Widespread evidence of cooperative DNA-binding by transcription factors in *Drosophila* development

Majid Kazemian[1,2], Hannah Pham[3], Scot A. Wolfe[3,4], Michael Brodsky[3,5,!], Saurabh Sinha[1,6,!]

[1]Department of Computer Science and [6]Institute of Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, IL, USA.

[2]Laboratory of Molecular Immunology, National Heart Lung and Blood Institute, National Institutes of Health, MD, USA.

[3]Program in Gene Function and Expression, [4]Department of Biochemistry and Molecular Pharmacology and [5]Department of Molecular Medicine, University of Massachusetts Medical School, MA, USA.

[!]Corresponding authors:

Saurabh Sinha sinhas@illinois.edu

Michael H. Brodsky Michael.brodsky@umassmed.edu

Running title: Identifying sequence signatures of TF interaction

Abbreviations:

TF	Transcription Factor
BR	Bound Region (of a transcription factor)
iTFs	“interacting TF signatures” (software tool)

Table S1. Correlation between computationally predicted and ChIP binding profiles. The first column shows the name of ChIP profile, with the three-letter abbreviation in parenthesis indicating the source of the ChIP profile (Mcc: modENCODE ChIP-chip, Mcs: modENCODE ChIP-seq, Bcs: BDTNP ChIP-seq, Bcc: BDTNP ChIP-chip, Fcs: Furlong ChIP-seq, Rcc: Rushlow ChIP-chip), and the number at the end indicating the replicate number, if applicable (1-7). The second column shows the name of the motifs used for computational prediction. All motifs are obtained from FlyFactorSurvey except the Giant (GT) motif, marked with asterisk, that is taken from (2). The third and fourth columns report the Pearson correlation co-efficient and its significance (p-value) between computationally predicted and ChIP binding profiles. The correlation is calculated over all non-overlapping segments of length 500bp that reside in highly accessible (90th percentile) regions during stage 5 of development. The fifth column indicates the number of overlaps between the top 2000 bound regions from ChIP profile and the top 2000 high scoring regions from STUBB profile in the accessible regions of the genome. The last column shows the fold enrichment of top ChIP bound regions in the top scoring regions from STUBB profile.

ChIP profile	Motif profile	Correlation	P-value	Overlap	FE
(Mcc)ubx	Ubx_FlyReg	0.41	0	608	1.70
(Bcc)mad.2	Mad_FlyReg	0.39	0	700	1.96
(Bcs)hb	hb_SOLEXA_5	0.3	0	653	1.83
(Bcs)bcd	Bcd_Cell	0.29	0	660	1.85
(Bcs)kr	Kr_NAR	0.26	0	628	1.76
(Mcc)eve	Eve_Cell	0.26	0	463	1.30
(Mcc)en	en_FlyReg	0.25	0	480	1.34
(Mcs)disco	disco-r-Cl1_SANGER_5	0.24	0	474	1.33
(Bcc)slp.1	slp1_NAR	0.21	2E-228	540	1.51
(Bcc)z.2	z_FlyReg	0.2	2.6E-225	515	1.44
(Bcc)D.1	D_NAR	0.2	3.6E-221	544	1.52
(Fcs)tin	Tin_SOLEXA	0.2	1E-210	501	1.39
(Bcc)cad	cad_FlyReg	0.19	1.2E-201	431	1.21
(Bcc)hkb.1	hkb_NAR	0.19	5.3E-196	619	1.73
(Bcs)gt	gt_bdtnp*	0.19	1.1E-189	551	1.54
(Bcc)sna.1	sna_SOLEXA_5	0.19	6.6E-189	570	1.60
(Mcc)bab	bab1_SANGER_5	0.19	2E-188	531	1.49
(Bcc)h.1	h_SANGER_5	0.17	9.5E-152	634	1.78
(Bcc)run.1	run_Bgb_NBT	0.16	1.6E-140	562	1.57
(Fcs)twi	twi_FlyReg	0.16	3.5E-132	665	1.85
(Bcc)tll.1	tll_NAR	0.15	2.4E-115	487	1.36
(Mcc)dll	Dll_Cell	0.14	8.9E-112	471	1.32
(Bcc)dl.3	dl_FlyReg	0.14	2.8E-102	437	1.22
(Bcc)ftz.3	ftz-f1_FlyReg	0.11	1.6E-67	458	1.28
(Mcc)GATAe	GATAe_SANGER_5	0.11	1.2E-63	421	1.18
(Bcc)prd.FQ	prd_NAR	0.09	1.5E-42	546	1.53
(Rcc)zld	vfl_SOLEXA_5	0.09	8.1E-41	637	1.78
(Bcc)da.2	da_SANGER_10	0.08	2.1E-35	577	1.62
(Mcc)inv	Inv_Cell	0.07	2.2E-30	350	0.98
(Bcc)shn.2	shn-F1-2_SANGER_5	0.07	4.2E-27	530	1.48
(Bcc)kni.2	kni_FlyReg	0.06	1.1E-19	466	1.31
(Mcc)ttk	ttk_NAR	-0.03	1	309	0.87
(Bcc)med.2	Med_FlyReg	-0.1	1	411	1.15

Table S2. [Separately uploaded Excel file] Spatial expression pattern of all 322 TFs. For each motif, shown is the list of term ids (separated by commas) and term names (separated by “;;”) that the corresponding TF gene is annotated with, as per BDGP.

