SI Appendix
Genome-wide identification of cis-regulatory motifs and modules underlying gene co-regulation using statistics and phylogeny

Hervé Rouault [*], **Khalil Mazouni** [† ‡], **Lydie Couturier** [† ‡], **Vincent Hakim** [*] and **François Schweisguth** [† ‡]

[*]Laboratoire de Physique Statistique, CNRS, Université Pierre et Marie Curie, École Normale Supérieure, 75231, Paris Cedex 05, France,[†]Institut Pasteur, Developmental Biology Dept., F-75015 Paris, France, and [‡]CNRS, URA2578, F-75015 Paris, France
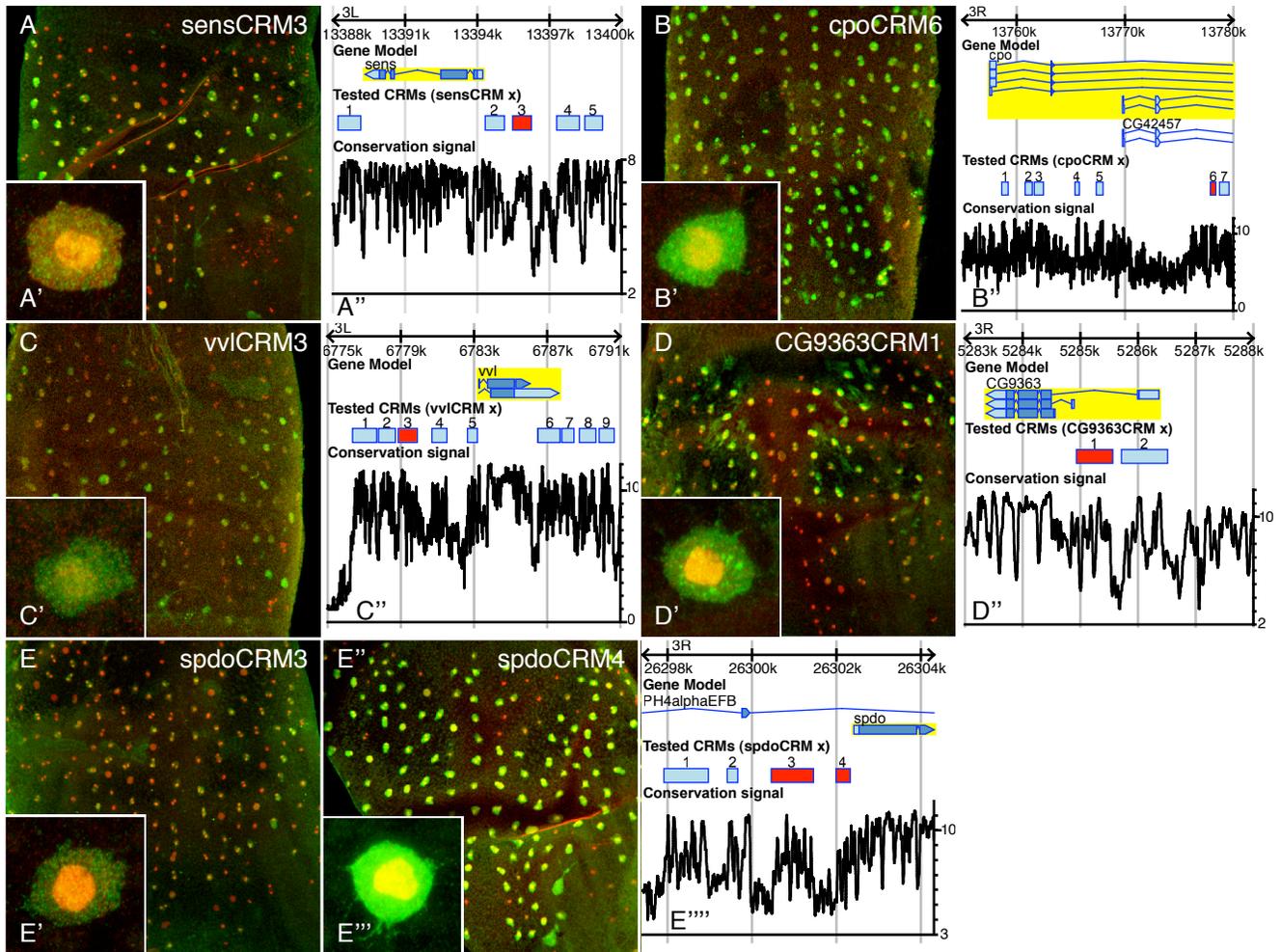
Fig. S1. Conservation-based identification of new SOP-specific CRMs Six novel CRMs were identified in the 20 kb region centered around the transcription start site of five selected genes known to be specifically expressed in SOPs of the pupal notum. The DNA fragments to be tested as CRMs were defined based on sequence conservation across the genome of 12 *Drosophila* species (bottom panels). Genome views showing exon/intron gene structure (gene model), position of the tested fragments and the conservation conservation signal. Genomic fragments with SOP-specific CRM activity are shown in red. CRM activity was monitored using a *lacZ* reporter gene. Cytoplasmic β-Galactosidase, green; nuclear Cut (red) as a SOP marker; DAPI in blue in high magnification views. Note that some SOPs have divided (as indicated by pairs of Cut-positive nuclei). A-A": sensEnh3 B-B": cpoEnh6 C-C": vvlEnh3 D-D": CG9363Enh1 E-E"": SpdoEnh3,SpdoEnh4
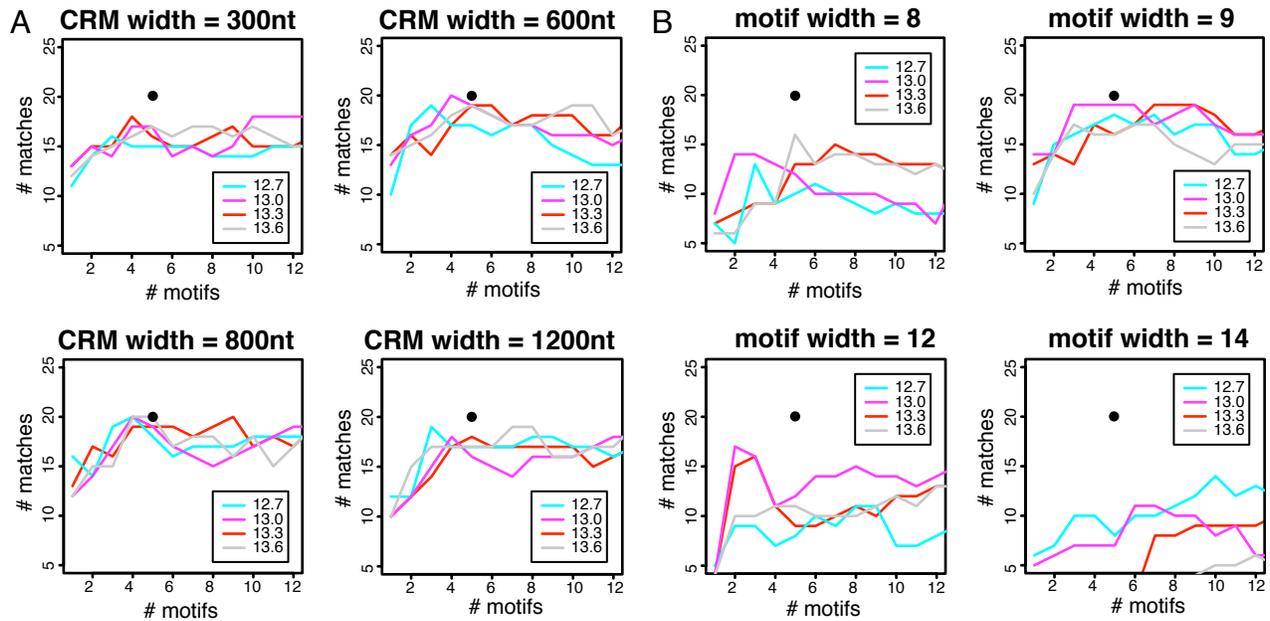
Fig. S2. Selection of optimal motif width and CRM width Similarly to Fig. 3-A, the number of predicted CRMs associated with a gene annotation related to sensory organs (number of matches in the y axis for the 100 top-ranked fragments; see section 3.3 of the supporting text) was plotted as a function of the number of motifs (1 to 12; x axis) for different S_{th} values (from 12.7 to 13.6). A. The curves are plotted for different CRM widths : 300, 600, 800, 1200 and for the same motifs as in Fig. 3-A. B. The curves are plotted for different motif widths : 8, 9, 12, 14 and for CRMs of width 1000nt. The solid black circles in each figure denote the number of matches obtained for the parameters chosen for the experimental validation.
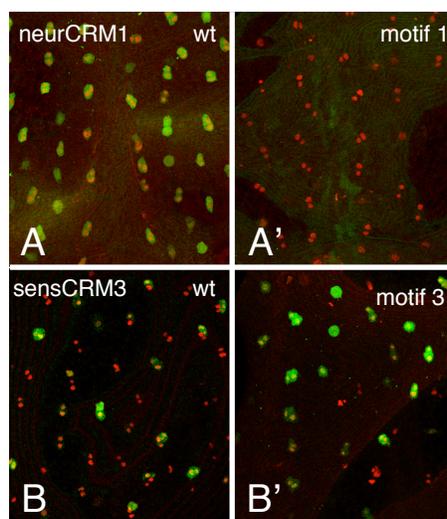

