized, while the positions of peaks were maintained.
(B) Effects of varying the bandwidth for density computation and threshold for peak detection. Symbols connected by dashed lines represent medians over 10 randomizations. The trend for HCNE density peaks to be centrally positioned within synteny blocks persisted at all parameter settings.The trend for RA-CDS peaks to be positioned close to synteny breaks disappeared at high density computation bandwidths, which appear too high for a relevant comparison with synteny blocks (Supplemental Fig. S4). Unless otherwise noted, results presented in this paper were obtained with a bandwidth of 30,000 and threshold of 0.8.
(C) The frequency of HCNE density peaks per sequence length increases with synteny block size. Based on their span in the *Dmel* genome, the synteny blocks on *Dmel* chromosomes 2, 3 and X were grouped into six groups of 149-150 blocks each. For each group, the range of synteny block spans and their total span are indicated on the x-axis. The y-value for each group shows the ratio between the total number of peaks and the total number of bases spanned by the synteny blocks in the group.
(D) Histogram of number of genes covered by synteny blocks that span HCNE density peaks. Out of all 136 synteny blocks that spanned a HCNE density peak, 120 (88%) also covered multiple genes. As in Table 1, we considered a gene to be covered by a synteny block if the block contained aligned sequence for at least 60% the gene's total CDS.