Table S3. [Separately uploaded Excel file] All 1926 TF-pairs (5% FDR) with site arrangement. Please refer to legend of Table 2 in main text for detailed information about columns.

Table S4. [Separately uploaded file] List of TF-pairs in each of 711 clusters from Figure 5. Each row represents a cluster and lists all the TF-pairs in that cluster, separated by tabs. The two TFs of a TF-pair are separated by “:”. The clusters are sorted by their size with the largest cluster appearing in the first row.

Table S5. Assessing site arrangements among TFs with known physical interaction. Shown are 13 TF-pairs with known PPI (7 homotypic and 6 heterotypic) that have strong site arrangement. Please refer to the legend of Table 2 in the main text for detailed information about columns.

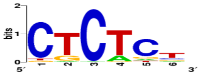
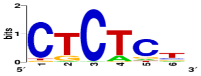
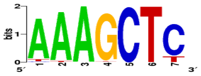
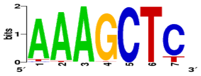
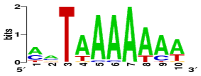
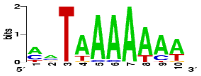
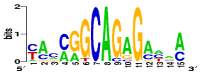
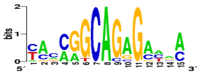
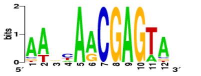
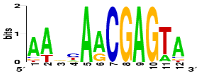
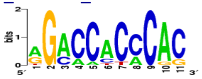
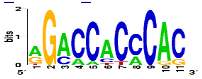
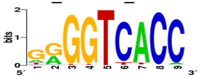
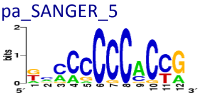
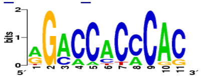
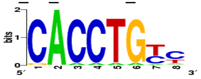
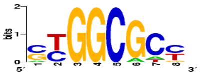
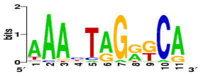
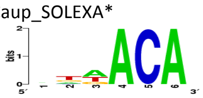
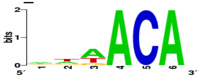
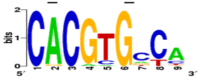
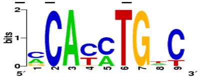
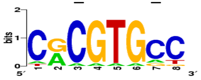
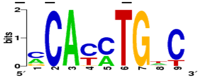
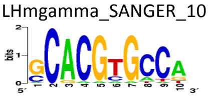
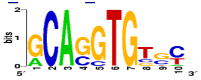
						0-10bp		**10-25bp**	
	Motif1	**Motif2**	**Family**	**Ori. type**	**Ori.**	**Dist.**	**OSD**	**Dist.**	**OSD**
Homotypic	Trl_FlyReg 	Trl_FlyReg 	(GAGA)	—> <— <— —> —> —> <— <—	- - 7e-15 7e-15	3e-22	- - 2e-27 2e-27	1e-12	- - 2e-11 2e-11
	lola_SOLEXA_5 	lola_SOLEXA_5 	(ZF-C2H2)	—> <— <— —> —> —> <— <—	- - - -	3e-10	0.004 - 1e-07 1e-07	3e-14	8e-04 3e-07 1e-06 1e-06
	jigr1_SANGER_5* 	jigr1_SANGER_5 	(MADF)	—> <— <— —> —> —> <— <—	- - - -	6e-07	7e-04 - 3e-04 3e-04	5e-06	- - 2e-05 2e-05
	Adf1_SANGER_5 	Adf1_SANGER_5 	(MADF)	—> <— <— —> —> —> <— <—	- - 1.4e-05 1.4e-05	1e-05	- - 0.003 0.003	-	- - - -
	CG12155_SANGER_5 	CG12155_SANGER_5 	(MADF)	—> <— <— —> —> —> <— <—	- - - -	4e-04	- - 5e-05 5e-05	-	- - 0.002 0.002
	Ci_SANGER_5 	Ci_SANGER_5 	(ZF-C2H2)	—> <— <— —> —> —> <— <—	- - - -	0.003	- - 2e-05 2e-05	-	- - - -
Heterotypic	CG8319_SOLEXA_2.5* 	Opa_SANGER_5 	(ZF-C2H2) (ZF-C2H2)	—> <— <— —> —> —> <— <—	- - - -	7e-06	- - 6e-04 0.004	-	- - - -
	Ci_SANGER_5 	nau_da_SANGER_5 	(ZF-C2H2) (bHLH)	—> <— <— —> —> —> <— <—	- - - -	1e-05	- - - 0.002	-	- - - -
	brk_FlyReg 	knrl_SANGER_5 	(BRK) (ZF-C4)	—> <— <— —> —> —> <— <—	- - - -	2e-05	0.001 - - -	-	- - - -
	Caup_SOLEXA* 	Mirr_SOLEXA 	(HOMEOBOX) (HOMEOBOX)	—> <— <— —> —> —> <— <—	- - - -	-	- - - -	6e-07	- - 3e-04 -
	HLHm3_SANGER_5 	ato_da_SANGER_10 	(bHLH) (bHLH)	—> <— <— —> —> —> <— <—	- - - -	2e-05	0.002 - - 0.002	-	- - - -
	HLHmbeta_SANGER_10* 	ato_da_SANGER_10 	(bHLH) (bHLH)	—> <— <— —> —> —> <— <—	- - - -	7e-05	- - - 0.003	-	- - - -
	HLHmgamma_SANGER_10* 	da_SANGER_10 	(bHLH) (bHLH)	—> <— <— —> —> —> <— <—	- - - -	-	- - - -	-	- - - 7e-05