Fig. S3. *In vivo* analysis of motifs 1 and 3 Site directed mutagenesis of the two sites detected as motif 1 in neurCRM1 strongly reduced the activity of this CRM in trangenic flies (A,A'). In contrast, mutagenesis of the three sites detected as motif 3 in sensCRM3 did not detectably change the activity of this CRM (B,B'). CRM activity was monitored in 17 hours APF pupae by anti-β-galactosidase antibody staining (green). Cut (red) was used as a nuclear marker for SOPs and its progeny cells.
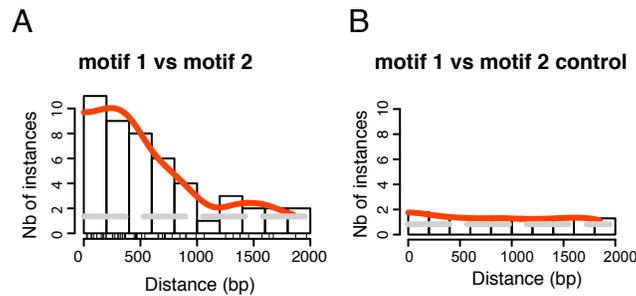
Fig. S4. Cross-correlation between motif 1 and 2 A. Genome-wide cross-correlation of conserved instances of motifs 1 and 2 in the *D. melanogaster* genome. The number of motif pairs (boxes) was plotted as a function of the distance (x axis) between the two motifs (bin size $=200$bp). The red curve was obtained by smoothing the histogram with a gaussian (width$=150$bp). The dashed line represented the average number of instances at very long distances. The co-occurrence of motifs 1 and 2 was shown by the histogram peak around zero. B. To assess the significance of the cross-correlation peak in A, we computed the cross-correlations between the original matrix 2 and 150 randomized versions of matrix 1 obtained by randomly shuffling its columns. The average cross-correlation of matrix 2 with the randomized versions of matrix 1 was displayed on the graph. The average (over the 150 randomized cross-correlations) difference in site number between the first bin $(0-200$ bp) and the 3 last bins $(1400-2000$ bp) was 0.47 with a standard deviation of 1.49. The distribution fitted well a gaussian and we did not observe values above 6.0. This led us to very conservatively estimate that $p < 0.005$ and to conclude that the observed correlated appearance of binding sites for matrix 1 and 2 was highly significant. Of note, while the binding site density was found to be comparable for matrix 1 and its randomized analogs, matrix 2 was found to have much more binding instances than its randomized versions. This potential bias prevented us from computing control correlations with randomized versions of motif 2.
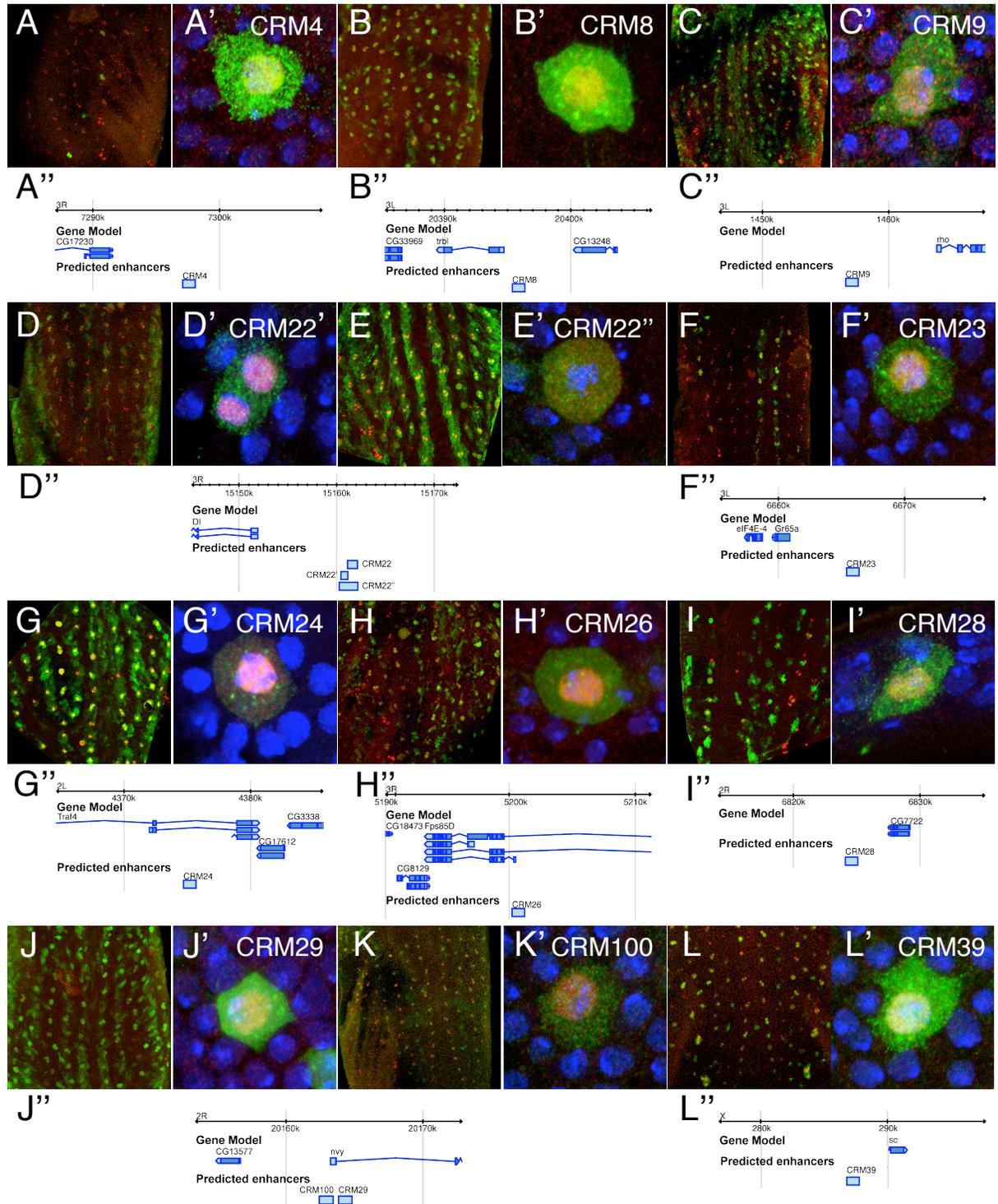
Fig. S5. *In vivo* analysis of predicted CRMs in the pupal notum Predicted CRMs were tested for their regulatory activity in the notum of 16-18hours APF transgenic pupae (all positive CRMs are shown, with the exception of CRM17). β-galactosidase expression is shown in top panels. For each CRM, a low and a high-magnification view is shown: Cytoplasmic β-galactosidase, green; nuclear Cut (red) as a SOP marker; DAPI in blue in high magnification views. The genomic position of the CRM is indicated by a blue box in the corresponding bottom panel. Eleven out of the 29 top-ranked CRMs directed expression in SOPs: CRM4 (A-A") CRM7 (Fig. 2), CRM8 (B-B"), CRM9 (C-C"; expression extended to PNCs), CRM20 (Fig. 4), CRM23 (F-F"), CRM24 (G-G"; expression extended to PNCs), CRM26 (H-H"), CRM28 (I-I"; expression was not strictly restricted to SOPs) and CRM29 (J-J"; note that expression extended to PNCs). Five additional CRMs were active in SOPs: two, CRM40 (Fig. 4) and CRM100 (K-K") were tested because they were found close to a functionally validated CRM, CRM20 and CRM29, respectively; three others, CRM22'/CRM22" (D-E'), CRM39 (L-L') and CRM41 (Fig. 2) were tested because they were located close to genes expressed in PNCs and up-regulated in SOPs, i.e. *Delta*, *scute* and *scabrous*, respectively. While the 1000 nt fragment tested as CRM22 was not active in our reporter assay, a larger 2.1 kb fragment encompassing CRM22, and referred to here as CRM22", was active in PNCs (D", E and E'). CRM22" also encompassed another 1000 nt fragment with a high score in our CRM prediction test. This fragment, noted here CRM22', was also active, albeit more weakly, in SOPs and PNCs. This is consistent with the notion that CRM22 contains some cis-regulatory information that contributes to the activity of CRM22".
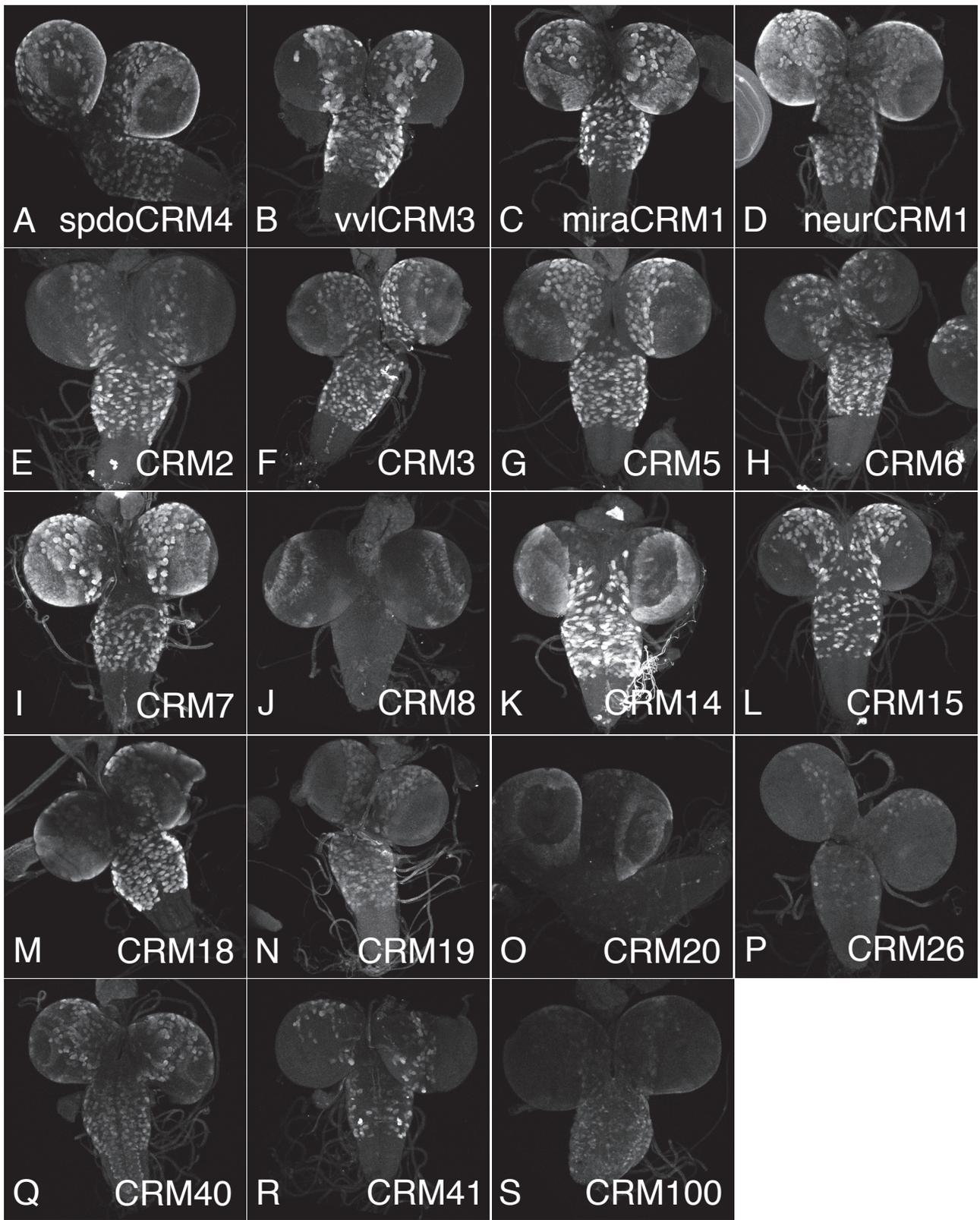
Fig. S6. *In vivo* analysis of predicted CRMs in the larval brain Predicted CRMs were tested for their regulatory activity in larval brain of third instar larvae. Several CRMs from the SOP training set, including spdoCRM4 (A), vvlCRM3 (B), miraCRM1 (C) and neurCRM1 (D) were active in larval neuroblasts. Thirteen of the 29 top-ranked CRMs were also directing β-galactosidase expression in neuroblasts. These included CRM2 (E), CRM3 (F), CRM5 (G), CRM6 (H), CRM7 (I), CRM (8), CRM14 (K), CRM15 (L), CRM18 (M), CRM19 (N), CRM20 (O), CRM24 (P) and CRM26 (Q). Additionally, three additional CRMs active in SOPs are also active in neuroblasts: CRM40 (Q), CRM41 (R) and CRM100 (S).
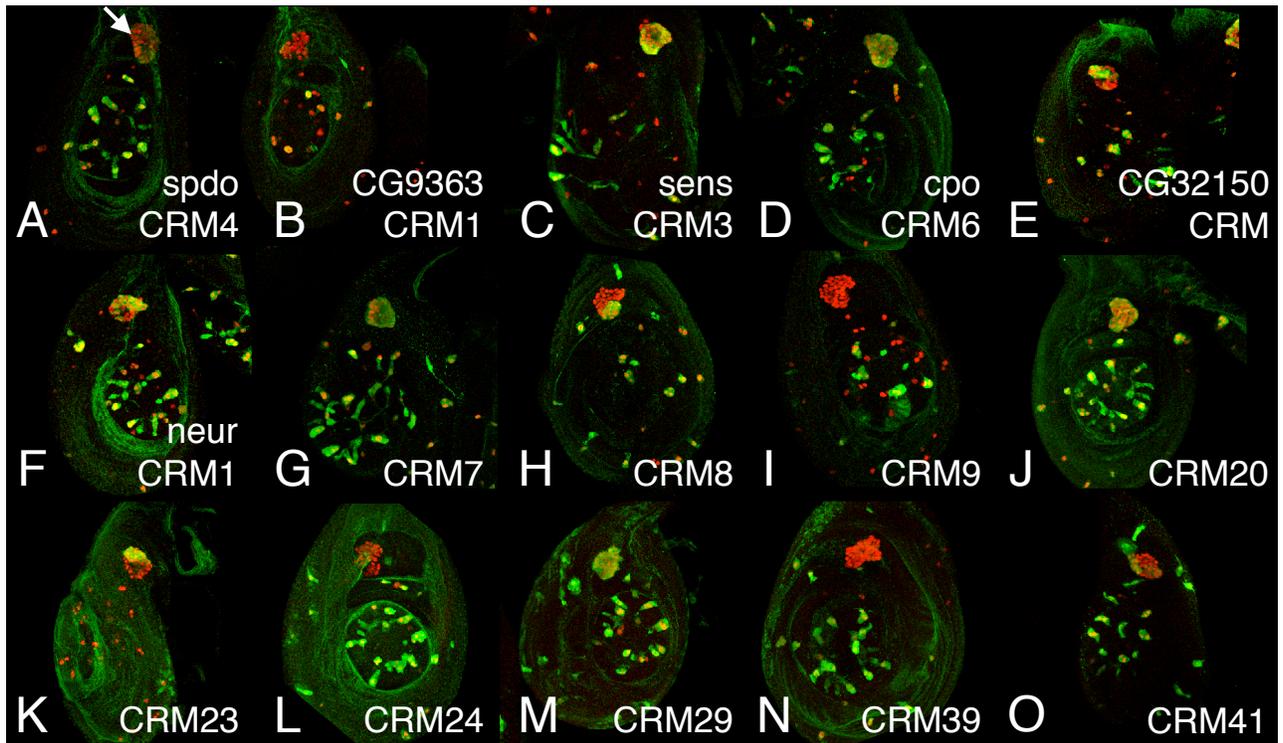
Fig. S7. *In vivo* **analysis of predicted CRMs in chordotonal SOPs of leg imaginal discs** The regulatory activity of CRMs from the SOP training set (A-F) and of predicted CRMs (G-O) was tested in in leg imaginal discs dissected from third instar larvae: Cytoplasmic β-galactosidase, green; Sens (red) as a SOP marker. Most CRMs active in SOPs of the pupal notum were also active in chordotonal (ch)-SOPs (arrow in A). A few CRMs, including CRM9 (J) and CRM39 (O), were active in External (E)-SOPs, that generate external sense organs, but not in ch-SOPs. In contrast with pupal notum E-SOPs that are specified by the proneural factors Achaete and Scute, ch-SOPs are specified by the proneural bHLH factor Atonal. Thus, CRMs active in both E-SOPs and Ch-SOPs are likely to be, directly or indirectly, regulated by both Atonal/Da and Ac(or Sc)/Da heterodimers (1). Since motif 2 of SensEnh3 can interact with Ato/Da, Ac/Da and Sc/Da heterodimers (2), regulation can be direct for all six CRMs expressed in both E-SOPs and Ch-SOPs that contain one to three copies of motif 2 (see Table S8).
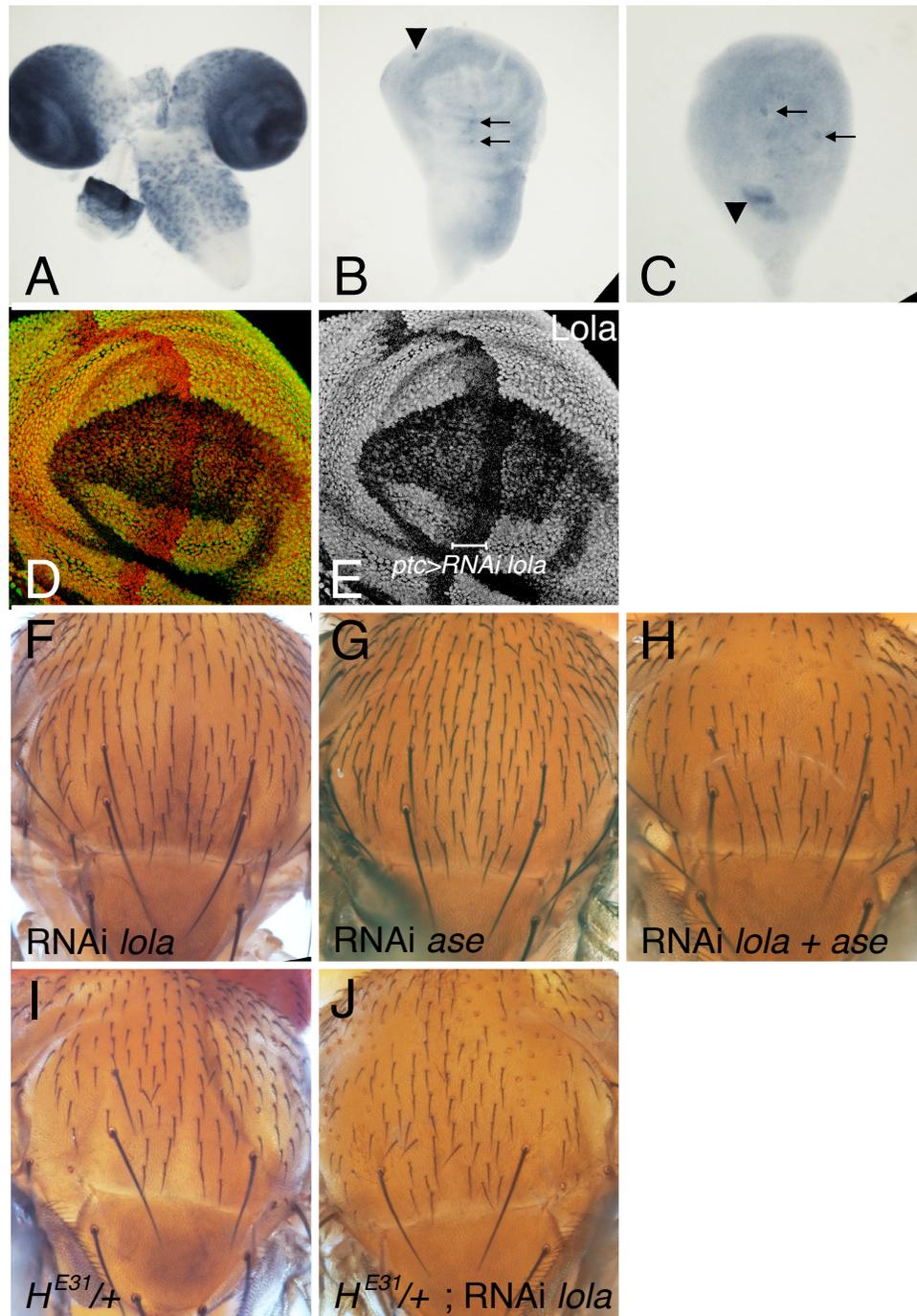
Fig. S8. *lola* **supporting characterization** (A-C) *In situ* hybridization analysis of *lola* transcript accumulation. *lola* transcripts were detected in neuroblasts of third instar larval brain (A) as well as in all cells of wing and leg imaginal discs. *lola* transcripts appeared to be more abundant in both E-SOPs (arrows) and ch-SOPs (arrowhead) of wing and leg imaginal discs. (D-E) Experimental validation of anti-Lola antibodies and RNAi-mediated inactivation of the *lola* gene. Immunostaining of a wing imaginal disc expressing a UAS-dsRNA construct against *lola* under the control of patched (ptc)-GAL4. The signal detected by the anti-Lola antibodies (green; DAPI in red) was very strongly reduced in ptc-GAL4 expressing cells (indicated by a bar), indicating that anti-Lola antibodies specifically recognized Lola and that the lola dsRNA construct efficiently down-regulated *lola* gene expression. (F-J) *lola* genetically interacts with *asense* and *Hairless*. RNAi-mediated inactivation of *lola* using Eq-GAL4 at 25°C had little effect on bristle development, with only a few bristles potentially missing (arrow in F). Similarly, RNAi-mediated inactivation of *asense* using Eq-GAL4 Gal80ts at 29°C did not result in a detectable bristle phenotype (G). In contrast, concomittant inactivation of the *lola* and *asense* genes using Eq-GAL4 Gal80ts at 29°C resulted in a strong bristle loss (H). Additionally, while the loss of a single copy of the *Hairless* gene had no significant effect on microchaete development (I), RNAi-mediated inactivation of *lola* using Eq-GAL4 at 25°C in H^{E31} heterozygous flies had a strong effect on bristle development, with many microchaetes showing a double-socket phenotype (J). This phenotype is indicative of a gain of Notch activity causing the transformation of shaft cells into socket cells (3).