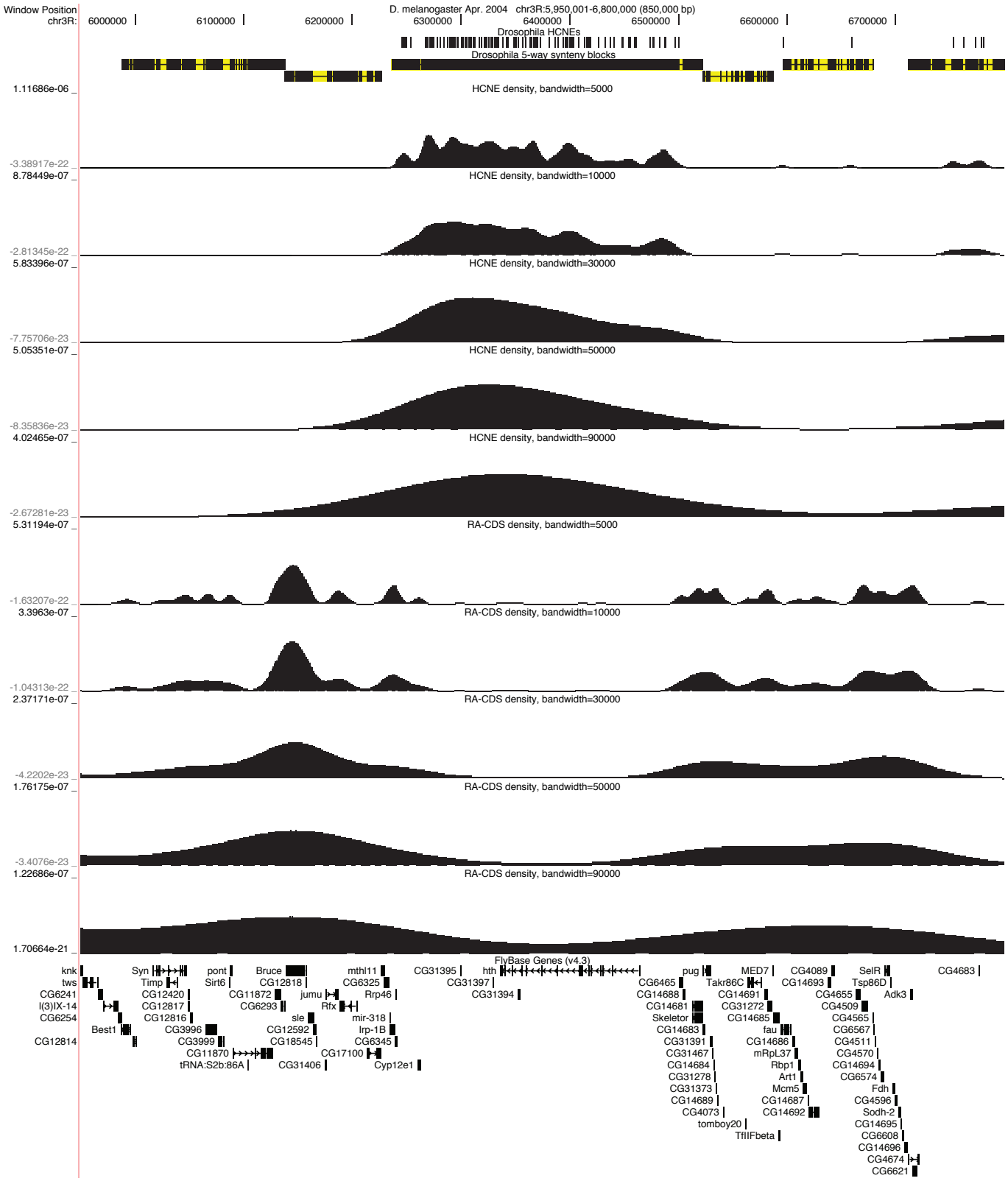

Figure S4. Examples of HCNE and RA-CDS density curves computed with different bandwidths. Shown is the *D. melanogaster* genomic region around the *homothorax* (*hth*) gene.

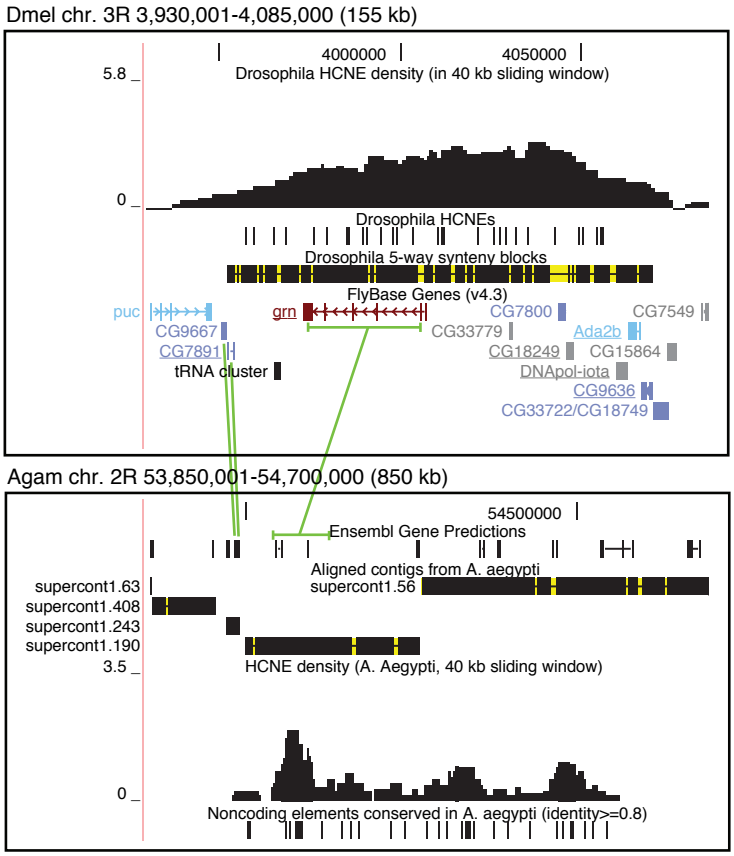

Figure S5. The grn locus in *Dmel* (upper panel), compared to a wider region in *Agam* (lower panel).
Notation as in Figure 2. *grn* and five other protein-coding genes (underlined) show strong evidence of being in conserved microsynteny among the five investigated flies. No functional relationship has been described between any of these genes, although both *grn* and *ada2b* have regulatory functions in development (Brown and Castelli-Gair Hombria, 2000, *Development* 127:4867; Qi et al., 2004, *Mol Cell Biol*, 24:8080). Between *Dmel* and *Agam*, *grn* has been maintained in microsynteny with two downstream genes. The HCNE density peak over the *grn* ortholog in *Agam* is the largest within the 850 kb region examined. Note the elevated gene density around the right border of the synteny block in *Dmel*.