Table S6. Selected instances of various site arrangements among TFs. The first 5 examples show orientation, distance, and OSD bias. The next 5 examples only have distance and OSD biases. Some of the cases show distance and/or OSD bias over multiple distance ranges (e.g. Sp1, Sp1). Please refer to legend of Table 2 in the main text for detailed information about columns.

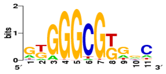
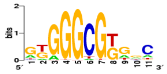
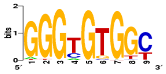
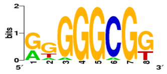
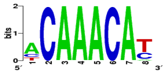
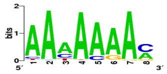
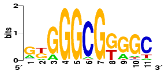
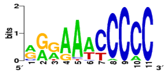
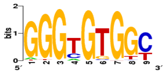
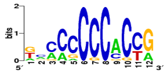
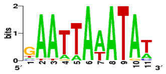
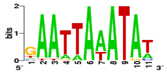
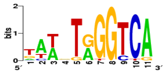
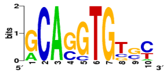
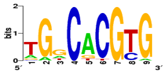
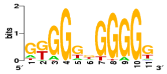
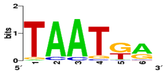
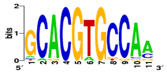
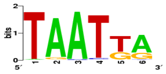
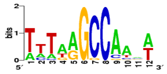
					0-10bp		10-25bp		25-50bp		50-100bp	
Motif1	**Motif2**	**Family**	**Ori. type**	**Ori.**	**Dist.**	**OSD**	**Dist.**	**OSD**	**Dist.**	**OSD**	**Dist.**	**OSD**
Sp1_SOLEXA_2.5 	Sp1_SOLEXA_2.5 	(zf-C2H2) (zf-C2H2)	—> <—	-	8e-29	-	5e-09	-	-	-	-	-
			<— —>	-		-		0.002		-		-
			—> —>	1.1e-09		7e-34		2e-07		-		-
			<— <—	1.1e-09		7e-34		2e-07		-		-
CG12029_SOLEXA_5 	btd_NAR 	(zf-C2H2) (zf-C2H2)	—> <—	-	2e-23	-	2e-04	-	-	-	-	-
			<— —>	-		-		-		-		-
			—> —>	2.6e-09		2e-18		5e-05		-		-
			<— <—	0.00033		1e-17		-		-		-
CG5953_SANGER_5 	jim_F1-9_SOLEXA_2.5 	(MADF) (zf-C2H2)	—> <—	-	1e-04	-	-	-	-	-	-	-
			<— —>	-		-		-		-		-
			—> —>	0.0041		1e-09		-		-		-
			<— <—	-		-		-		-		-
CG5669_SOLEXA_5 	Rel_SANGER_5 	(zf-C2H2) (RHD_TIG)	—> <—	1.2e-05	3e-05	2e-07	-	-	-	-	-	-
			<— —>	-		0.002		-		-		-
			—> —>	-		-		-		-		-
			<— <—	-		-		-		-		-
CG12029_SOLEXA_5 	Opa_SANGER_5 	(zf-C2H2) (zf-C2H2)	—> <—	-	8e-08	3e-07	-	-	-	-	-	-
			<— —>	5.9e-06		9e-05		-		-		-
			—> —>	-		-		-		-		-
			<— <—	-		-		-		-		-
bab1_SANGER_5 	bab1_SANGER_5 	(HTH_psq) (HTH_psq)	—> <—	-	3e-09	2e-06	4e-07	-	-	-	-	-
			<— —>	-		-		-		-		-
			—> —>	-		4e-04		4e-05		-		-
			<— <—	-		4e-04		4e-05		-		-
Eip75B_SANGER_5 	da_SANGER_10 	(zf-C4) (bHLH)	—> <—	-	-	-	-	-	-	-	-	-
			<— —>	-		-		3e-07		-		-
			—> —>	-		-		-		-		-
			<— <—	-		-		-		-		-
HLHmdelta_SANGER_10 	l3neo38_SOLEXA_2.5 	(bHLH) (zf-C2H2)	—> <—	-	-	-	-	-	0.002	1e-06	-	-
			<— —>	-		-		-		-		-
			—> —>	-		-		-		-		-
			<— <—	-		-		-		-		-
Dfd_SOLEXA 	HLHm7_SANGER_5 	(Homeobox) (bHLH)	—> <—	-	-	-	-	-	-	-	2e-05	0.002
			<— —>	-		-		-		-		0.004
			—> —>	-		-		-		-		-
			<— <—	-		-		-		-		-
Awh_SOLEXA 	Mes2_SANGER_5 	(Homeobox) (MADF)	—> <—	-	-	-	-	-	-	-	0.002	2e-06
			<— —>	-		-		-		-		-
			—> —>	-		-		-		-		-
			<— <—	-		-		-		-		-

Table S7. [Separately uploaded Excel file] Relationships based on site arrangement biases. For each TF, shown is the number of TFs with site arrangement bias (second column). The third and fourth columns report the TF family and additional binding domain(s) if any, respectively.