Table S1. The SOP training set: validated CRMs.

Id.	Coordinate			Neighboring SOP	Source
	chromosome	start	stop	specific gene	
CG32150CRM	3L	15839629	15840789	CG32150	Reeves *et al* (4)
chnCRM	2R	11019807	11020918	*charlatan (chn)*	"
miraCRM	3R	15756362	15757274	*miranda (mira)*	"
PFECRM	2L	18013214	18015611	*reduced ocelli (rdo)*	"
neurCRM1	3R	4850827	4850970	*neuralized (neur)*	Gomes *et al* (5)
neurCRM2	3R	4852116	4853004	"	"
phylCRM1	2R	10320543	10322141	*phyllopod (phyl)*	Pi *et al* (6)
phylCRM2	2R	10322623	10324621	"	"
CG9363CRM1	3R	5284935	5285565	CG9363	this study
spdoCRM3	3R	26300460	26301460	*sanpodo (spdo)*	"
spdoCRM4	3R	26301990	26302330	"	"
cpoCRM6	3R	13777879	13778379	*couch potato (cpo)*	"
vvlCRM3	3L	6778859	6779909	*ventral vein lacking (vvl)*	"
sensCRM3	3L	13395475	13396245	*senseless (sens)*	"

Coordinates of the 14 validated SOP CRMs in our training set. The given coordinates correspond to the *D. melanogaster* genome assembly v. 5.

Table S2. The SOP training set: conserved sequences close to some SOP genes.

Id.	Coordinate			Neighboring SOP
	chromosome	start	stop	specific gene
CG9363CRM2	3R	5285715	5286515	CG9363
CG32392CRM1	3L	6757396	6758246	CG32392
spdoCRM1	3R	26297910	26298960	*sanpodo (spdo)*
spdoCRM2	3R	26299410	26299660	"
cpoCRM1	3R	13758629	13759229	*couch potato (cpo)*
cpoCRM2	3R	13760779	13761429	"
cpoCRM3	3R	13761629	13762479	"
cpoCRM4	3R	13765379	13765779	"
cpoCRM5	3R	13767329	13767979	"
cpoCRM7	3R	13778729	13779579	"
vvlCRM1	3L	6776359	6777679	*ventral vein lacking (vvl)*
vvlCRM2	3L	6777779	6778709	"
vvlCRM4	3L	6780709	6781529	"
vvlCRM5	3L	6782639	6783179	"
vvlCRM6	3L	6786509	6787709	"
vvlCRM7	3L	6787809	6788459	"
vvlCRM8	3L	6788759	6789659	"
vvlCRM9	3L	6789839	6790659	"
svCRM1	4	1108593	1109363	*shaven (sv)*
svCRM2	4	1109593	1110093	"
svCRM3	4	1110993	1111443	"
insvCRM1	2L	2575086	2575406	*insensitive (insv)*
insvCRM2	2L	2576496	2576756	"
insvCRM3	2L	2576906	2577256	"
sensCRM1	3L	13388205	13389155	*senseless (sens)*
sensCRM2	3L	13394325	13395125	"
sensCRM4	3L	13397295	13398245	"
sensCRM5	3L	13398475	13399205	"
chnCRM1	2R	11000170	11001220	*charlatan (chn)*
chnCRM2	2R	11015320	11015670	"
chnCRM3	2R	11022520	11023420	"

Coordinates of the 31 sequences in our SOP training set that were chosen on the basis of their conservation and their proximity to known SOP but that did not direct reporter gene expression in SOPs. The given coordinates correspond to the *D. melanogaster* genome assembly v. 5.

Table S3. The PNC training set.

Id.	Coordinate			Neighboring PNC specific gene	Source
	chromosome	start	stop		
malphaCRM	3R	21835602	21836613	*E(spl) region transcript mα (mα)*	B. Castro *et al* (7)
EsplCRM	3R	21864872	21865973	*Enhancer of split (E(spl))*	"
HLHm5CRM	3R	21855458	21856354	*E(spl) region transcript mα (HLHm5)*	M. Lecourtois and F. Schweisguth (8)
m4CRM	3R	21850216	21850717	*E(spl) region transcript m4 (m4)*	A. M. Bailey and J. W. Posakony (8)
BrdCRM	3L	14964319	14965768	*Bearded (Brd)*	A. Singson *et al* (9)
edlCRM	2R	14558811	14560190	*ETS-domain lacking (edl)*	N. Reeves and J. W. Posakony (4)
traf4CRM	2L	4374718	4375544	*TNF-receptor-associated factor 4 (Traf4)*	"
sizCRM	3L	21059048	21060958	*schizo (siz)*	"

Coordinates of the sequences that compose our PNC training set. The given coordinates correspond to the *D. melanogaster* genome assembly v. 5.

Table S4. Predicted SOP motifs.

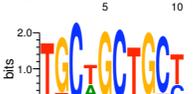| Starting site	Score	χ^2 score	Site density on the training set	Site density on the background	Logo
CAACCCCTAT	68.7	16.2	6.89×10^{-4}	1.67×10^{-5}	
ATGGCAGCAG	63.7	652	1.17×10^{-3}	1.17×10^{-4}	
ACCGCGTGCC	49	7.57	5.52×10^{-4}	2.1×10^{-5}	
CAGCTGATGA	43.2	1.33	6.89×10^{-4}	6.41×10^{-5}	
CCGGCAGCCC	36.1	21.6	4.83×10^{-4}	7.07×10^{-5}	
AGGCGCAGCT	35.8	19	6.2×10^{-4}	9.94×10^{-5}	
CGAAAAAAAA	35.4	917	9.65×10^{-4}	4.22×10^{-4}	
CGCACCAAAC	31.7	6.37	4.14×10^{-4}	3.52×10^{-5}	
CGGTGGCAGC	31	6.99	5.52×10^{-4}	6.5×10^{-5}	
GCAAATCGCA	30.5	18.2	5.52×10^{-4}	8.9×10^{-5}	
GGGTGTCCTT	37.8	3.71×10^6	7.58×10^{-4}	4.2×10^{-4}	
TGCTCCTGCA	33.5	1.55×10^3	6.89×10^{-4}	2.4×10^{-4}	
GGGGTCCTCC	31.7	2.66×10^3	4.14×10^{-4}	4.47×10^{-5}	
CCCCCCCCCT	31.6	2.81×10^4	6.2×10^{-4}	1.73×10^{-4}	
AGGACGAGAC	31.5	4.42×10^6	4.83×10^{-4}	5.17×10^{-5}	
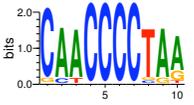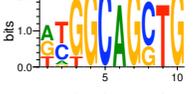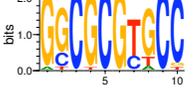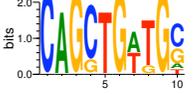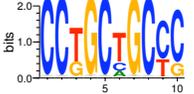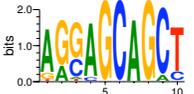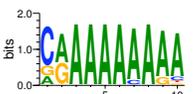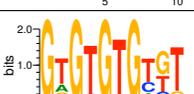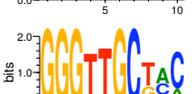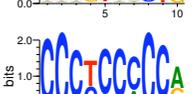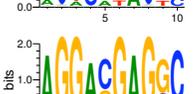
The first ten top-ranked motifs obtained with the SOP training set are displayed in the top part of the table. The five top-ranked motifs corresponding to repeated sequences are displayed in the bottom part of the table. The score column corresponds to the score of motifs defined in supporting text, section 2.5.3. The χ^2 score is defined in supporting text, section 2.5.2. The site densities correspond to a site detection threshold of $S_{\text{th}} = 13.3$.

