Supplementary Material

This document contains the supplementary material for the document "An alignment-free method to identify candidate orthologous enhancers in multiple *Drosophila* genomes". It includes:

- Information on phylogenetic tree reweighting of completed *Drosophila* genomes with labeled branch lengths.
- Detailed results on the D/V dataset with results from alignment-free and alignment-based methods.
- Results on the shadow enhancer detection in non-melanogaster species.
- Scanning results of *eve* enhancers against Sepsid genomes.

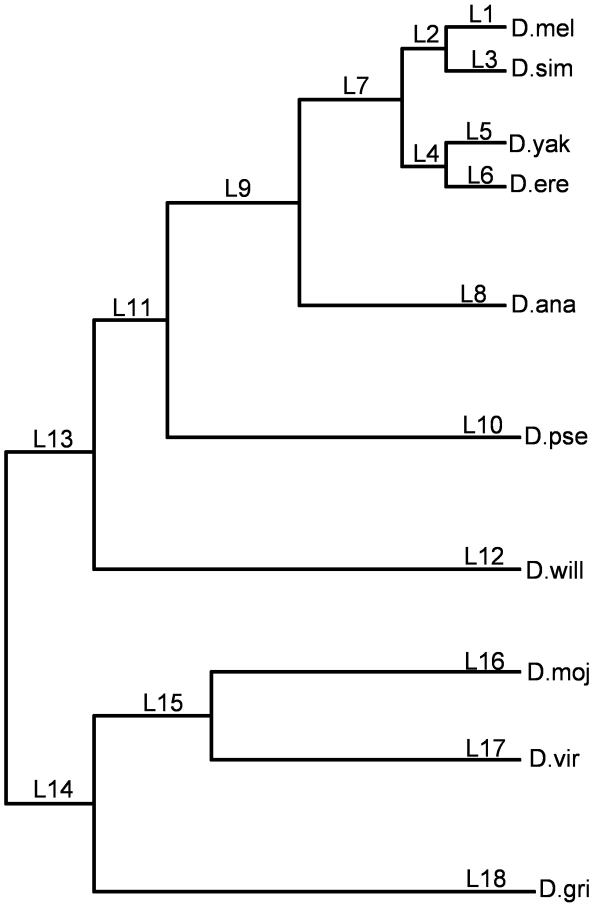
1. Phylogenetic tree re-weighted score (PRS) calculation

To take the evolutionary distance and relationship into account, we considered the available evolutionary distance information that was obtained from a set of 5067 orthologous genes identified using a tBlastn→Genewise→Blastp approach (http://www.danielpollard.com/tree.html). We first normalized all branch lengths used in a particular combination of species to sum up to 1. Figure (1) shows the tree and labeled branches for the completed *Drosophila* genomes (not including *D. persimilis* and *D. sechellia* which are very closely related to *D. pseudoobscura* and *D. simulans*, respectively). We then compute a phylogenetic-tree-reweighting score (PRS) based on the normalized path length between the two compared species. As examples, the pair-wise PRS between *D. melanogaster* and *D. pseudoobscura* is obtained by,

$$PRS(D.mel \leftrightarrow D.pse) = [((L1 + L2 + L7)/6) + (L9/5) + L10)] \qquad (1)$$

where L1, L2 and L7 are the common paths for all 6 diverged (non-melanogaster-subgroup) species, and their contribution is therefore normalized by all 6 species considered for the analysis. The path L9 is traversed by the 5 non-melanogaster-subgroup species (i.e. except *D. ana*) and so L9 is normalized by 5. L10 is the path traversed only by *D. pse*.

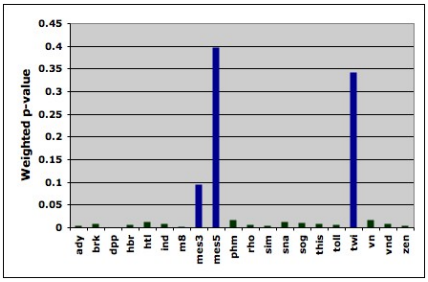
Figure 1: Phylogenetic tree of *Drosophila* species with indicated branch lengths.



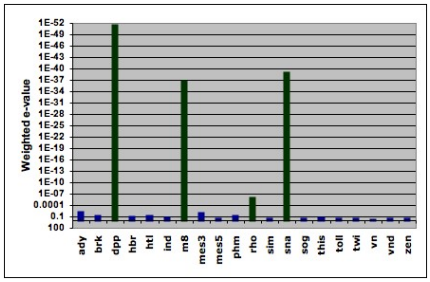
2. Results of orthologous enhancer detection of Dorsal Ventral enhancers of *Drosophila*

A set of 20 enhancer sequences that control dorsal-ventral (D/V) patterning of the *Drosophila* embryo were taken from a previous study (Papatsenko *et al.,* 2005). These enhancers are directly regulated by different concentration of Dorsal, which is a sequence-specific transcription factor that is distributed in a broad nuclear gradient across the D/V axis of the early *Drosophila* embryo. Together with this data set, orthologous sequences of 18 enhancers have been identified in evolutionarily diverged species like *D. pseudoobascura, D. virilis*, and *D. mojavensis*. The positions of enhancers in the orthologous loci were annotated through local alignment procedures, and the positions of conserved blocks in

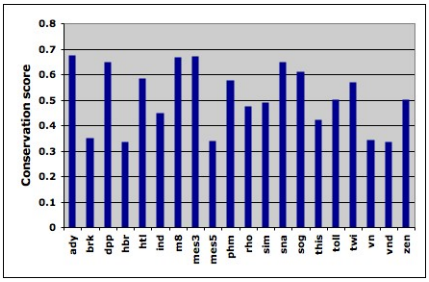
Figure 2: Performance of different methods in detecting enhancers based on conservation across multiple species. (a) Alignment-free approach; (b) BLAST; (c) phastCons. Predicted candidates that exceeded significance threshold are indicated in green bars and failed candidates are represented in blue.