Table S8. [Separately uploaded Excel file] Measurement of direct, in vitro interaction between TF pairs. All the constructs are provided in the second worksheet.

Table S9. [Separately uploaded Excel file] DNA binding measurements for five homo or heterotypic TF-TF interactions. All the constructs, oligos, and measurements are included in the separate worksheets.

Table S10. [Separately uploaded Excel file] Detailed DNA binding measurements for five homo or heterotypic TF-TF interactions. All the constructs, oligos, and measurements are included in the separate worksheets.

Table S11. Fisher Exact test contingency table. We divided the adjacent binding site pairs from the "real" and "background" data sets into those within a range of *d* bp and those outside range *d* bp.

#Adjacent pairs	d ~ range	d <> range
Real	a	b
Background	c	d

Figure S1. Comparison between iTFs and SpaMo. For each method, the graph shows the number of predicted TF-pairs (out of 100) with distance bias, as a function of false positive rate. Blue and red lines correspond to homotypic and heterotypic TF-pairs, respectively. SpaMo was run with default parameter settings (see methods) here, unlike in Figure 2 where SpaMo was run with settings more comparable to those of iTFs.

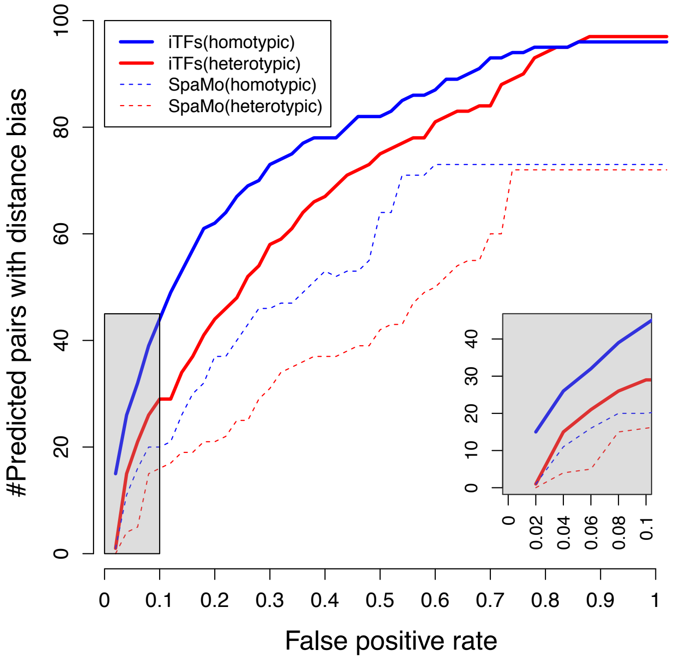


Figure S2. Site arrangement biases. (A) Each bar represents the average number of partners for each TF in that family before removal of “similar” TF-pairs. (Only TF families with more than 10 TFs included in our analysis are shown.) (B,C) The scatter plots show the number of partners with and without removal of “similar” TF-pairs for each TF (B) and each Family (C).