Table S5. Matrices associated to the predicted SOP motifs.
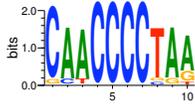

```
A [ 0.009   0.883   0.883   0.009   0.009   0.009   0.009   0.009   0.875   0.711 ]
C [ 0.914   0.096   0.007   0.981   0.981   0.981   0.981   0.057   0.006   0.007 ]
G [ 0.067   0.007   0.007   0.006   0.006   0.006   0.006   0.061   0.109   0.212 ]
T [ 0.009   0.011   0.102   0.009   0.009   0.009   0.009   0.875   0.009    0.07 ]
```

*
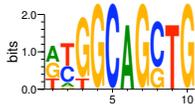

```
A [  0.38   0.094   0.004   0.003   0.003   0.996   0.003   0.004   0.003   0.003 ]
C [ 0.003   0.359   0.002   0.002   0.991   0.002   0.002   0.684   0.002   0.002 ]
G [ 0.371    0.05   0.844   0.991   0.002   0.002   0.991   0.308   0.002   0.991 ]
T [ 0.246   0.497   0.147   0.003   0.003   0.003   0.003   0.004   0.996   0.003 ]
```

*
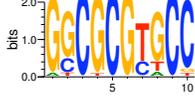

```
A [ 0.046   0.008   0.006   0.006   0.006   0.006   0.008   0.071   0.006   0.007 ]
C [ 0.004   0.222   0.981   0.004   0.981   0.004   0.272   0.006   0.981   0.896 ]
G [ 0.942   0.712   0.004   0.952   0.004   0.981   0.005   0.771   0.004   0.052 ]
T [ 0.006   0.059   0.006   0.039   0.006   0.006   0.715   0.151   0.006   0.045 ]
```

*
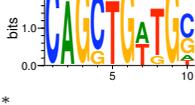

```
A [ 0.005   0.986   0.005   0.007   0.005   0.005   0.498   0.006   0.005   0.107 ]
C [ 0.991   0.003   0.003   0.787   0.003   0.003   0.005   0.004   0.003   0.478 ]
G [ 0.003   0.003   0.991   0.197   0.003   0.991   0.005   0.158   0.991   0.311 ]
T [ 0.005   0.005   0.005   0.007   0.986   0.005   0.493   0.832   0.005   0.104 ]
```

*
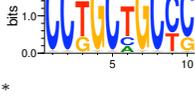

```
A [ 0.012   0.012   0.015   0.012   0.012   0.145   0.012   0.012   0.015   0.013 ]
C [ 0.971   0.971    0.01   0.008   0.971    0.19   0.008   0.971   0.671   0.779 ]
G [ 0.008   0.008   0.297   0.971   0.008   0.011   0.971   0.008    0.01   0.194 ]
T [ 0.012   0.012   0.678   0.012   0.012   0.654   0.012   0.012   0.305   0.013 ]
```

*
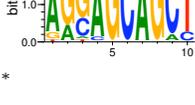

```
A [ 0.824   0.251   0.042   0.919   0.005   0.005   0.986   0.005   0.087   0.007 ]
C [ 0.004   0.004   0.302   0.075   0.004   0.991   0.004   0.004   0.905   0.172 ]
G [  0.13   0.741   0.613   0.004   0.981   0.004   0.004   0.981   0.004   0.004 ]
T [ 0.045   0.006   0.041   0.006   0.005   0.005   0.005   0.005   0.006   0.815 ]
```

*
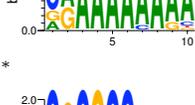

```
A [ 0.144   0.567   0.986   0.986   0.986   0.986   0.901   0.986    0.84   0.832 ]
C [ 0.671   0.005   0.005   0.005   0.005   0.005   0.087   0.005   0.005   0.081 ]
G [ 0.177   0.422   0.005   0.005   0.005   0.005   0.005   0.005   0.144   0.042 ]
T [ 0.007   0.007   0.007   0.007   0.007   0.007   0.007   0.007   0.007   0.044 ]
```

*
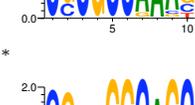

```
A [ 0.008    0.01   0.008   0.007   0.008   0.008   0.901   0.986   0.866    0.37 ]
C [ 0.981   0.181   0.981   0.005   0.981   0.981   0.005   0.005   0.062   0.202 ]
G [ 0.005   0.803   0.005   0.981   0.005   0.005   0.089   0.005    0.06   0.303 ]
T [ 0.007    0.01   0.007   0.007   0.007   0.007   0.007   0.007   0.008   0.124 ]
```

*
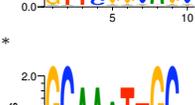

```
A [ 0.016   0.016   0.587   0.015   0.014   0.014   0.014   0.967   0.014   0.014 ]
C [ 0.819   0.011   0.011   0.129   0.009   0.009   0.961   0.009   0.009   0.961 ]
G [ 0.147   0.819   0.011    0.01   0.961   0.961   0.009   0.009   0.961   0.009 ]
T [ 0.016   0.154   0.391   0.849   0.014   0.014   0.014   0.014   0.014   0.014 ]
```

*
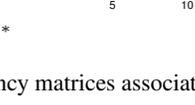

```
A [ 0.009   0.009   0.976   0.976   0.754   0.009   0.012   0.009   0.009   0.538 ]
C [ 0.006   0.981   0.006   0.006   0.008   0.006   0.211   0.006   0.981   0.268 ]
G [ 0.981   0.006   0.006   0.006   0.008   0.006   0.326   0.981   0.006   0.008 ]
T [ 0.009   0.009   0.009   0.009   0.231   0.976   0.451   0.009   0.009   0.185 ]
```

*

Position frequency matrices associated to the first ten top-ranked motifs obtained with the SOP training set.

Table S6. Predicted PNC motifs.

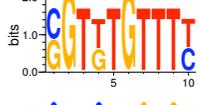| Starting site	Score	χ^2 score	Site density on the training set	Site density on the background	Logo
TGGGAGAAAC	63.3	4.86	1.1×10^{-3}	2.4×10^{-5}	
AAACAGCTGC	46.9	3.33	9.91×10^{-4}	3.86×10^{-5}	
ACCCAAAAAC	31.7	10.9	8.81×10^{-4}	1.02×10^{-4}	
ATGCGTGGGA	27.7	8.12	5.51×10^{-4}	3×10^{-5}	
CCTTTTACGC	25.5	0.787	4.41×10^{-4}	1.47×10^{-5}	
GATGTGTTTT	25.4	3.61	5.51×10^{-4}	4.28×10^{-5}	
CAACATGTGC	23.2	18.8	5.51×10^{-4}	3.83×10^{-5}	
CTTGGCTAGC	19	10.3	4.41×10^{-4}	5.3×10^{-5}	
GCGACAGCTG	18.7	12	4.41×10^{-4}	5.16×10^{-5}	
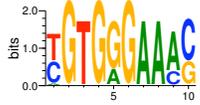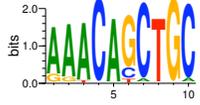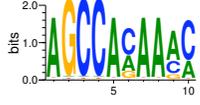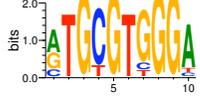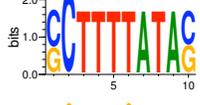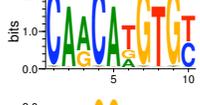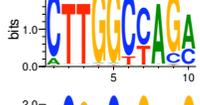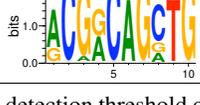
The first nine top-ranked motifs obtained with the PNC training set are displayed. The site densities correspond to a site detection threshold of $S_{\text{th}} = 13.3$.

Table S7. Matrices associated to the predicted PNC motifs.
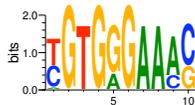
A	[0.043	0.005	0.005	0.005	0.244	0.005	0.986	0.986	0.731	0.006]
C	[0.427	0.003	0.003	0.003	0.004	0.003	0.003	0.003	0.259	0.638]
G	[0.004	0.991	0.003	0.991	0.749	0.991	0.003	0.003	0.004	0.348]
T	[0.526	0.005	0.986	0.005	0.006	0.005	0.005	0.005	0.006	0.006]

*
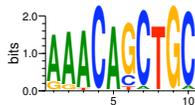
A	[0.832	0.901	0.947	0.006	0.986	0.008	0.047	0.006	0.006	0.054]
C	[0.005	0.004	0.004	0.981	0.004	0.159	0.942	0.004	0.004	0.933]
G	[0.152	0.091	0.004	0.004	0.004	0.771	0.004	0.004	0.981	0.004]
T	[0.008	0.006	0.048	0.006	0.006	0.062	0.006	0.986	0.006	0.007]

*
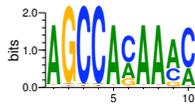
A	[0.957	0.017	0.017	0.017	0.957	0.308	0.957	0.957	0.611	0.294]
C	[0.011	0.011	0.952	0.952	0.011	0.522	0.011	0.011	0.263	0.677]
G	[0.011	0.952	0.011	0.011	0.011	0.151	0.011	0.011	0.109	0.012]
T	[0.017	0.017	0.017	0.017	0.017	0.019	0.017	0.017	0.017	0.018]

*
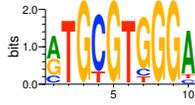
A	[0.482	0.009	0.009	0.008	0.009	0.009	0.01	0.009	0.009	0.815]
C	[0.158	0.006	0.006	0.819	0.006	0.006	0.111	0.006	0.006	0.081]
G	[0.35	0.006	0.981	0.006	0.981	0.006	0.779	0.981	0.981	0.007]
T	[0.01	0.976	0.009	0.164	0.009	0.976	0.099	0.009	0.009	0.092]

*
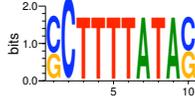
A	[0.011	0.01	0.01	0.01	0.01	0.01	0.976	0.01	0.976	0.012]
C	[0.583	0.971	0.007	0.007	0.007	0.007	0.007	0.007	0.007	0.541]
G	[0.393	0.007	0.007	0.007	0.007	0.007	0.007	0.007	0.007	0.435]
T	[0.011	0.01	0.976	0.976	0.976	0.976	0.01	0.976	0.01	0.012]

*
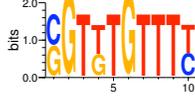
A	[0.012	0.01	0.01	0.013	0.01	0.01	0.01	0.01	0.01	0.012]
C	[0.492	0.006	0.006	0.009	0.006	0.006	0.006	0.006	0.006	0.437]
G	[0.483	0.971	0.006	0.418	0.006	0.971	0.006	0.006	0.006	0.008]
T	[0.012	0.01	0.976	0.56	0.976	0.01	0.976	0.976	0.976	0.542]

*
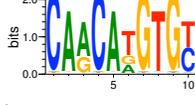
A	[0.015	0.967	0.665	0.015	0.967	0.206	0.015	0.015	0.015	0.02]
C	[0.961	0.01	0.011	0.961	0.01	0.014	0.01	0.01	0.01	0.472]
G	[0.01	0.01	0.308	0.01	0.01	0.214	0.961	0.01	0.961	0.013]
T	[0.015	0.015	0.016	0.015	0.015	0.566	0.015	0.967	0.015	0.495]

*

A	[0.096	0.015	0.015	0.015	0.015	0.014	0.016	0.967	0.018	0.666]
C	[0.879	0.01	0.01	0.01	0.01	0.853	0.644	0.01	0.143	0.307]
G	[0.01	0.01	0.01	0.961	0.961	0.009	0.011	0.01	0.819	0.011]
T	[0.014	0.967	0.967	0.015	0.015	0.12	0.329	0.015	0.018	0.016]

*

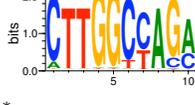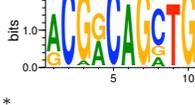|										
A	[0.715	0.011	0.07	0.463	0.011	0.976	0.01	0.166	0.01	0.01]
C	[0.008	0.971	0.007	0.01	0.971	0.007	0.007	0.638	0.007	0.007]
G	[0.266	0.007	0.914	0.512	0.007	0.007	0.971	0.178	0.007	0.971]
T	[0.011	0.01	0.01	0.015	0.01	0.01	0.011	0.016	0.976	0.011]

*
Position frequency matrices associated to the first nine top-ranked motifs obtained with the PNC training set.

Table S8. Genome-wide prediction of novel SOP specific CRMs.

Id.	Score	chrom.	start	stop	Closest gene	1	2	3	4	5	E-SOP (notum)	ch-SOP (leg discs)	NB (larval brain)
CRMTS (chnCRM)	27.52	2R	11019873	11020872	chn	1	5	0	0	2	-	-	-
CRM1	14.78	3R	25090483	25091482	stg	1	2	0	0	1	- (a)	-	-
CRM2	13.40	3L	900791	901790	CG34057	1	1	1	0	0	-	-	+
CRM3	13.39	2L	15425361	15426360	wor	0	3	0	1	0	-	-	+
CRM4	13.29	3R	7297094	7298093	CG17230	0	4	0	0	0	+	+	-
CRM5	12.11	3L	715989	716988	CG13896	1	1	0	1	0	-	-	+
CRM6	12.11	2L	12682650	12683649	pdm2	1	1	0	1	0	-	-	+
CRMTS (neurCRM1)	12.01	3R	4850333	4851313	neur	1 (g)	2	0	0	0	-	-	-
CRM7	12.01	2R	13363226	13364225	CG6520	1	2	0	0	0	+	+	+
CRM8	12.01	3L	20395366	20396365	trbl	1	2	0	0	0	+	+	+
CRM9	12.01	3L	1456583	1457582	rho	0	2	1	0	0	+ (b)	-	-
CRM10	11.36	2L	18191784	18192783	CG31746	0	2	1	0	0	-	-	-
CRM11	11.36	3L	1439214	1440213	SA-2	0	2	1	0	0	-	-	-
CRMchn' (c)	11.36	2R	11002969	11003968	chn	2	0	0	0	0	-	-	-
CRM12	10.73	3L	21203102	21204101	CG10510	2	0	0	0	0	-	-	-
CRM13	10.73	2R	14015868	14016867	Dip3	2	0	0	0	0	-	-	-
CRM14	10.73	3L	14568001	14569000	CG13479	2	0	0	0	0	-	-	+
CRM15	10.73	2R	15150809	15151808	CG7229	2	0	0	0	0	-	-	+
CRM16	10.73	3L	3785200	3786199	CG32264	2	0	0	0	0	-	-	-
CRM17	10.17	2R	20352047	20353046	CG3492	0	1	0	2	0	+	-	-
CRM18	10.17	2R	18821282	18822281	CG3788	0	1	0	2	0	-	-	+
CRMTS (phylCRM2)	10.08	2R	10323044	10324043	phyl	1	0	1	0	0	-	-	-
CRM19	10.08	X	354242	355241	ase	1	0	1	0	0	- (d)	-	+
CRM20	10.08	2R	6424677	6425676	lola	1	0	1	0	0	+	+	+
CRM21	10.08	2R	18008174	18009173	mei-S332	1	0	1	0	0	-	-	-
CRM22	10.08	3R	15161125	15162124	Dl	1	0	1	0	0	- (e)	-	-
CRM23	10.07	3L	6665397	6666396	Gr65a	0	2	0	1	0	+	+	+
CRM24	10.07	2L	4374703	4375702	Traf4	0	2	0	1	0	+ (b,f)	+	+
CRM25	10.07	3R	8345739	8346738	CG18554	0	2	0	1	0	-	-	-
CRM26	10.07	3R	5200244	5201243	Fps85D	0	2	0	1	0	+ (b)	-	+
CRM27	10.07	2L	17564670	17565669	CG7094	0	2	0	1	0	+	-	-
CRM28	10.07	2R	6824090	6825089	CG7722	0	3	0	0	0	+	-	-
CRM29	9.97	2R	20163828	20164827	nvy	0	3	0	1	0	+	+	+
CRMTS (cpoCRM6)	9.97	3R	13777573	13778572	cpo	1 (g)	0	0	1	0			
CRMTS (CG9363CRM1)	8.79	3R	5284700	5285699	CG9363	1	1	0	0	0			
CRM39	8.69	X	286744	287743	sc	1	1	0	0	0	+	-	-
CRM40	8.69	2R	6435347	6436346	lola	1	0	0	0	0	+	+	+
CRM41	8.69	2R	8655980	8656979	CG12374 (h)	1	1	0	0	0	+	+	+
CRMTS (spdoCRM3)	8.04	3R	26300357	26301356	spdo	0	1	1	0	0			
CRMTS (sensCRM3)	8.04	3L	13395416	13396396	sens	0	1	1 (g)	0	0			
CRMTS (spdoCRM4)	7.49	3R	26301660	26302659	spdo	0	0	0	0	1 (g)			
CRMTS (vvlCRM3)	6.75	3L	6778928	6779927	vvl	0	1	0	1	0			
CRM100	6.75	2R	20162435	20163434	nvy	0	1	0	1	0	+	+	+