(a)



(b)



(c)

enhancers were mapped using the motif extraction algorithm MEME (Bailey *et al.,* 1995). These enhancers therefore constitute a dataset with known functionality in *D. melanogaster* and manually curated orthologous regions; they are therefore useful to evaluate whether our approach can identify the correct locations of orthologous enhancer regions without any knowledge of binding sites. This dataset does not contain negative examples, but the locations of enhancers in other genomes has been manually curated.

Applying our method to 6 non-melanogaster genomes, we obtained the results shown in Figure (2a). For 17 candidates, the combined scores exceeded the significance threshold. However, we failed to detect enhancers for Mes5, Mes3 and Twi, possibly because their sequence lengths are shorter than 250 bp. We compared these results to the alignment-based methods (BLAST and phastCons) and the results are shown in Figure (2b,2c). As observed from the A/P patterning data (discussed in the main paper), these methods largely failed to assign significant similarity scores to these enhancers. We were therefore successful at pinpointing the location of enhancers in diverged genomes, and the success rate indicates a very good performance on enhancers of typical size.

3. Results on the shadow enhancer detection in non-melanogaster species

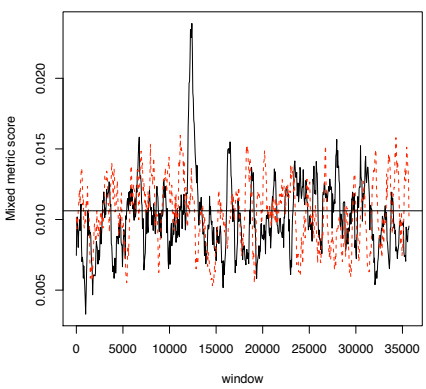
In addition to the primary D/V enhancers, we have also applied our method to analyze so-called "shadow enhancers". A recent study found that Dorsal target genes are regulated by secondary "shadow" enhancers (Hong *et al.,* 2008), often located in a distal genomic location. A set of genes (brk, htl, sog, vnd) that contains two enhancers for expression in the D/V axis were considered. We first identified orthologous intergenic regions from non-melanogaster genomes (*D. ana, D. pse, D. wil, D. moj, D. vir, and D. gri*). We then scanned the known primary and shadow enhancer along the corresponding orthologous intergenic region. For enhancers htl_primary, htl_shadow, vnd_primary, we were not able to identify the orthologous flanking genes in the more distantly related genomes of *D. wil, D. moj, D. vir, D. gri*. We could make predictions with significant p-value for the other enhancers (Shown in table 1). The borderline candidates are indicated in red color. Our result suggests that our method is capable of identifying the shadow enhancers as well.

Table 1: Evaluation on the shadow enhancers. The first column shows the enhancer name, the corresponding p-values of the non-melanogaster species are shown in the following columns.

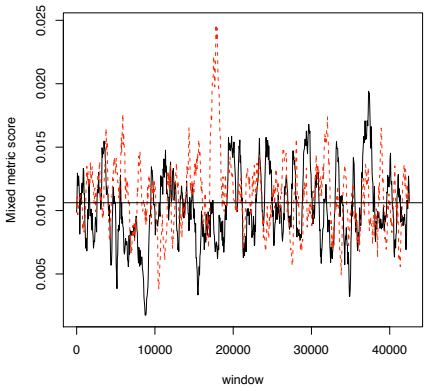
Enhancers	**D.ana**	**D. pse**	**D. will**	**D. moj**	**D. vir**	**D. gri**
brk_primary	0.0276	0.0038	0.002	0.0056	0.0026	0.0059
brk_shadow	0.0071	0.0062	0.0029	0.0046	0.0038	0.0037
htl_primary	0.0033	0.0031	0.0029	-	-	0.009
htl_shadow	0.0438	0.0038	0.003	0.0055	0.0035	-
sog_primary	0.0682	0.0642	0.0872	0.0121	0.0076	0.0354
sog_shadow	0.0049	0.0062	0.0044	0.0051	0.004	0.0029
vnd_primary	0.0306	0.0363	-	0.0021	0.0021	0.0026
vnd_shadow	0.009	0.0049	0.0021	0.003	0.0026	0.0024

4. Scanning result of Eve enhancers against Sepsid genomes

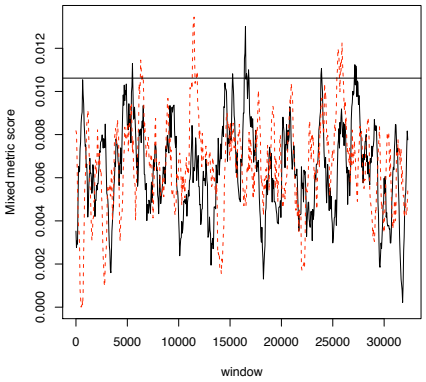
Figure 3: Scanning examples of Eve_stripe_3+7, Eve_stripe_4+6 and Eve_stripe_2 against Sepsid genomes. (a) Eve_stripe_3+7 scanned against *Themira putris* (b) Scanning result of Eve_stripe_4+6 against *Sepsis cynipsea* (c) Eve_stripe_2 searched against *Dicranosepsis spp.*



(a)



(b)



(c)