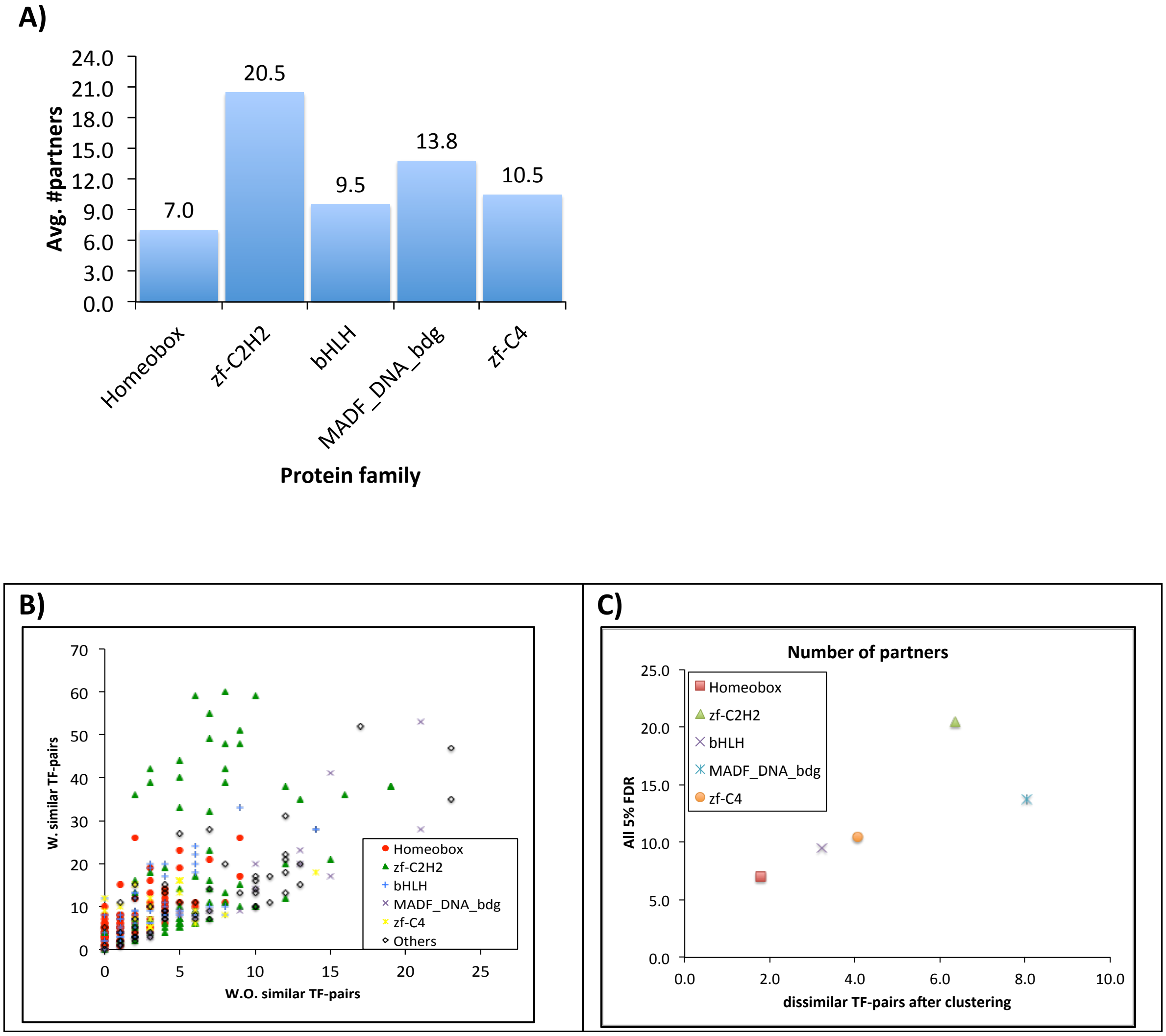
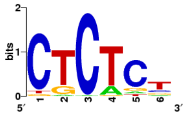
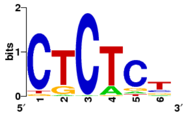
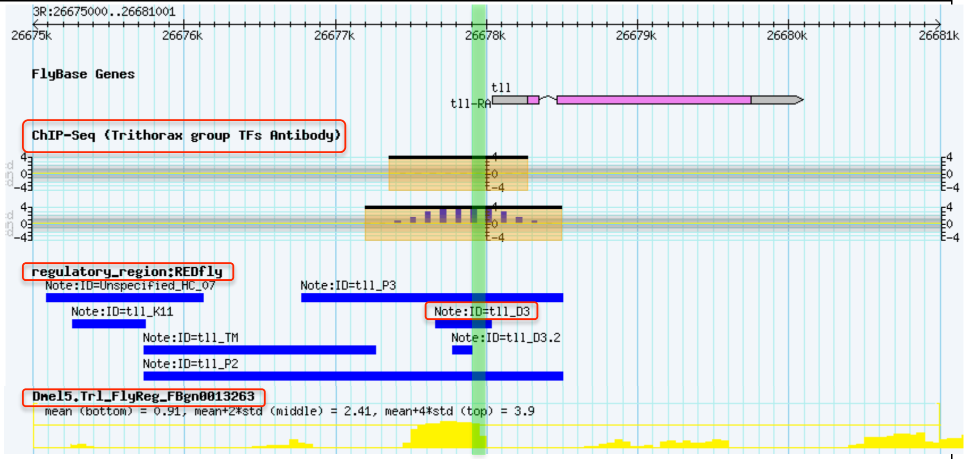
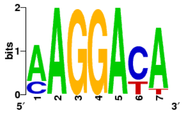