The 29 best-scoring predicted CRMs are shown together with four additional fragments ranking between positions 39 and 100 that were also tested. The 9 top-ranking 1kb-long fragments overlapping the training set (marked TS) have also been displayed for comparison. For each fragment, the number of conserved sites above S_{th} = 13.3 is shown for each of the five top-ranked motifs, as well as its found pattern of expression in our transgenic fly reporter assay. E-SOP : external SOP of the pupal notum; ch-SOP : chordotonal SOP in leg imaginal discs; NB : neuroblasts in larval brain. (a) CRM1 is included within a larger genomic fragment with CRM activity in the embryonic PNS (10). (b) These CRMs are also active in PNCs. (c) This CRM overlaps the *chn* locus and was not tested. (d) CRM19 is located close to a SOP-specific gene. (e) CRM22 is included within a 2.1 kb fragment, defined here as CRM22'' (Fig. S5) with CRM activity in PNCs. (f) CRM24 overlaps with traf4CRM (Table S3). (g) neurCRM1 and cpoCRM6 contain two instances of motif 1, one conserved and one not conserved (see section 2.4 of the Supporting text for a definition of the conservation requirements); sensCRM3 contains three instances of motif3, one conserved and two not conserved; spdoCRM4 contains two conserved instances of motif 5 when realigned with muscle. Functional analysis of predicted motifs involved the mutations of both conserved and not conserved sites (see Figs. 2 and S3). (h) CRM41 is located 3' to CG12374 and 5' to the *scabrous* (*sca*) gene.

1. Powell L, Deaton A, Wear M, Jarman A (2008) Specificity of Atonal and Scute bHLH factors: analysis of cognate E box binding sites and the influence of Senseless. *Genes to Cells* 13:915.
2. Jafar-Nejad H, et al. (2003) Senseless acts as a binary switch during sensory organ precursor selection. *Genes & development* 17:2966.
3. Bang A, Posakony J (1992) The Drosophila gene Hairless encodes a novel basic protein that controls alternative cell fates in adult sensory organ development. *Genes & development* 6:1752.
4. Reeves N, Posakony J (2005) Genetic programs activated by proneural proteins in the developing Drosophila PNS. *Developmental cell* 8:413–425.
5. Gomes JE, Corado M, Schweisguth F (2009) Van Gogh and Frizzled act redundantly in the Drosophila sensory organ precursor cell to orient its asymmetric division. *PLoS ONE* 4:e4485.
6. Pi H, Huang S, Tang C, Sun Y, Chien C (2004) phyllopod is a target gene of proneural proteins in Drosophila external sensory organ development. *Proceedings of the National Academy of Sciences* 101:8378.
7. Castro B, Barolo S, Bailey A, Posakony J (2005) Lateral inhibition in proneural clusters: cis-regulatory logic and default repression by Suppressor of Hairless. *Development* 132:3333.
8. Lecourtois M, Schweisguth F (1995) The neurogenic suppressor of hairless DNA-binding protein mediates the transcriptional activation of the enhancer of split complex genes triggered by Notch signaling. *Genes & development* 9:2598.
9. Singson A, Leviten M, Bang A, Hua X, Posakony J (1994) Direct downstream targets of proneural activators in the imaginal disc include genes involved in lateral inhibitory signaling. *Genes & development* 8:2058.
10. Lehman D, et al. (1999) Cis-regulatory elements of the mitotic regulator, string/Cdc25. *Development* 126:1793.

<center>

Supporting text
Genome-wide identification of cis-regulatory motifs and modules underlying gene co-regulation, using statistics and phylogeny

</center>

1 Modelization of transcription factor affinity for DNA

1.1 Transcription factor frequency matrix

The DNA-binding specificity of a transcription factor (TF) T is represented by a frequency matrix $\mathbf{w}$ [1]. The matrix $\mathbf{w}$ specifies the frequency $w_{b,i}$ at which a base b ($b =$A, T, C or G) is found at position i, $1 \leq i \leq W$, in a set of properly aligned DNA binding sites $\mathbf{s} = (s_1, s_2, \cdots, s_W)$ for the factor T.

This representation [1] implicitly assumes that the affinity of a base for a transcription factor (TF) is independent of the other bases present in the binding site [2]. Although this may not be strictly true [3, 4], the number of sites found on the training set (see main text) corresponding to the best ranked matrices does not exceed a few dozens and does not allow the inference of further correlations.

1.2 Sites associated to a frequency matrix

Each frequency matrix corresponds to a position weight matrix (PWM; see [1]) :

$$\epsilon_{b,i} = \log_2 \frac{w_{b,i}}{\pi_b} \tag{1}$$

where π_b is the mean frequency of the base b within intergenic regions, ($\pi_{A,T} = 0.30$ and $\pi_{C,G} = 0.20$ as measured on the "background sequences",

<center>18</center>

see subsection 3.1). PWMs serve to infer the relative affinity of TFs for DNA sequences [2]. DNA sequences are assumed to be binding sites for a TF if they have a sufficiently high affinity. To this end, a score threshold S_{th} is introduced and a sequence of width W is assumed to be a site corresponding to the considered PWM if :

$$\sum_{i=1}^{W} \epsilon_{b(i),i} > S_{th} \tag{2}$$

where $b(i)$ is the base present at position i on the site sequence. As detailed in the main text, we typically used S_{th} values between 12 and 14 .

Reverse complement   A sequence corresponding to a given PWM can *a priori* recognize sites located on both DNA strands. We shall assume in the following that the recognized sites are not biased toward a particular stand. Therefore, we shall assume that a sequence of the sequenced strand also corresponds to a site i of the considered PWM if :

$$\sum_{i=1}^{W} \epsilon_{\bar{b}(W-i+1),i} > S_{th} \tag{3}$$

where $\bar{b}(i)$ is the complementary base of $b(i)$. Hence, the set of sites corresponding to a PWM is the set of sequences verifying either (2) or (3).

2   Algorithm for PWM inference

The goal of the algorithm described here is to infer PWMs and their corresponding binding sites, from a collection of intergenic sequences, the training set, with no *a priori* knowledge of the TFs involved. The training set consists of sequences for a given species (*D. melanogaster* in the present work). Conservation with other species (the 11 other sequenced *Drosophilae* species here) is used both to enrich the training set with orthologous sequences and to focus on PWMs that have conserved binding sites in different species.

2.1   Overview of the algorithm

The algorithm designed to build the matrices from the training set proceeds in several steps:

i) First, at each base position p in the training set, a sequence $\mathbf{s}$ of width W starting at p is extracted, and an initial approximative matrix is built using this unique sequence.

ii) The training set (consisting of *D. melanogaster* CRM sequences only) is exhaustively scanned for sites corresponding to the previously determined approximative matrix, *i.e.* for sites that have a score higher than S_{th}. For each found site, orthologous sites are searched in the 11 other sequenced *Drosophilae* species. These orthologous sites are combined to obtain a refined frequency matrix using phylogenetic information and a model of transcription factor binding site evolution. The procedure is iterated to converge on a final frequency matrix.

iii) The set of obtained PWMs is pruned by eliminating redundant PWMs and PWMs that correspond to repeated sequences by analyzing the statistics of their binding sites on a set of "background" intergenic sequences. The remaining set of PWMs is ranked according to the deviation of their bindings statistics on the validated enhancers of the training set, from what would be expected from their binding statistics on the background set.

The implementation of these steps as well as some technical assumptions are detailed below.

2.2 Bayesian inference of PWMs and choice of a *prior*

2.2.1 Bayesian inference

In the core part of the algorithm, the detection of sites corresponding to a PWM is used to refine this PWM. This is done in a "Bayesian" way [5]: the probability that a frequency matrix has a particular form is modified by the successive detection of binding sites. Namely, for a given frequency matrix $\mathbf{w}$, one can compute the probability $\mathcal{P}(\mathbf{s}|\mathbf{w})$ that one of its binding sites has the sequence $\mathbf{s} = (s_1, \cdots, s_W)$,

$$\mathcal{P}(\mathbf{s}|\mathbf{w}) = \prod_i w_{s_i,i} \tag{4}$$

If the probability of the different matrix forms is $\mathcal{P}(\mathbf{w})$, finding that $\mathbf{s}$ is a binding site of the searched matrix changes the probability of the different matrix forms to $\mathcal{P}(\mathbf{w}|\mathbf{s})$. The posterior probability $\mathcal{P}(\mathbf{w}|\mathbf{s})$ follows from

Bayes' rule for conditional probability,

$$\mathcal{P}(\mathbf{w}|\mathbf{s}) \propto \mathcal{P}(\mathbf{s}|\mathbf{w})\mathcal{P}(\mathbf{w}) \tag{5}$$

where the proportionality sign simply means that $\mathcal{P}(\mathbf{w}|\mathbf{s})$ should be normalized.

2.2.2 *Prior*

In order to start the process, one needs to decide what the probability is *a priori* that the frequency matrix has a particular form $\mathbf{w}$. For convenience, the *a priori* probability distribution, "the *prior*", that the matrix $\mathbf{w}$ has the column ($\{w_{b,i}\} \equiv \{w_{A,i}, w_{T,i}, w_{C,i}, w_{G,i}\}$) at position i, is chosen, as often, to be a Dirichlet distribution [6],

$$\mathcal{P}(\{w_{b,i}\}) = \frac{w_{A,i}^{\alpha-1} w_{T,i}^{\alpha-1} w_{C,i}^{\beta-1} w_{G,i}^{\beta-1}}{B(\alpha, \alpha, \beta, \beta)} \; \delta\left(1 - \sum_b w_{b,i}\right) \tag{6}$$

where the normalizing term $B(\alpha, \alpha, \beta, \beta)$ is the quadrinomial Beta function, the index b runs over the four bases types and the δ-function ensures that the sum of their probabilities is equal to 1 in each column of the matrix $\mathbf{w}$. The exponents associated with complementary bases are chosen equal in agreement with our assumption on reverse complement sites (see paragraph 1.2).

Two further assumptions fully determine the exponents α and β of the prior (Eq. (6) .

First, it is assumed that the *a priori* base frequencies at each position, in the set of frequency matrices, are equal to the base frequency in the background (*i.e.* that TF binding sites have no systematic bias in base composition),

$$\langle w_b \rangle_{\text{Prior distribution}} = \pi_b \tag{7}$$

This imposes that :
$$\frac{\alpha}{\beta} = \frac{\pi_{\text{A,T}}}{\pi_{\text{C,G}}} \tag{8}$$

A second condition on α and β arises from requiring that a frequency matrix contains on average a prescribed amount of information (*i.e.* deviate from the background frequencies). We require, consistent with our site detection method (see subsection 1.2), that the average information content of

21

a frequency matrix over the *prior*, is equal to the score threshold S_{th} defined above.