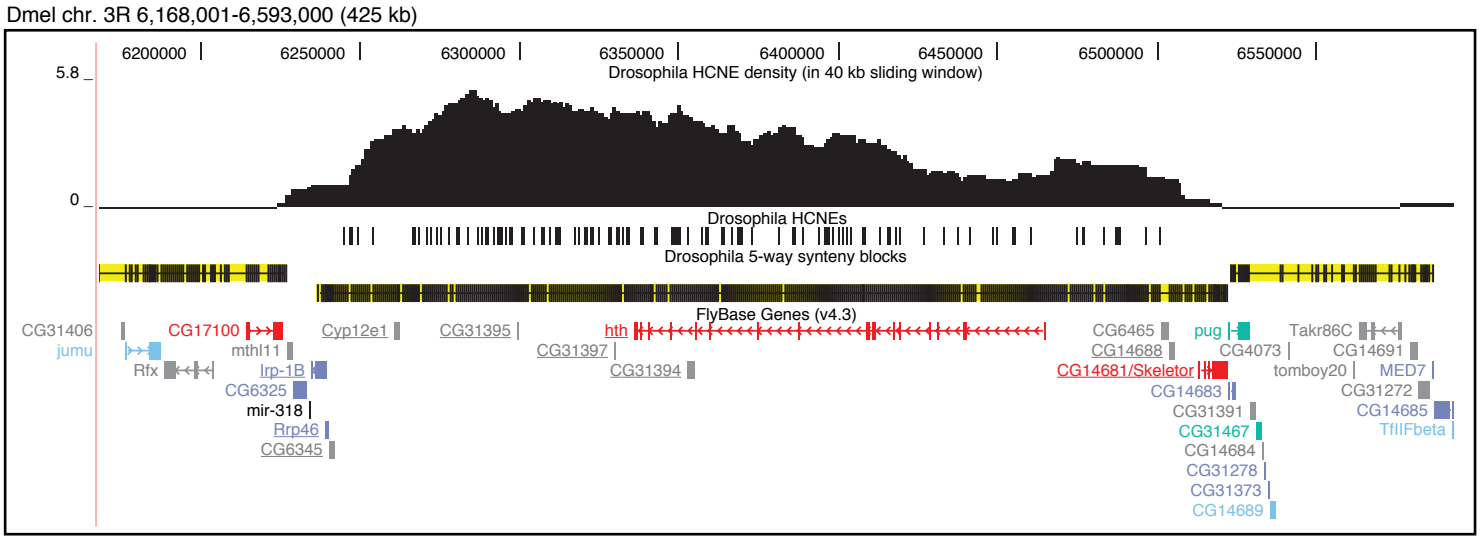

Figure S6. The *hth* locus in *Dmel*.