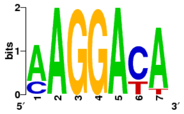
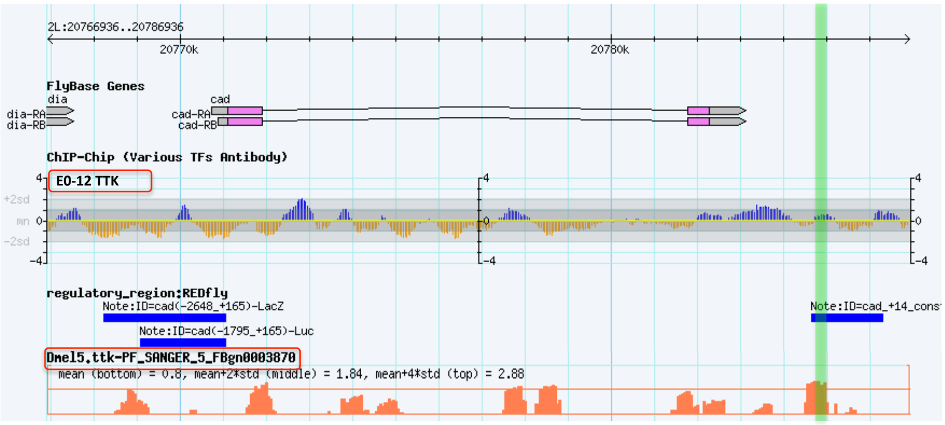
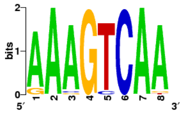
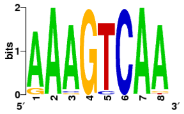
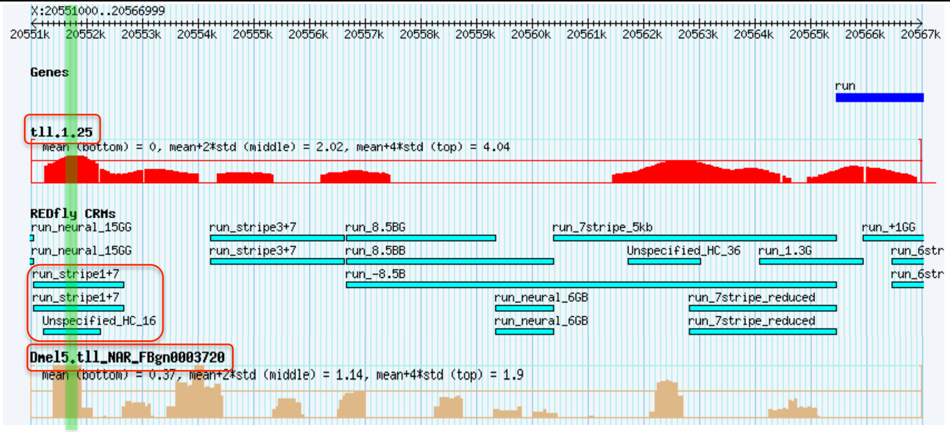