The information content IC of a matrix $\mathbf{w}$ is defined as :

$$\text{IC}(w) = \sum_{i,b} w_{b,i} \log_2(w_{b,i}/\pi_b) \tag{9}$$

where the sum runs over i, the index of the W possible column positions, and over b, which denotes the four possible base types. With the chosen *prior*, all columns contribute equally to the mean information content which can thus be written as :

$$\langle \text{IC} \rangle_{\text{Prior distribution}} = W \int \left(\prod_b dw_b \right) \sum_b w_b \log_2(w_b/\pi_b) \mathcal{P}(\{w_b\}) \tag{10}$$

The aforementioned condition translates into $\langle \text{IC} \rangle = S_{th}$. It leads upon performing the integrals and using Eq. (8) to :

$$2\pi_{A,T} \left[\psi(\alpha+1) - \psi\left(\frac{\alpha}{\pi_{A,T}} + 1 \right) - \ln(\pi_{A,T}) \right]$$
$$+ 2\pi_{C,G} \left[\psi\left(\frac{\pi_{C,G}\alpha}{\pi_{A,T}} + 1 \right) - \psi\left(\frac{\alpha}{\pi_{A,T}} + 1 \right) - \ln(\pi_{C,G}) \right] = S_{th} \log(2)/W \tag{11}$$

where ψ is the digamma function [7]. Eq. (11) determines the exponent α (and β) as a function of the information content *a priori* required for PWMs.

2.3   Initial matrices

The first step of the algorithm is, for each position p on the training set, to extract the sequence $\mathbf{s}$ of width W starting at p, and to build an approximative form for a matrix that would bind this particular sequence. Using Bayesian inference (Eq. (5)) and the Dirichlet *prior* (Eq. (6)), one obtains for the probability distribution of the matrices $\mathbf{w}$ that bind the sequence $\mathbf{s} = (s_1, ..., s_W)$

$$\mathcal{P}(\mathbf{w}|\mathbf{s}) \propto \prod_i w_{s_i,i} w_{A,i}^{\alpha-1} w_{T,i}^{\alpha-1} w_{C,i}^{\beta-1} w_{G,i}^{\beta-1} \tag{12}$$

22

The initial matrix $\mathbf{w}^{(in)}$ is chosen as the mean of the distribution (12) :

$$w_{b,i}^{(in)} = \frac{\delta_{s(i),b} + \alpha(\delta_{A,b} + \delta_{T,b}) + \beta(\delta_{C,b} + \delta_{G,b})}{1 + 2\alpha + 2\beta} \tag{13}$$

where $\delta_{A,b} = 1$ if $b = A$ and 0 when $b \neq A$. In other words, an initial matrix $\mathbf{w}^{(\mathbf{in})}$ is build for each sequence of the training set using pseudo-counts α for (A,T) and β for (C, G).

2.4 Matrix refinement

The second step of the algorithm consists in refining the initial matrix (13) using the training set sequences and conservation with orthologous species. This proceeds as follows.

2.4.1 Scan of the training set.

For a given initial matrix, the training set is exhaustively scanned to find all the N_1 corresponding sites $\mathbf{s}^{D.mel;j} = (s_1^{D.mel;j}, \cdots, s_W^{D.mel;j})$, $j = 1, \cdots, N_1$, *i.e.* sites that have a score higher than the threshold S_{th} for the initial matrix $\mathbf{w}^{(\mathbf{in})}$. Then, for each found binding site, orthologous sites are sought in the eleven other sequenced *Drosophilae* species. Only orthologous sites with a score above a milder threshold $S'_{th} < S_{th}$ are retained (the value $S'_{th} = S_{th} - 1.5$ was used). This allows some flexibility in the refinement process and it facilitates the retention of information coming from orthologous sequences. At the same time, it eliminates cases where no orthologous sequence is present for the considered CRM, either because the sequencing procedure left a hole, or because the regulatory sequence has disappeared through evolution, and cases where an orthologous sequence is present but in which the particular site under consideration has no orthologous counterpart.

Orthologous sites are sought on orthologous sequences of width W+40nt centered on the base aligned with the center of the site present in *D. melanogaster*. The possibility of shifts within the alignments is introduced to ensure robustness against errors coming from the alignements themselves, like spurious insertion-deletion introduction in the sites or shifts in the alignment. If several orthologous sites with a score higher than S'_{th} are found in one species (within the W+40nt window), only the site with the highest score is taken into account. If no orthologous site is found (because the orthologous site or CRM is absent), we simply ignore the species for that particular site.

2.4.2 Conservation requirements.

In order to reduce the noise coming from sequences that are poorly conserved and present in multiple copies by chance in the reference species (that may correspond to non-functional sites), we chose to keep only conserved sites, as defined below, in the refinement process. We defined a site as conserved if orthologous instances are found in at least three distant species, including *D. melanogaster*. We defined 5 groups of closely related species : {*melanogaster, simulans, sechellia, yakuba, erecta*}, {*ananassae*}, {*pseudoobscura, persimillis*}, {*willistoni*}, {*mojavensis, virilis, grimshawi*}. A site instance must be found in at least three of these five groups for the site to be considered as conserved. This conservation requirement reduces the N_1 sites in *D. melanogaster* to N conserved sites.

2.4.3 Matrix estimation using conserved binding sites.

The previous steps provide N conserved binding sites corresponding to a frequency matrix in the *D. melanogaster* training set aligned with their orthologous counterparts in the eleven other sequenced species, $s^{\sigma;j} = (s_1^{\sigma;j}, \cdots, s_W^{\sigma;j})$ where j is the index of the site ($j = 1, \cdots, N$), and ($\sigma = D.mel, \cdots, D.grim$) is the species index. The obtention of a refined frequency matrix requires computing the probability that each one of these N sites in the reference species and its orthologous sites in the other species are sites for a given frequency matrix $\mathbf{w}$. To this aim, we adopt here a simple evolutionary model for TF binding sites previously used in ref. [8, 9]. It assumes that the frequency matrices of orthologous transcription factors in different species and their common ancestor are identical. Then, when a point mutation occurs during the course of evolution in a TF binding site, it is assumed that the binding site is drawn at random among the possible binding sites (with all the others bases unchanged). In other words, the mutated base is chosen at random among the 4 different bases with probabilities equal to those of the column of the TF frequency matrix corresponding to the mutating base. This model translates into a simple mathematical form for the transition probabilities between a base b and a base b' at the i-th position in a binding site, for an ancestor and a daughter species at a phylogenetic distance of d,

$$p_{b \to b'} = q\delta_{b,b'} + (1 - q)\, w_{b',i} \tag{14}$$

where the proximity $q = \exp(-d)$ is the probability that no mutation has occurred between the two considered species.

Given a frequency matrix $\mathbf{w}$ and a species phylogenetic tree, this model gives the probability $\mathcal{P}(\{s_i^{\sigma;\alpha}\}|\mathbf{w})$ of observing the collection of bases $\{s_i^{\sigma;\alpha}\}$ at position i of the α-th binding site in all species in which the site is detected. This is done recursively [10] by computing backward in time, the probability $\mathcal{P}^m(s_i^\alpha = b|\mathbf{w})$ of a phylogenetic tree leading to the observed bases, in which a mother species m has base b at the i-th position of the site α, knowing the corresponding tree probabilities, $\mathcal{P}^{d1}(s_i^\alpha = b|\mathbf{w})$ and $\mathcal{P}^{d2}(s_i^\alpha = b|\mathbf{w})$, for its two daughter species $d1, d2$

$$
\begin{aligned}
\mathcal{P}^m(s_i^\alpha = b|\mathbf{w}) &= \left[q_{m,d1}\mathcal{P}^{d1}(s_i^\alpha = b|\mathbf{w}) + (1 - q_{m,d1}) \sum_{b'} w_{b',i}\mathcal{P}^{d1}(s_i^\alpha = b'|\mathbf{w}) \right] \\
&\times \left[q_{m,d2}\mathcal{P}^{d2}(s_i^\alpha = b|\mathbf{w}) + (1 - q_{m,d2}) \sum_{b'} w_{b',i}\mathcal{P}^{d2}(s_i^\alpha = b'|\mathbf{w}) \right]
\end{aligned}
$$

where $q_{m,d1}$ and $q_{m,d2}$ are the proximities between the mother and two daughters species. After climbing the whole species phylogenetic tree, this provides the probability of the tree starting from different bases at the i-th position of the site α in the species common ancestor $\mathcal{P}^{ca}(s_i^\alpha = b|\mathbf{w})$. Finally the probability $\mathcal{P}(\{s_i^{\sigma;\alpha}\}|\mathbf{w})$ of the observed collection of bases at the i-th position of the α-th site given the weight matrix $\mathbf{w}$, is obtained as,

$$
\mathcal{P}(\{s_i^{\sigma;\alpha}\}|\mathbf{w}) = \sum_b w_b \mathcal{P}^{ca}(s_i^\alpha = b|\mathbf{w}) \tag{15}
$$

The likelihood of a frequency matrix $\mathbf{w}$ for the whole collection of binding sites is computed from the individual probabilities $\mathcal{P}(\{s_i^{\sigma;\alpha}\}|\mathbf{w})$ by assuming that the evolution of the different bases in a binding site occurred independently as well as the evolution of different binding sites,

$$
\mathcal{P}(\mathbf{w}|\{\mathbf{s}^{\sigma;\alpha}\}) = \prod_{1 \leq \alpha \leq N} \prod_{1 \leq i \leq W} \mathcal{P}(\{s_i^{\sigma;\alpha}\}|\mathbf{w}) \, \mathcal{P}(\mathbf{w}) \tag{16}
$$

where the product on the right-hand side runs over the W positions of the N aligned conserved binding sites.

To estimate the best matrix that accounts for the observed sites and alignments, we use maximum likelihood, that is we take the matrix $\mathbf{w}$ that maximises the left-hand side of Eq.(16). This keeps the complexity of the algorithm within a numerically accessible range. The previously determined

Dirichlet exponents of the *prior* are changed accordingly so that the maximum likelihood estimate matches the mean estimate in the case of independent sites (sites without alignments) :

$$\mathcal{P}(\mathbf{w}|\{\mathbf{s}^{\sigma;\alpha}\}) = \prod_{1\leq\alpha\leq N} \prod_{1\leq i\leq W} \mathcal{P}(\{s_i^{\sigma;\alpha}\}|\mathbf{w}) \prod_{1\leq j\leq W} w_{A,j}^{\alpha} w_{T,j}^{\alpha} w_{C,j}^{\beta} w_{G,j}^{\beta} \qquad (17)$$

The numerical maximization is performed by using the Nelder and Mead simplex algorithm implemented in the *GNU Scientific Library* [11].

2.4.4 Iterative refinement

Once the refined matrix is obtained from the maximum likelihood estimation, it is again iteratively used to scan for sites in the training set until this process converges to a frequency matrix $w_{b,i}$. This type of algorithm sometimes leads to trapping of the solution into unwanted local optima. To avoid that, each frequency matrix $w_{b,i}$ is transformed to another matrix $w'_{b,i}$:

$$w'_{b,i} = \frac{w_{b,i} + \alpha(\delta_{A,b} + \delta_{T,b}) + \beta(\delta_{C,b} + \delta_{G,b})}{1 + 2\alpha + 2\beta} \qquad (18)$$

The algorithm is run a second time starting from $\mathbf{w}'$ until convergence.

2.5 Pruning and ordering the set of obtained PWMs.

The previous algorithm produces a large number of PWMs. Some of them are shifted duplicate of each other, some others appear to correspond to repeated sequences. The set of obtained PWMs thus needs to be pruned and the significance of the remaining ones assessed. These steps are described below.

2.5.1 Proximity between matrices

We start by defining a notion of proximity between frequency matrices that we call "strict proximity". It assumes that the matrices are well aligned and well oriented. We relax this constraint later.

The "strict proximity" between the two matrices $\mathbf{w}^{(1)}$ and $\mathbf{w}^{(2)}$ is defined by comparing the set of binding sites common to the two matrices, to the sets of binding sites for each one of them,

$$\text{Prox}(\mathbf{w}^{(1)}, \mathbf{w}^{(2)}) = 2\frac{\mathcal{P}\left\{[S\left(\mathbf{s}, \mathbf{w}^{(1)}\right) > S_{th}] \text{ and } [S\left(\mathbf{s}, \mathbf{w}^{(2)}\right) > S_{th}]\right\}}{\mathcal{P}\{S(\mathbf{s}, \mathbf{w}^{(1)}) > S_{th}\} + \mathcal{P}\{S(\mathbf{s}, \mathbf{w}^{(2)}) > S_{th}\}} \qquad (19)$$

where $\mathcal{P}\{S(\mathbf{s}, \mathbf{w}) > S_{th}\}$ is the probability that a sequence $\mathbf{s}$ drawn at random with the background frequencies (π_b, $b = A, C, G, T$), has a score $S(\mathbf{s}, \mathbf{w})$ above the threshold S_{th} for the frequency matrix $\mathbf{w}$. Similarly, the numerator of the expression (19), $\mathcal{P}\{[S(\mathbf{s}, \mathbf{w}^{(1)}) > S_{th}]$ and $[S(\mathbf{s}, \mathbf{w}^{(2)}) > S_{th}]\}$, is the probability that a sequence is a binding site for both $\mathbf{w}^{(1)}$ and $\mathbf{w}^{(2)}$.

Given two matrices $\mathbf{w}^{(1)}$ and $\mathbf{w}^{(2)}$, $\text{Prox}(\mathbf{w}^{(1)}, \mathbf{w}^{(2)})$ could, in principle, be numerically computed by drawing a large ensemble of sequences. We find it more convenient and numerically much faster to use an analytic approximation $\text{Prox}_{as}(\mathbf{w}^{(1)}, \mathbf{w}^{(2)})$ that is asymptotically exact as the width W of the PWMs grows (in the limit where the mean information per matrix column is finite).