Notation as in Figure 2. The developmental regulatory homeobox gene *homothorax* (*hth*) occupies one of the most HCNE-dense synteny blocks in the *Dmel* genome. Its human orthologs *MEIS1* and *MEIS2* are surrounded by an exceptional number of noncoding elements conserved across vertebrates (Sandelin et al., 2004, *BMC Genomics* 5:99). Dmel *hth* is in conserved microsynteny with at least nine other genes (underlined) among the five investigated flies. Six of these genes encode protein of diverse classes: Irp-1B, an iron regulatory protein uniformly expressed during development (Muckenthaler et al., 1998, *Eur J Biochem* 254:230); Rrp46, a component of the RNA-processing exosome (Graham et al., 2006, *Mol Biol Cell* 17:1399); Cyp12e1, a cytochrome P450 protein; CG6465, a putative peptidase; CG14688, a putative phytanoyl-CoA dioxygenase; and Skeletor, a chromosomal protein that relocalizes during mitosis (Walker et al., 2000, *J Cell Biol* 151:1401). The remaining three genes are unannotated. The function of an additional protein expressed from the *CG14681/skeletor* transcriptional unit is unknown (Walker et al., 2000). Micro-RNA 318, which was cloned from adult flies (Aravin et al., 2003, *Dev Cell*, 5:337), aligns to the *hth* locus in all five flies, but at different locations relative to *irp-1b*, indicating that local rearrangements have occurred at the edge of the synteny block. *hth* is the only gene in the synteny block known to have a key regulatory role in development. *hth* is in conserved microsynteny with the gene *Rrp46* among flies, mosquitos, bees and beetles, although there is no evidence for a functional relationship between *hth* and *Rrp46*, which encodes a component of the RNA-processing exosome (Graham et al., 2006). In *Dmel* the two genes are separated by a region of 95 kb that is HCNE-dense and gene-sparse, containing only four genes, all of which are in conserved microsynteny with *hth* among the investigated flies. Similarly, the honeybee (*Apis mellifera*) orthologs of *hth* and *Rrp46* are separated by a gene-sparse ~130 kb region (devoid of matches to known *Dmel* proteins in the UCSC Genome Browser). In *Agam* and the beetle *Tribolium castaneum*, the intervening regions also appear to be gene-free, but are smaller (~8 kb and ~10 kb, respectively; estimated from alignments of *Dmel* transcripts to the 2005-10-11 version of the *T. castaneum* assembly, using the BLAST interface at the Baylor HGSC website, www.hgsc.bcm.tmc.edu). In *Aaeg*, there appears to be a gene desert of at least 45 kb downstream of *hth*. (Since this gene desert is at the end of a supercontig, we could not determine whether *hth* and *Rrp46* are linked in *Aaeg*.) Taken together, these data suggest that a regulatory region is present downstream of *hth* in many insect species, underlies microsynteny conservation of multiple fly genes, but has been extensively modified in evolution and possibly lost in some insect species, including *Agam*. Consistent with the variations among insects in the size of the region separating *hth* and *Rrp46*, regulatory sequences appear to have diverged greatly at the mosquito *hth* locus. Little noncoding sequence is highly conserved between the *hth* loci of *Agam* and *Aaeg*: using the same detection criteria as for the *ct, tup* and *grn* loci shown in Figure 2 and Supplemental Figure S5, we only found three HCNEs within *hth* introns and none in the region between *hth* and *Rrp46*. We reasoned that closer sequence comparisons may be required to detect the regulatory elements at the mosquito *hth* locus, so we compared the *Aaeg* supercontig harboring the locus to trace reads from the genome of *Culex pipiens*, which is more closely related to *Aaeg* than *Agam*. Indeed, the introns of *Aaeg hth* contain numerous elements conserved in *Culex* at higher levels than most surrounding exons (Supplemental Fig. S7), confirming that HCNEs are abundant at the *hth* locus in both flies and mosquitos.

