Formation of Regulatory Modules by Local Sequence Duplication

Armita Nourmohammad and Michael Lässig

Supporting Figures S1 - S3

Figure S1: **Motif detection in sequence segments (schematic).** The figure shows a configuration of correlated sequence sites of length $\ell = 10$ bp and distance $r = 14$ bp from each other. Pairs of correlated sites have the following properties: (i) The average mutual similarity between aligned nucleotides is larger than a given threshold, $c \geq c_{\min} = 0.8$. (ii) The left sites (and, hence, also the right sites) of all pairs have no common nucleotides. This condition is necessary in order to avoid overcounting of mutual similarity in overlapping site pairs. (iii) The sum of the mutual similarities of all pairs in the set is maximal. In the example shown, there are three different motifs with reoccurring sequence patterns marked by different colors (red, blue, green). To illustrate the alignment of the site pairs, we shift the whole sequence by $r = 14$ bp in the second row. The left and right site of each motif are shown in boldface in the first and the second row, respectively. Mismatches between aligned sites of the same motif are shown in boldface gray letters. The flanking regions separating the correlated sequence pairs are shown in smaller font.

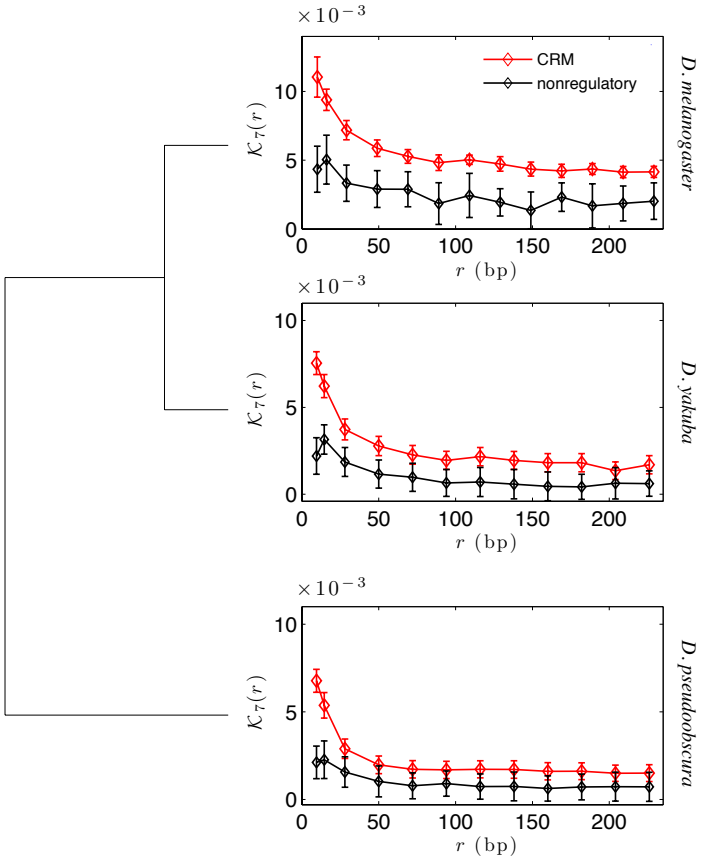

Figure S2: **Sequence similarity in regulatory modules of 3 *Drosophila* species.** Distance-dependent similarity information $K_7(r)$ for motif length $\ell = 7$ in regulatory modules (red) and in generic intergenic sequence (black), evaluated in *D. melangaster* and in the homologous regions of *D. yakuba* and *D. pseudoobscura* (see Materials and Methods). These data show a consistent pattern of overall amplitudes and of decay lengths.

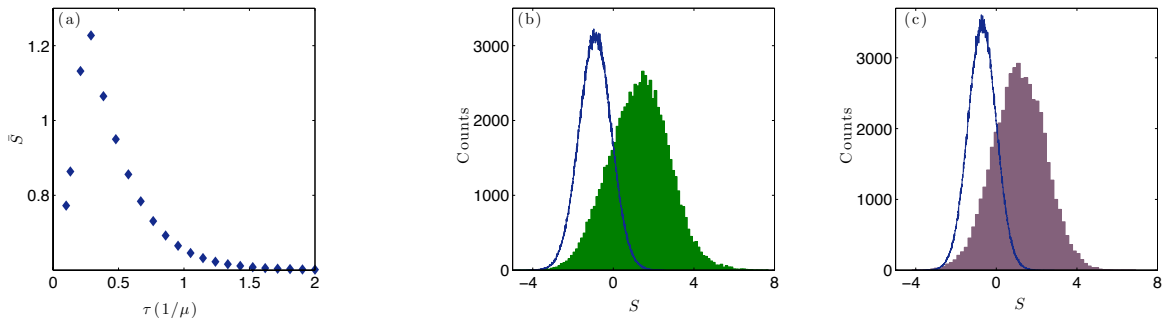

Figure S3: **Tests of the duplication inference method.** We simulate binding site pairs (**a**, **b**) evolving by common descent or by independent descent, as described in Materials and Methods. (a) Dependence of the total duplication score Σ on the time parameter τ for an ensemble of 150000 site pairs of common descent. This function has a pronounced maximum at a value $\tau_{\rm ML} \approx 0.3/\mu$, which is close to the mean divergence time $\bar{\tau} = 0.4/\mu$ since duplication. (b) Distributions of the score S (with $\tau = \tau_{\rm ML}$) for pairs of sites binding different factors. The distribution for sites of common descent (filled curve) is distinguished from the distribution for sites with independent descent (solid curve) by its increased score average, $\langle S \rangle - \langle S \rangle_0 = 2.1$, and by its increased width. (c) Same as (b) for pairs of sites binding the same factor. The distribution for sites of common descent (filled curve) has again an increased average, $\langle S \rangle - \langle S \rangle_0 = 1.6$, and an increased width.