Before giving the expression of $\text{Prox}_{as}(\mathbf{w}^{(1)}, \mathbf{w}^{(2)})$, we first introduce some useful functions. For a matrix $\mathbf{w}$, we define the real functions $f(\mathbf{w})$ and $g(\mathbf{w})$ by

$$f(\mathbf{w}) = -\beta S_{th} + \sum_{j=1,\cdots,W} \ln\left[\sum_b \pi_b \exp(\beta \epsilon_{b,j})\right] \qquad (20)$$

$$g(\mathbf{w}) = \left[\beta^2 \sum_{j=1,\cdots,W} \frac{\sum_{b,c} \pi_b \pi_c (\epsilon_{b,j} - \epsilon_{c,j})^2 \exp[\beta(\epsilon_{b,j} + \epsilon_{c,j})]}{\left[\sum_b \pi_b \exp(\beta \epsilon_{b,j})\right]^2}\right]^{-1/2} \qquad (21)$$

in which the sum over b and c corresponds to sums over the four bases, $\epsilon_{b,j}$ is the PWM associated to $\mathbf{w}$ (Eq. (1)) and β is a function of $\mathbf{w}$ (or equivalently of $\epsilon_{b,j}$) implicitly defined by

$$S_{th} = \sum_{j=1,\cdots,W} \frac{\sum_b \pi_b \epsilon_{b,j} \exp(\beta \epsilon_{b,j})}{\sum_b \pi_b \exp(\beta \epsilon_{b,j})} \qquad (22)$$

Similarly, for two matrices $(\mathbf{w}^{(1)}, \mathbf{w}^{(2)})$, we define the real functions $h(\mathbf{w}^{(1)}, \mathbf{w}^{(2)})$ and $k(\mathbf{w}^{(1)}, \mathbf{w}^{(2)})$

$$h(\mathbf{w}^{(1)}, \mathbf{w}^{(2)}) \;=\; -(\gamma_1 + \gamma_2)S_{th} + \sum_j \ln\left[\sum_b \pi_b \exp(\gamma_1 \epsilon_{b,j}^{(1)} + \gamma_2 \epsilon_{b,j}^{(2)})\right] \tag{23}$$

$$k(\mathbf{w}^{(1)}, \mathbf{w}^{(2)}) \;=\; \left[\gamma_1^2 \gamma_2^2 \sum_{j,j',a,b,c,d} \frac{\pi_a \pi_b \pi_c \pi_d \left[(\epsilon_{a,j}^{(1)} - \epsilon_{b,j}^{(1)})(\epsilon_{c,j'}^{(2)} - \epsilon_{d,j'}^{(2)}) - (\epsilon_{c,j'}^{(1)} - \epsilon_{d,j'}^{(1)})(\epsilon_{a,j}^{(2)} - \epsilon_{b,j}^{(2)})\right]^2}{\left[\sum_b \pi_b \exp(\gamma_1 \epsilon_{b,j}^{(1)} + \gamma_2 \epsilon_{b,j}^{(2)})\right]^2 \left[\sum_b \pi_b \exp(\gamma_1 \epsilon_{b,j'}^{(1)} + \gamma_2 \epsilon_{b,j'}^{(2)})\right]^2}\right.$$
$$\left. \times \exp\left[\gamma_1(\epsilon_{a,j}^{(1)} + \epsilon_{b,j}^{(1)} + \epsilon_{c,j'}^{(1)} + \epsilon_{d,j'}^{(1)}) + \gamma_1(\epsilon_{a,j}^{(2)} + \epsilon_{b,j}^{(2)} + \epsilon_{c,j'}^{(2)} + \epsilon_{d,j'}^{(2)})\right]\right]^{-1/2} \tag{24}$$

where the indices a, b, c and d run over the four bases and γ_1 and γ_2 are implicitly defined as a function of $\mathbf{w}^{(1)}$ and $\mathbf{w}^{(2)}$ by the following equations,

$$S_{th} \;=\; \sum_{j=1,\cdots,W} \frac{\sum_b \pi_b \epsilon_{b,j}^{(1)} \exp(\gamma_1 \epsilon_{b,j}^{(1)} + \gamma_2 \epsilon_{b,j}^{(2)})}{\sum_b \pi_b \exp(\gamma_1 \epsilon_{b,j}^{(1)} + \gamma_2 \epsilon_{b,j}^{(2)})} \tag{25}$$

$$S_{th} \;=\; \sum_{j=1,\cdots,W} \frac{\sum_b \pi_b \epsilon_{b,j}^{(2)} \exp(\gamma_1 \epsilon_{b,j}^{(1)} + \gamma_2 \epsilon_{b,j}^{(2)})}{\sum_b \pi_b \exp(\gamma_1 \epsilon_{b,j}^{(1)} + \gamma_2 \epsilon_{b,j}^{(2)})} \tag{26}$$

Given two matrices $\mathbf{w}^{(1)}$ and $\mathbf{w}^{(2)}$, these functions allow us to compute the analytic approximation $\mathrm{Prox}_{as}(\mathbf{w}^{(1)}, \mathbf{w}^{(2)})$ of the strict proximity as,

$$\mathrm{Prox}_{as}(\mathbf{w}^{(1)}, \mathbf{w}^{(2)}) = \frac{2\sqrt{2}}{\pi} \frac{k(\mathbf{w}^{(1)}, \mathbf{w}^{(2)}) \exp\left[h(\mathbf{w}^{(1)}, \mathbf{w}^{(2)})\right]}{g(\mathbf{w}^{(1)}) \exp\left[f(\mathbf{w}^{(1)})\right] + g(\mathbf{w}^{(2)}) \exp\left[f(\mathbf{w}^{(2)})\right]} \tag{27}$$

A derivation of Eq. (27) is provided at the end of this subsection, for the convenience of the reader.

To take into account potential differences in the alignments of the frequency matrices, or in their orientation, $\mathrm{Prox}_{as}(\mathbf{w}^{(1)}, \mathbf{w}^{(2)})$ is computed for all the possible alignments of the two matrices (with a maximum shift of 3 nt) in the two possible orientations. When shifted matrices are compared, they are completed by additional columns with the background frequencies (i. e. with no specifity). The proximity between the two matrices is obtained simply by taking the maximum over the obtained strict proximities. Two PWMs are considered duplicates of each other (i. e. correspond to two overlapping set of sites) if, and only if, their proximity is higher than a chosen threshold. For the results presented here, this proximity threshold was chosen to be

1/10 and among duplicates the best-scoring matrix was kept. The β, γ_1 and γ_2 parameters have been computed by optimizing the equations Eq. (20,23). This has been implemented using the Brent algorithm for equation Eq. (20) and the Fletscher-Reeves conjugate gradient algorithm for equation Eq. (23) [11].

We conclude this subsection by a derivation of Eq. (27) using standard statistical mechanics techniques (similar calculations in a related context can be found, for instance, in OG Berg's appendix to [2] or in [12]).

The probability $\mathcal{P}\{S(\mathbf{w}, \mathbf{s}\} > S_{th})$ can be written

$$\mathcal{P}\{S(\mathbf{w}, \mathbf{s}) > S_{th}\} = \sum_{\mathbf{s}} p(\mathbf{s})\Theta(S(\mathbf{w}, \mathbf{s}) - S_{th}) \tag{28}$$

where $p(\mathbf{s})$ is the probability of drawing the sequence $\mathbf{s}$ and $\Theta(x)$ is the Heaviside function , $\Theta(x) = 1$ if $x > 0$ and $\Theta(x) = 0$ otherwise. The sum of sequences can be explicitly performed in a usual way by introducing an integral representation for the Heaviside function

$$\Theta(x - S_{th}) = \int_{S_{th}}^{+\infty} du \int_{-\infty}^{+\infty} \frac{d\lambda}{2\pi} \exp[i\lambda(x - u)] \tag{29}$$

Substitution in Eq. (28) and averaging over sequences leads to

$$\mathcal{P}\{S(\mathbf{w}, \mathbf{s}) > S_{th}\} = \int_{S_{th}}^{+\infty} du \int_{-\infty}^{+\infty} \frac{d\lambda}{2\pi} \exp\left\{-i\lambda u + \sum_j \ln\left[\sum_b \pi_b \exp(i\lambda\epsilon_{b,j})\right]\right\} \tag{30}$$

The integral on λ, the r. h. s. of Eq. (30), can be estimated by the method of steepest-descent in the limit where W, the width of the PWM, is large. We denote by $F(u, \lambda)$ the argument of the exponential in Eq. (30)

$$F(u, \lambda) = -i\lambda u + \sum_j \ln\left[\sum_b \pi_b \exp(i\lambda\epsilon_{b,j})\right] \tag{31}$$

The saddle-point is given by $\partial_\lambda F(u, \lambda) = 0$. We ultimately find that the u-integral is dominated by values close to the threshold S_{th} and we are considering restrictive and attainable values of S_{th} (i.e. below the value obtained by taking the base with the maximum $\epsilon_{b,j}$ at each column j). In this case,

a solution of the saddle-point equation is obtained for a purely imaginary $\lambda = -i\beta$ with $\beta > 0$ implicitly defined as a function of u by [1] ,

$$u = \sum_{j=1,\cdots,W} \frac{\sum_b \pi_b \epsilon_{b,j} \exp(\beta \epsilon_{b,j})}{\sum_b \pi_b \exp(\beta \epsilon_{b,j})} \tag{32}$$

The integral around the saddle point is performed by expanding $F(u,\lambda)$ around $\lambda = -i\beta$ as

$$F(u,\lambda) = F(u,-i\beta) + \frac{1}{2}(\lambda + i\beta)^2 \, \partial_\lambda^2 F(u,\lambda = -i\beta) \tag{33}$$

with

$$
\begin{aligned}
\partial_\lambda^2 F(u,-i\beta) &= -\sum_j \left\{ \frac{\sum_b \pi_b \epsilon_{b,j}^2 \exp(\beta \epsilon_{b,j})}{\sum_b \pi_b \exp(\beta \epsilon_{b,j})} - \left[\frac{\sum_b \pi_b \epsilon_{b,j} \exp(\beta \epsilon_{b,j})}{\sum_b \pi_b \exp(\beta \epsilon_{b,j})} \right]^2 \right\} \\
&= -\frac{1}{2} \sum_{j=1,\cdots,W} \frac{\sum_{b,c} \pi_b \pi_c (\epsilon_{b,j} - \epsilon_{c,j})^2 \exp[\beta(\epsilon_{b,j} + \epsilon_{c,j})]}{\left[\sum_b \pi_b \exp(\beta \epsilon_{b,j}) \right]^2}
\end{aligned}
\tag{34}
$$

Performing the gaussian integral on λ readily gives

$$\mathcal{P}\{S(\mathbf{w},\mathbf{s}) > S_{th}\} = \int_{S_{th}}^{+\infty} \frac{du}{\sqrt{2\pi}} \frac{1}{\sqrt{|\partial_\lambda^2 F(u,-i\beta)|}} \exp[F(u,-i\beta)] \tag{35}$$

The remaining integral over u can also be performed by the method of steepest descent. It is intuitively clear that it is dominated by the neighbourhood of S_{th}, its lowest bound. It can also be directly checked that $F(u,-i\beta)$ is a decreasing function of u by computing its derivative,

$$\frac{d}{du} F(u,-i\beta) = \frac{\partial}{\partial u} F(u,-i\beta) = -\beta \tag{36}$$

Although β is a function of u, the total derivative over u in Eq. (36) reduces to a partial derivative, since β is an extremum of the partial derivative over λ (Eq. (26)). Finally, one obtains

$$\mathcal{P}\{S(\mathbf{w},\mathbf{s}) > S_{th}\} = \frac{1}{\sqrt{2\pi}} \frac{1}{\beta \sqrt{|\partial_\lambda^2 F(S_{th},-i\beta)|}} \exp[F(S_{th},-i\beta)] \tag{37}$$

---

[1] We assume, here, that this solution is the dominant saddle-point for the evaluation of the integral.