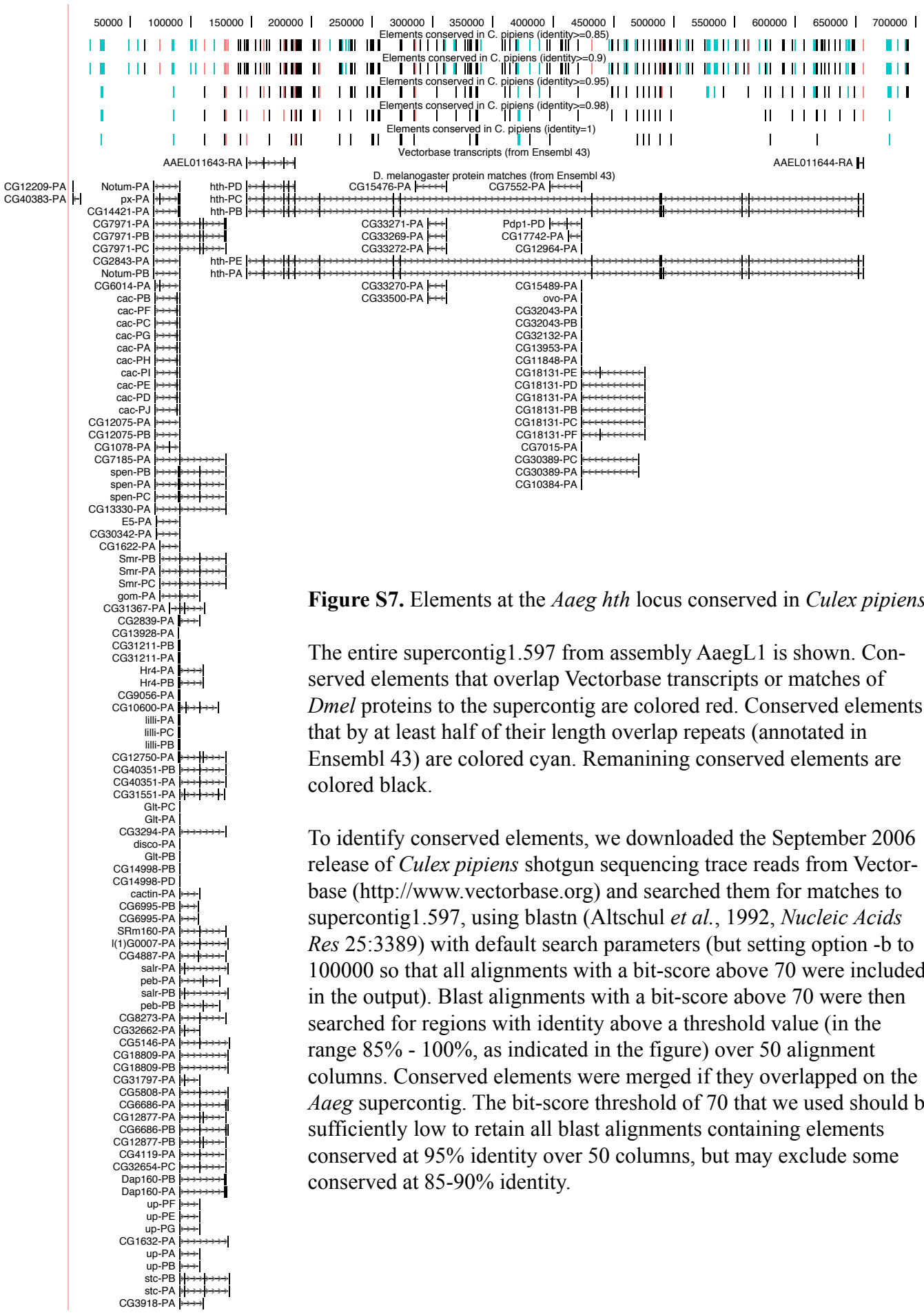

Figure S7. Elements at the *Aaeg hth* locus conserved in *Culex pipiens*

The entire supercontig1.597 from assembly AaegL1 is shown. Conserved elements that overlap Vectorbase transcripts or matches of *Dmel* proteins to the supercontig are colored red. Conserved elements that by at least half of their length overlap repeats (annotated in Ensembl 43) are colored cyan. Remanining conserved elements are colored black.

To identify conserved elements, we downloaded the September 2006 release of *Culex pipiens* shotgun sequencing trace reads from Vectorbase (http://www.vectorbase.org) and searched them for matches to supercontig1.597, using blastn (Altschul *et al.*, 1992, *Nucleic Acids Res* 25:3389) with default search parameters (but setting option -b to 100000 so that all alignments with a bit-score above 70 were included in the output). Blast alignments with a bit-score above 70 were then searched for regions with identity above a threshold value (in the range 85% - 100%, as indicated in the figure) over 50 alignment columns. Conserved elements were merged if they overlapped on the *Aaeg* supercontig. The bit-score threshold of 70 that we used should be sufficiently low to retain all blast alignments containing elements conserved at 95% identity over 50 columns, but may exclude some conserved at 85-90% identity.