Figure S3. Genome browser view of selected co-bound regions (oligos) for five TF-pairs with cooperative binding to DNA. The motifs of selected TF-pairs are shown in the first column. The genome browser view of selected regions includes four tracks, Flybase genes, experimental ChIP profile(s), REDfly enhancers, and STUBB scores for the motifs. The filled green box represents the candidate oligo.

TF-Pair	Genome browser view
TRL  TRL 	
TTK  TTK 	
TLL  TLL 	

2L:20766936..20786936
20770k
20780k
FlyBase Genes
dia
dia-RA
dia-RB
cad
cad-RA
cad-RB
ChIP-Chip (Various TFs Antibody)
E0-12 TTK
4
+2sd
mn 0
-2sd
-4
regulatory_region:REDfly
Note:ID=cad(-2648_+165)-LacZ
Note:ID=cad(-1795_+165)-Luc
Note:ID=cad_+14_cons
Dmel5.ttk-PF_SANGER_5_FBgn0003870
mean (bottom) = 0.8, mean+2*std (middle) = 1.84, mean+4*std (top) = 2.88

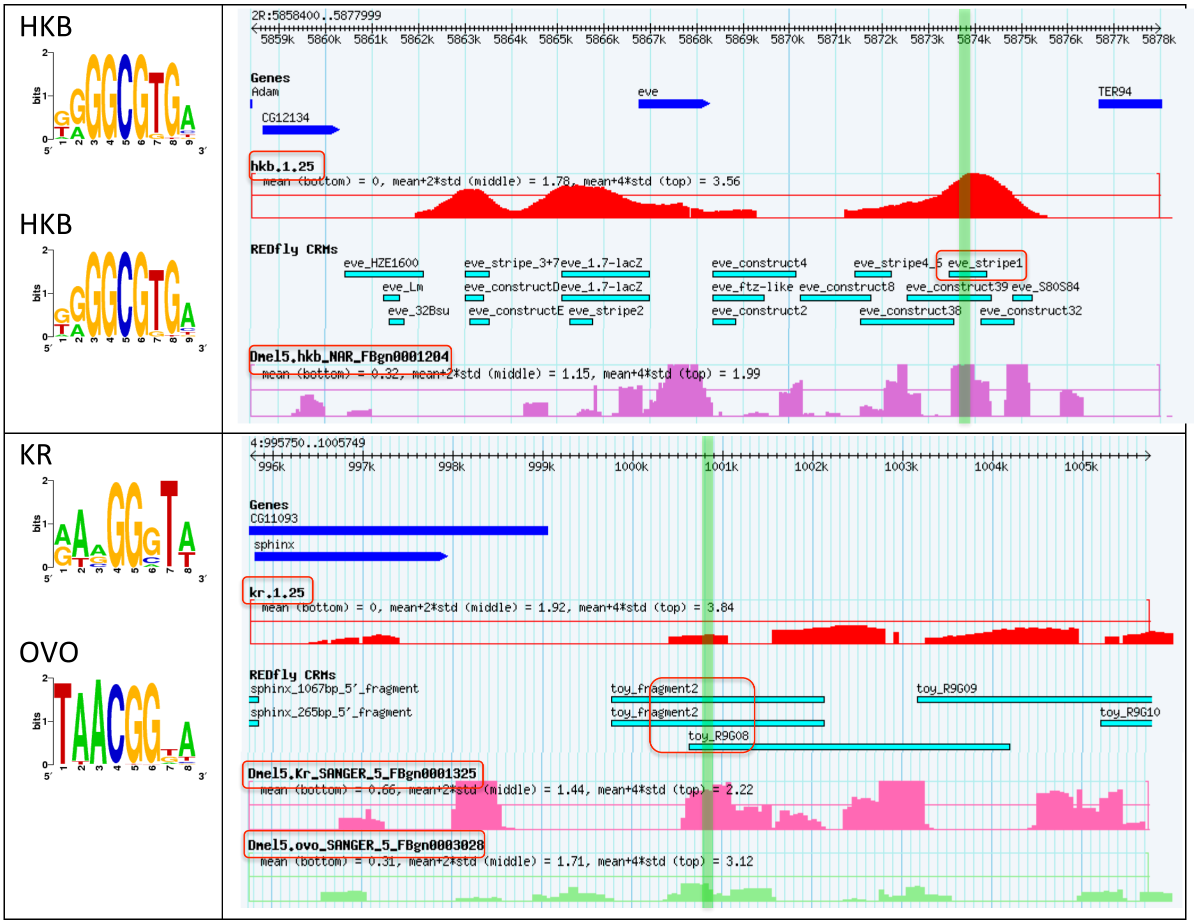

HKB
HKB
KR
OVO
bits
5′
3′
2R:5858400..5877999
5859k 5860k 5861k 5862k 5863k 5864k 5865k 5866k 5867k 5868k 5869k 5870k 5871k 5872k 5873k 5874k 5875k 5876k 5877k 5878k
Genes
Adam
CG12134
eve
TER94
hkb.1.25
mean (bottom) = 0, mean+2*std (middle) = 1.78, mean+4*std (top) = 3.56
REDfly CRMs
eve_HZE1600
eve_Lm
eve_32Bsu
eve_stripe_3+7
eve_constructD
eve_constructE
eve_1.7-lacZ
eve_1.7-lacZ
eve_stripe2
eve_construct4
eve_ftz-like
eve_construct2
eve_construct8
eve_stripe4_6
eve_construct39
eve_construct38
eve_stripe1
eve_S80S84
eve_construct32
Dmel5.hkb_NAR_FBgn0001204
mean (bottom) = 0.32, mean+2*std (middle) = 1.15, mean+4*std (top) = 1.99
4:995750..1005749
996k 997k 998k 999k 1000k 1001k 1002k 1003k 1004k 1005k
Genes
CG11093
sphinx
kr.1.25
mean (bottom) = 0, mean+2*std (middle) = 1.92, mean+4*std (top) = 3.84
REDfly CRMs
sphinx_1067bp_5′_fragment
sphinx_265bp_5′_fragment
toy_fragment2
toy_fragment2
toy_R9G08
toy_R9G09
toy_R9G10
Dmel5.Kr_SANGER_5_FBgn0001325
mean (bottom) = 0.66, mean+2*std (middle) = 1.44, mean+4*std (top) = 2.22
Dmel5.ovo_SANGER_5_FBgn0003028
mean (bottom) = 0.31, mean+2*std (middle) = 1.71, mean+4*std (top) = 3.12]

Figure S4. Distance preference profiles for five selected TF-pairs. The motifs of selected TF-pairs are shown in the first column. The second column plots the distance and OSD preferences (-log p-value) as a function of inter-site spacing between the two motifs. (Black dashed line corresponds to distance bias and colored lines correspond to OSD bias, for different orientations as shown in inset.)