30

or with the notations of Eq. (20) and (21)

$$\mathcal{P}\{S(\mathbf{w}, \mathbf{s}) > S_{th}\} = \frac{1}{\sqrt{\pi}} g(\mathbf{w}) \exp[f(\mathbf{w})] \tag{38}$$

The two-matrix binding probability $\mathcal{P}\{S(\mathbf{s}, \mathbf{w}^{(1)}) > S_{th})$ and $S(\mathbf{s}, \mathbf{w}^{(2)}) > S_{th})\}$ can be computed in a fully analogous way. We denote it by the shorter notation $\mathcal{P}\{\mathbf{w}^{(1)}, \mathbf{w}^{(2)}\}$ and sketch here the main steps of its computation. First, it can be written

$$\mathcal{P}\{\mathbf{w}^{(1)}, \mathbf{w}^{(2)}\} = \sum_{\mathbf{s}} p(\mathbf{s}) \, \Theta(S(\mathbf{w}^{(1)}, \mathbf{s}) - S_{th}) \, \Theta(S(\mathbf{w}^{(2)}, \mathbf{s}) - S_{th}) \tag{39}$$

After the introduction of integral representations for the two Θ-functions (Eq. (29)), the average over sequences can be explicitly performed to obtain

$$\mathcal{P}\{\mathbf{w}^{(1)}, \mathbf{w}^{(2)}\} = \int_{S_{th}}^{+\infty} du_1 \int_{S_{th}}^{+\infty} du_2 \int_{-\infty}^{+\infty} \frac{d\lambda_1}{2\pi} \int_{-\infty}^{+\infty} \frac{d\lambda_2}{2\pi} \exp[H(u_1, u_2, \lambda_1, \lambda_2)] \tag{40}$$

with

$$H(u_1, u_2, \lambda_1, \lambda_2) = -i\lambda_1 u_1 - i\lambda_2 u_2 + \sum_j \ln\left[\sum_b \pi_b \exp(i\lambda_1 \epsilon_{b,j}^{(1)} + i\lambda_2 \epsilon_{b,j}^{(2)})\right] \tag{41}$$

The double integral on λ_1, λ_2 can, as before, be performed by steepest descent. The saddle point $(\lambda_1, \lambda_2) = (-i\gamma_1, -i\gamma_2)$ is determined by the following two equations

$$u_1 = \sum_{j=1,\cdots,W} \frac{\sum_b \pi_b \, \epsilon_{b,j}^{(1)} \exp(\gamma_1 \epsilon_{b,j}^{(1)} + \gamma_2 \epsilon_{b,j}^{(2)})}{\sum_b \pi_b \exp(\gamma_1 \epsilon_{b,j}^{(1)} + \gamma_2 \epsilon_{b,j}^{(2)})} \tag{42}$$

$$u_2 = \sum_{j=1,\cdots,W} \frac{\sum_b \pi_b \epsilon_{b,j}^{(2)} \exp(\gamma_1 \epsilon_{b,j}^{(1)} + \gamma_2 \epsilon_{b,j}^{(2)})}{\sum_b \pi_b \exp(\gamma_1 \epsilon_{b,j}^{(1)} + \gamma_2 \epsilon_{b,j}^{(2)})} \tag{43}$$

The argument of the exponential is expanded around the saddle-point as

$$
\begin{aligned}
H(u_1, u_2, \lambda_1, \lambda_2) &= H(u_1, u_2, -i\gamma_1, -i\gamma_2) + \frac{1}{2}(\lambda_1 + i\gamma_1)^2 \, \partial_{\lambda_1}^2 H \\
&+ (\lambda_1 + i\gamma_1)(\lambda_2 + i\gamma_2) \, \partial_{\lambda_1 \lambda_2} H + \frac{1}{2}(\lambda_2 + i\gamma_2)^2 \, \partial_{\lambda_2}^2 H + \cdots
\end{aligned}
$$

Gaussian integration over λ_1, λ_2 leads to

$$\mathcal{P}\{\mathbf{w}^{(1)}, \mathbf{w}^{(1)}\} = \int_{S_{th}}^{+\infty} du_1 \int_{S_{th}}^{+\infty} \frac{du_2}{2\pi} \frac{1}{\sqrt{dH_2}} \exp[H(u_1, u_2, -i\gamma_1, -i\gamma_2)] \quad (44)$$

where

$$dH_2 = [\partial_{\lambda_1}^2 H \partial_{\lambda_2}^2 H - (\partial_{\lambda_1\lambda_2} H)^2]_{(\lambda_1 = -i\gamma_1, \lambda_1 = -i\gamma_1)}$$

$$= \frac{1}{8} \sum_{j,j'} \frac{\sum_{a,b,c,d} \pi_a \pi_b \pi_c \pi_d \left[(\epsilon_{a,j}^{(1)} - \epsilon_{b,j}^{(1)})(\epsilon_{c,j'}^{(2)} - \epsilon_{d,j'}^{(2)}) - (\epsilon_{c,j'}^{(1)} - \epsilon_{d,j'}^{(1)})(\epsilon_{a,j}^{(2)} - \epsilon_{b,j}^{(2)}) \right]^2}{\left[\sum_b \pi_b \exp(\gamma_1 \epsilon_{b,j}^{(1)} + \gamma_2 \epsilon_{b,j}^{(2)}) \right]^2 \left[\sum_b \pi_b \exp(\gamma_1 \epsilon_{b,j'}^{(1)} + \gamma_2 \epsilon_{b,j'}^{(2)}) \right]^2}$$

$$\times \exp[\gamma_1(\epsilon_{a,j}^{(1)} + \epsilon_{b,j}^{(1)} + \epsilon_{c,j'}^{(1)} + \epsilon_{d,j'}^{(1)}) + \gamma_2(\epsilon_{a,j}^{(2)} + \epsilon_{b,j}^{(2)} + \epsilon_{c,j'}^{(2)} + \epsilon_{d,j'}^{(2)})] \quad (45)$$

Finally, the integration over u_1 and u_2 in Eq. (44) can also be performed using the method of steepest descent. As in the single matrix case, it is dominated by the neighbourhood of $u_1 = u_2 = S_{th}$. With $(\partial_{u_1} H = -\gamma_1, \partial_{u_2} H = -\gamma_2)$, for $u_1 = u_2 = S_{th}$, one obtains

$$\mathcal{P}\{\mathbf{w}^{(1)}, \mathbf{w}^{(2)}\} = \frac{1}{2\pi\gamma_1\gamma_2} \frac{1}{\sqrt{dH_2}} \exp[H(S_{th}, S_{th}, -i\gamma_1, -i\gamma_2)] \quad (46)$$

or with the notations of Eq. (23,24)

$$\mathcal{P}\{\mathbf{w}^{(1)}, \mathbf{w}^{(2)}\} = \frac{\sqrt{2}}{\pi} k(\mathbf{w}^{(1)}, \mathbf{w}^{(2)}) \exp\left[h(\mathbf{w}^{(1)}, \mathbf{w}^{(2)}) \right] \quad (47)$$

The sought expression of Eq. (27) for $\mathrm{Prox}_{as}(\mathbf{w}^{(1)}, \mathbf{w}^{(2)})$ directly follows from the obtained asymptotic expression of Eq. (47) for $\mathcal{P}\{\mathbf{w}^{(1)}, \mathbf{w}^{(2)}\}$ combined to the asymptotics of Eq. (38) for $\mathcal{P}\{S(\mathbf{s}, \mathbf{w}^{(1)}) > S_{th}\}$ and for $\mathcal{P}\{S(\mathbf{s}, \mathbf{w}^{(2)}) > S_{th}\}$.

2.5.2 Elimination of Motifs sampled from simple repeats

The training set, before being scanned, had been masked against simple repeats (annotation from Flybase obtained with Repeat Masker [13]) . However, we observed in our first attempts that still many of the obtained PWMs had binding sites that matched simple repeats. This introduced a large amount of noise in the CRM inference (subsection 2.6) and led us to develop a method to remove these PWMs using their binding site statistics on background sequences.

A characteristic of simple repeats is that they lead to non-Poisson distributions of binding sites : when a site is detected, there is a high probability that another site is detected after a multiple of the repeat period. Based on this feature, we have designed a quantitative way to remove the corresponding PWMs. First, the binding sites of each PWM (that is nucleotide sequences verifying Eq. (2) or (3)) are determined on a large background set of $N_{bg} = 10^4$ intergenic sequences, each of length $L_{ig} = 2000$ nt (subsection 3.1). Second, this data is used to compute for each frequency matrix w, the mean concentration $\lambda_w^{(bg)}$ of its binding sites on the background set. Last, for each PWM, the observed distribution of motifs on the background set is compared to what would be expected for a Poisson distribution with the same concentration of binding sites. For a frequency matrix **w** with a mean concentration $\lambda_w^{(bg)}$ of binding sites, one would expect from a Poisson distribution, $N_w^{(p)}(j)$ intergenic sequences in the background set containing j binding sites of w, with

$$N_w^{(p)}(j) = N_{bg}\frac{(\lambda_w^{(bg)} L_{ig})^j}{j!} \exp(-\lambda_w^{(bg)} L_{ig}) \tag{48}$$

For each frequency matrix **w**, the proximity of the distribution of the observed number $N_w(j)$ of background sequences with j binding sites to the ideal Poisson distribution (48) can be quantitatively assessed by computing the χ^2-like value,

$$\chi^2(w) = \sum_j \frac{[N_w(j) - N_w^{(p)}(j)]^2}{N_w^{(p)}(j)}\Theta(N_w(j)) \tag{49}$$

where again Θ is the Heaviside function. That is, in the computation of $\chi^2(w)$ the sum is restricted to non-zero values of $N_w(j)$. Retaining frequency matrices with a $\chi^2(w)$ below a threshold value of 10^3 produced satisfactory results (see table S4).

2.5.3 Matrix scoring

After the elimination of redundant PWMs and of the PWMs corresponding to simple repeats, the significance of the large number of remaining ones need to be assessed.

After the simple repeat elimination step, the remaining PWM have binding sites which are approximately Poisson-distributed in the set of background intergenic sequences (see subsection 3.1). It is thus possible to assess

the PWM significance, and rank them, by quantifying how much the distribution of their binding sites on the validated enhancers of the training set (v.e.t.s.) deviates from the expected Poisson distribution. This is done by computing, for each frequency matrix $\mathbf{w}$, the Poisson log-likelihood on the v.e.t.s :

$$Pl(w) = - \sum_{t \in \{\text{v.e.t.s.}\}} \log \left(\frac{(L\lambda_w^{(bg)})^{k_t} \exp(-L\lambda_w^{(bg)})}{k_t!} \right) \qquad (50)$$

where k_t is the number of instances of m on the sequence t of the v.e.t.s. . The computed $Pl(m)$ serves to rank the motifs.

2.6   CRM scoring at the genome scale

The set of obtained PWMs was used to detect SOP-specific CRMs on a genome wide scale.

First, for the 15 first ranked PWMs, conserved binding sites instances were sought and determined in the whole *D. melanogaster* genome as described previously for the training set. In order to do that, the Mavid Mercator alignment (see subsection 3.2) was used without further refinement, but after masking *D. melanogaster* genomic sequences for coding sequences.

Then, to predict CRMs, the masked *D. melanogaster* genomic sequence was chopped into 1kbp fragments (one every 50bp). Each fragment E was scored according to its content in binding sites with the score of a fragment defined by the log odds score :

$$S(E) = \sum_{\text{PWM } w} n_w(E) \ln \left[\frac{\lambda_w^{(tr)}}{\lambda_w^{(bg)}} \right] \qquad (51)$$

where $n_w(E)$ is the number of conserved binding sites of the frequency matrix $\mathbf{w}$ in the fragment E. Although, it would have been possible to use other algorithms for ranking putative enhancers given a set of PWM (e. g. [14, 15, 16]), the formula (51) was chosen both for its simplicity and for consistency between the conservation requirements imposed on the binding sites for PWM determination and fragment ranking.

2.7   Implementation of the algorithm

The developed programs have been written in C++ and are available upon request. They have been executed on an octoprocessor Intel Xeon machine

with 32 Go RAM.

3 Data

3.1 Intergenic regions

Intergenic sequences used to evaluate site statistics in non-specific regions are extracted from 10000 non-overlapping sequences of 2000bp drawn randomly from the *D. melanogaster* genome. Repeated sequences were not masked to better discriminate PWM arising from simple repeats.

3.2 Alignments

The alignments used in the analysis have been generated by Mercator (an orthology mapping program) and MAVID (a multiple alignment program) on the 12 drosophila genomes (CAF1). They have been downloaded from the AAAWiki web site (`http://rana.lbl.gov/drosophila/`). The orthologous sequences for the characterized CRMs have been extracted from this datasets and realigned using MUSCLE [17] for more refinement.

3.3 Assigning putative CRMs to GO terms

CRM ranking at the genome scale was described in section 2.6. In order to bio-informatically annotate these ranked putative CRMs (Fig. 3 of the main text), we associated to each one, the gene with the transcriptional start site closest to the center of the considered fragment. The fragment was then annotated as "SOP" when it was associated to a named gene with GO terms related to SOP developpement ("Sensory mother cell" and "Sensory organ"). These annotations by phenotype data have been obtained from Flybase [18]. Genes appearing as mere CG were not considered in the annotation part.

References

[1] Stormo G (2000) DNA binding sites: representation and discovery. *Bioinformatics* 16:16–23.

[2] Berg OG, von Hippel PH (1987) Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.* 193:723–750.

[3] Bulyk M, Johnson P, Church G (2002) Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic acids research* 30:1255.

[4] Tomovic A, Oakeley E (2007) Position dependencies in transcription factor binding sites. *Bioinformatics* 23:933.

[5] Cox D (2006) *Principles of statistical inference* (Cambridge Univ Pr).

[6] Durbin R, Eddy S, Krogh A, Mitchison G (1998) *Biological sequence analysis: Probabilistic models of proteins and nucleic acids* (Cambridge Univ Pr).

[7] Abramowitz M, Stegun I (1965) *Handbook of mathematical functions with formulas, graphs, and mathematical table* (Courier Dover Publications).

[8] Sinha S, van Nimwegen E, Siggia ED (2003) A probabilistic method to detect regulatory modules. *Bioinformatics* 19 Suppl 1:i292–301.

[9] Siddharthan R, Siggia E, van Nimwegen E (2005) PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput Biol* 1:e67.

[10] Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of molecular evolution* 17:368–376.

[11] Galassi M. *et al* (2009) *GNU Scientific Library Reference Manual*, Third edition.

[12] Djordjevic M, Sengupta AM, Shraiman BI (2003) A biophysical approach to transcription factor binding site discovery. *Genome Res.* 13:2381–2390.

[13] Bao Z, Eddy SR (2002) Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res.* 12:1269–1276.

[14] Berman BP, et al. (2002) Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. *Proc. Natl. Acad. Sci. U.S.A.* 99:757–762.

[15] Rebeiz M, Reeves NL, Posakony JW (2002) Score: a computational approach to the identification of cis-regulatory modules and target genes in whole-genome sequence data. site clustering over random expectation. *Proc Natl Acad Sci U S A* 99:9888–93.

[16] Sinha S, van Nimwegen E, Siggia ED (2003) A probabilistic method to detect regulatory modules. *Bioinformatics* 19 Suppl 1:292–301.

[17] Edgar R (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* 32:1792.

[18] Wilson RJ, Goodman JL, Strelets VB (2008) FlyBase: integration and improvements to query tools. *Nucleic Acids Res.* 36:D588–593.