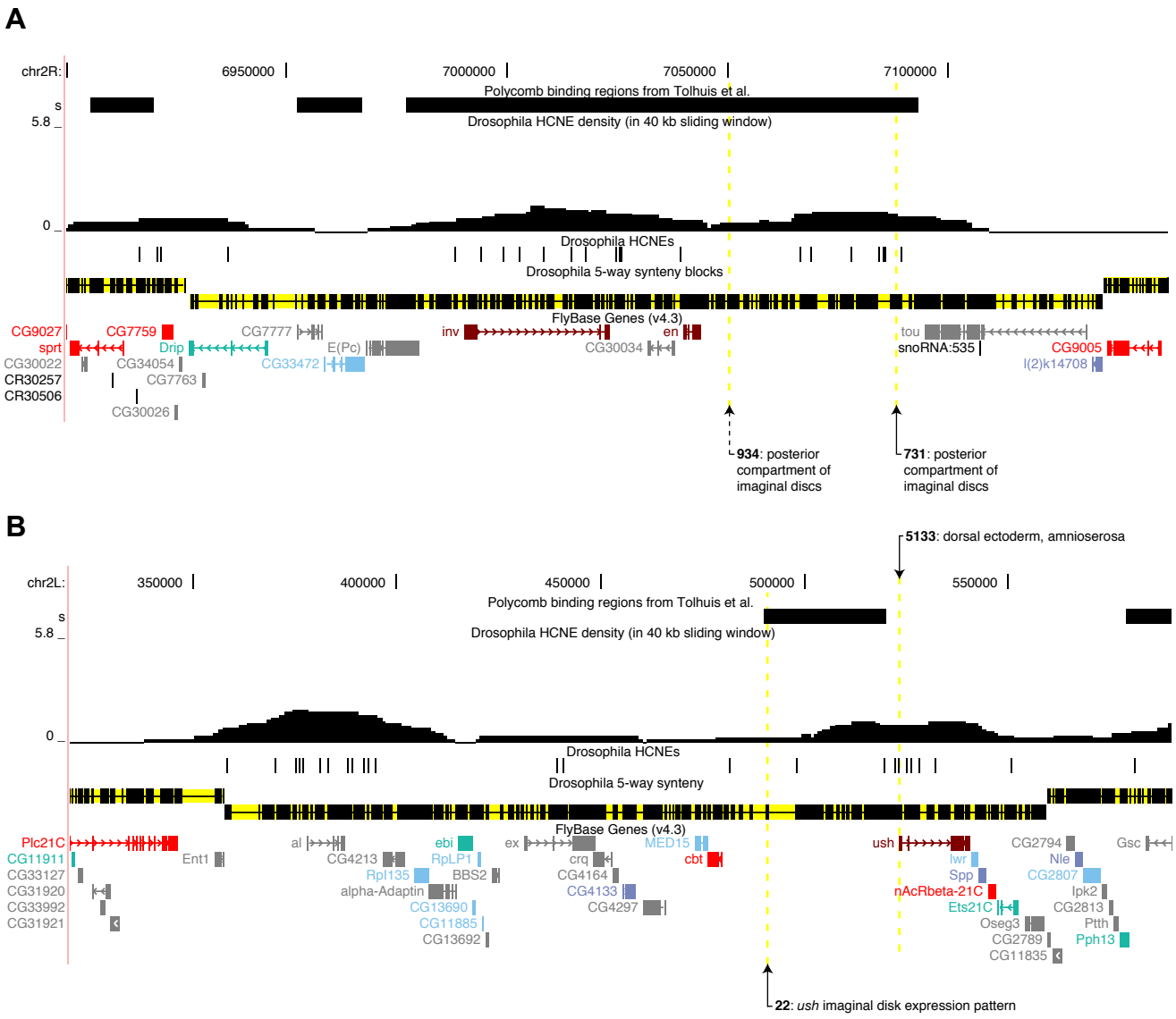

Figure S8. Enhancer trap insertions described by Hayashi et al.
Arrows indicate locations of enhancer trap insertions, and are labeled with strain number and expression pattern as described by Hayashi et al. (2002, *Genesis* 34:58). Gene models are colored by predicted core promoter type as in Figure 2.
(A) Insertion 731 is about 44 kb upstream of *engrailed* (*en*) and expressed in the posterior compartment of imaginal discs, similar to *en* (reviewed by Hidalgo, 1996, *Trends Genet* 12:1). On the contrary, the neighboring gene *toutatis* (*tou*) is expressed ubiquitously in wing imaginal discs (Vanolst et al., 2005, *Development* 132:4327). Insertion 934 has a similar expression pattern as insertion 731 although the two insertions are about 40 kb apart. We failed to find flanking sequence for insertion 934 in any database, and could therefore not determine its orientation and exact position; the dashed arrow indicates its approximate position based on Figure 2 in Hayashi et al. (2002).
(B) Insertion 22 is about 33 kb upstream of *u-shaped* (*ush*), on the opposite strand, and captures the expression pattern of *ush* in imaginal disks, but not in the embryo. Insertion 5133 is about 50 bp upstream of *ush*, on the same strand, and has an embryonic expression pattern similar to that of *ush* (Fosset et al., 2000, *PNAS* 97:7348).