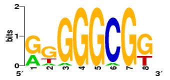
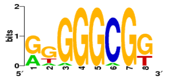
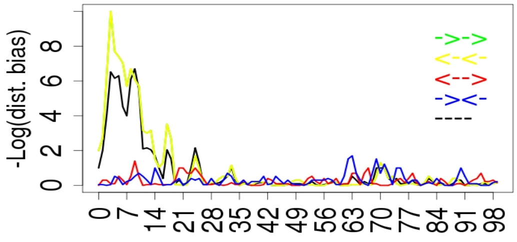
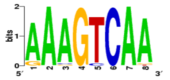
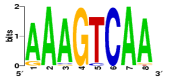
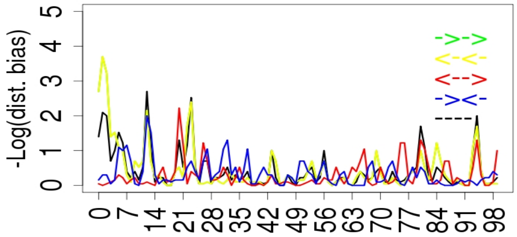
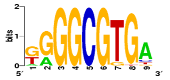
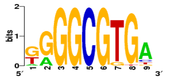
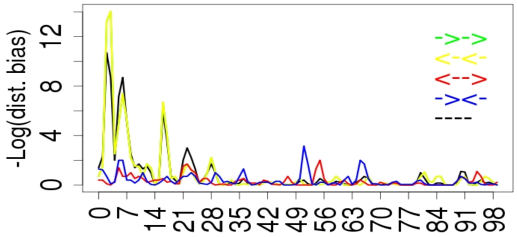
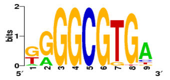
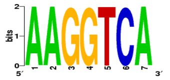
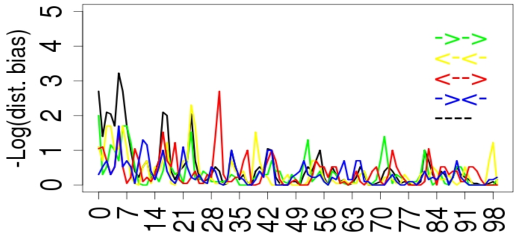
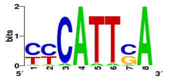
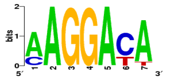
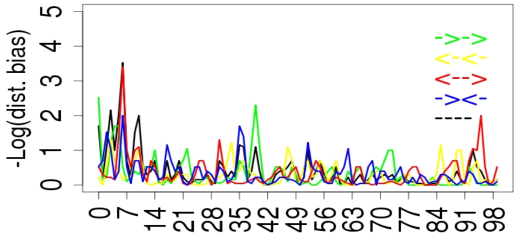
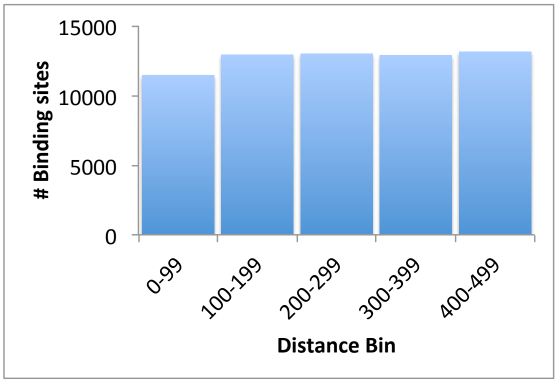
TF-pair	Distance preference
BTD  BTD 	
TLL  TLL 	
HKB  HKB 	
HKB  FTZ-F1 	
CIC  TTK-PF 	

Figure S5. Combined histogram of binding site locations in sequences analyzed by iTFs. For each of the 10 TFs with most number of partners as reported in manuscript, we examined the top 500 predicted bound regions of that TF, and annotated binding sites of the TF in each of these 500-bp regions. All of these binding sites were collected together and the distribution of their location within the 500-bp region is shown here.



Supplementary Note 1. Calculating the significance of abundance/depletion of site arrangement biases in a TF family. For each TF (out of N=322), we recorded the number of partners with site arrangement bias (Table S7). We counted the number of TFs that have >= 5 partners, and denoted this by *m*. We then counted the total number of TFs and the number of TFs with ≥ 5 partners in each TF family, denoted by *n* and *k* respectively. We finally calculated the significance of abundance of site arrangement biases in a TF family using a Hypergeometric test of significance, H(k, N, n, m). To obtain the significance of depletion of site arrangement in a TF family, we set *m* and *k* to be the number of TFs with no partner and the number of TFs of the family that had no partners.

REFERENCES

1. Li, X.Y., MacArthur, S., Bourgon, R., Nix, D., Pollard, D.A., Iyer, V.N., Hechmer, A., Simirenko, L., Stapleton, M., Luengo Hendriks, C.L. *et al.* (2008) Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm. *PLoS Biol*, **6**, e27.
2. Kaplan, T., Li, X.Y., Sabo, P.J., Thomas, S., Stamatoyannopoulos, J.A., Biggin, M.D. and Eisen, M.B. (2011) Quantitative models of the mechanisms that control genome-wide patterns of transcription factor binding during early Drosophila development. *PLoS Genet*, **7**, e1001290.
3. MacArthur, S., Li, X.Y., Li, J., Brown, J.B., Chu, H.C., Zeng, L., Grondona, B.P., Hechmer, A., Simirenko, L., Keranen, S.V. *et al.* (2009) Developmental roles of 21 Drosophila transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions. *Genome Biol*, **10**, R80.
4. Zinzen, R.P., Girardot, C., Gagneur, J., Braun, M. and Furlong, E.E. (2009) Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature*, **462**, 65-70.
5. Nien, C.Y., Liang, H.L., Butcher, S., Sun, Y., Fu, S., Gocha, T., Kirov, N., Manak, J.R. and Rushlow, C. (2011) Temporal coordination of gene networks by Zelda in the early Drosophila embryo. *PLoS Genet*, **7**, e1002339.
6. Celniker, S.E., Dillon, L.A., Gerstein, M.B., Gunsalus, K.C., Henikoff, S., Karpen, G.H., Kellis, M., Lai, E.C., Lieb, J.D., MacAlpine, D.M. *et al.* (2009) Unlocking the secrets of the genome. *Nature*, **459**, 927-930.
7. Schuettengruber, B., Ganapathi, M., Leblanc, B., Portoso, M., Jaschek, R., Tolhuis, B., van Lohuizen, M., Tanay, A. and Cavalli, G. (2009) Functional anatomy of polycomb and trithorax chromatin landscapes in Drosophila embryos. *PLoS Biol*, **7**, e13.