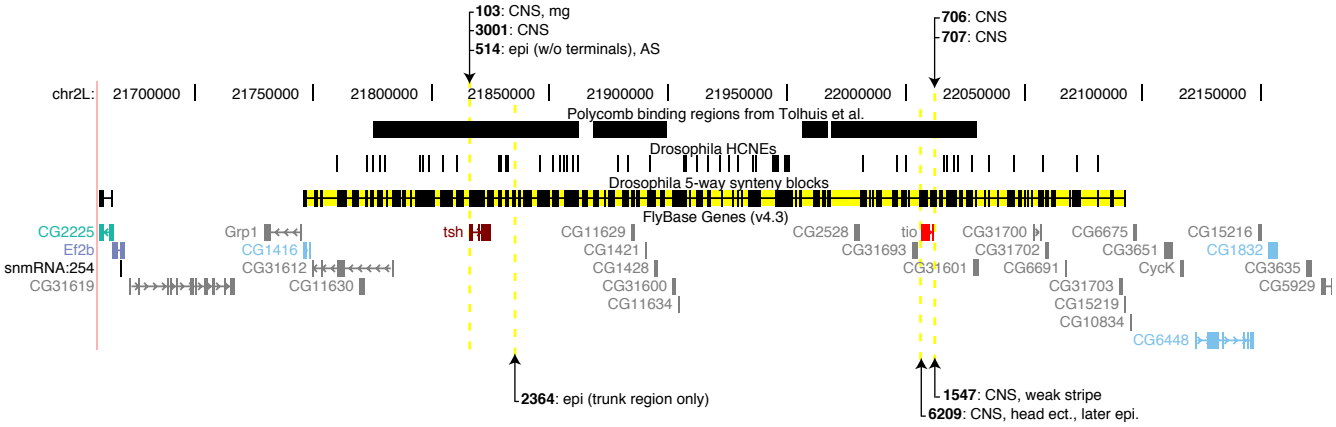

Figure S9. Enhancer trap insertions around *Dmel* genes *teashirt* (*tsh*) and *tiptop* (*tio*)
Arrows indicate locations of enhancer trap insertions in the same orientation as *tsh* and *tio* (arrows above panel) and the reverse orientation (arrows below panel), and are labeled with strain number and annotated embryonic expression pattern from GETDB (http://flymap.lab.nic.gac.jp). Gene models are colored by predicted core promoter type as in Figure 2. Insertions 103, 3001 and 514 are all within 300 bp of the annotated transcription start site for *tsh*. Insertion 2364 is about 10 kb downstream of *tsh* on the opposite strand. Consistent with the expression annotation for these insertions, *tsh* is expressed in the part of the trunk of the embryo that gives rise to the central nervous system (CNS) and epidermis, and in some other regions, including part of the visceral mesoderm around the midgut (Fasano et al., 1991, *Cell* 64:63). Insertion 6209 is about 350 bp upstream of *tio*. Insertions 706, 707 and 1547 are all about 200 bp downstream of *tio*. *tio* is expressed in parts of the embryonic CNS (Laugier et al., 2005, *Dev Biol* 283:446), consistent with the expression annotation for the surrounding insertions. For this figure, we combined the *tio* and *CR33987* gene models from FlyBase, because *CR33987* appears to contain the sequence for the first exon of *tio*.