Motif-blind, genome-wide discovery of cis-regulatory modules in Drosophila and mouse

(Supplementary Information)

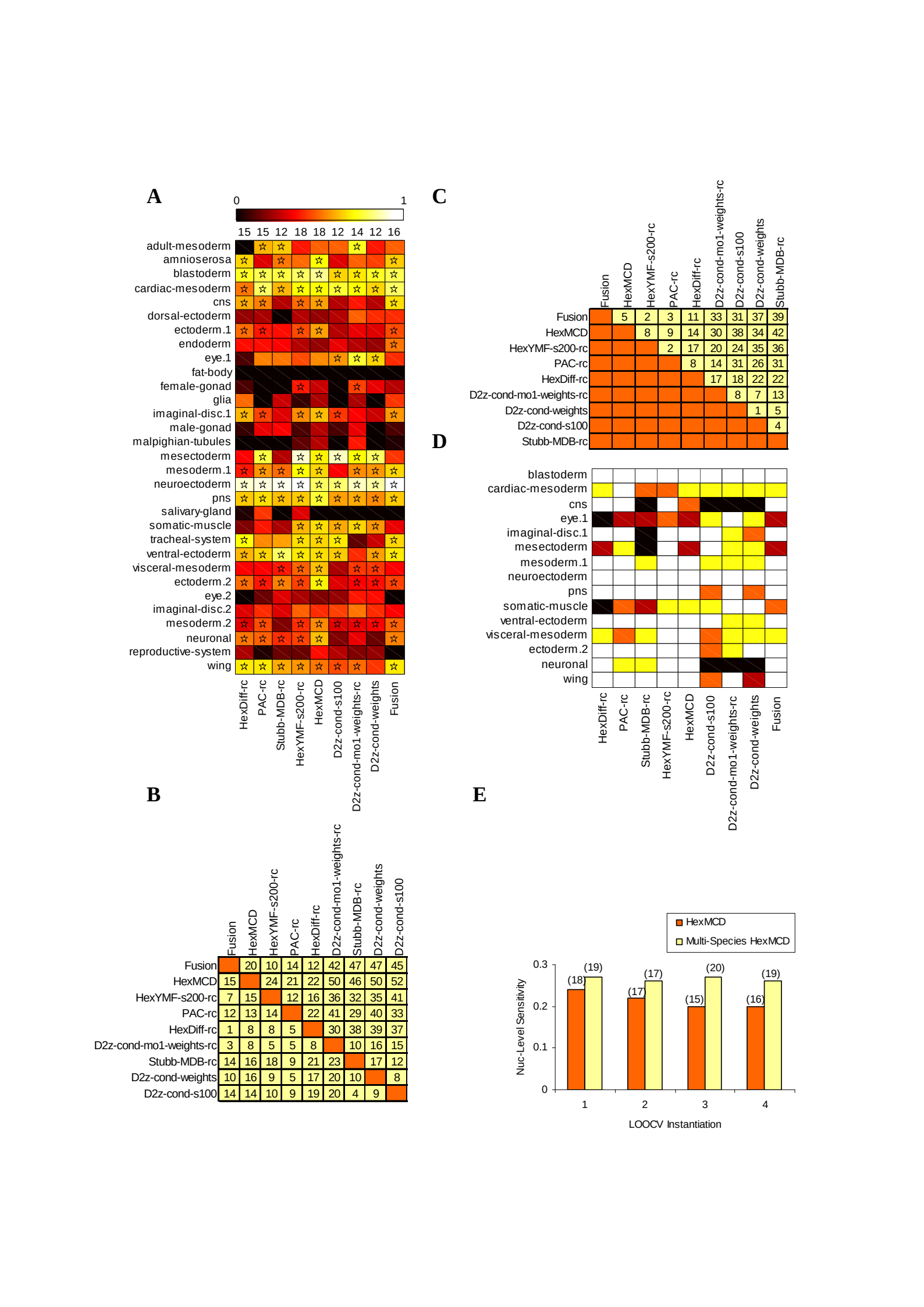
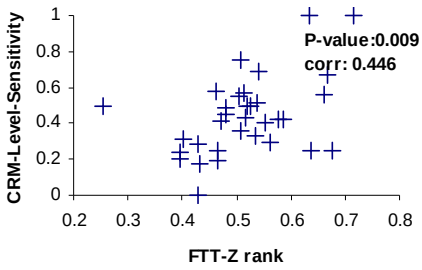
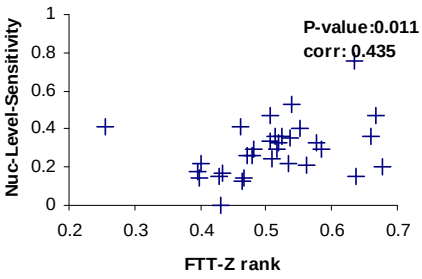
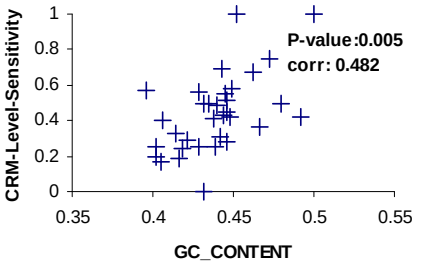

A
0
1
15 15 12 18 18 12 14 12 16
adult-mesoderm
amnioserosa
blastoderm
cardiac-mesoderm
cns
dorsal-ectoderm
ectoderm.1
endoderm
eye.1
fat-body
female-gonad
glia
imaginal-disc.1
male-gonad
malpighian-tubules
mesectoderm
mesoderm.1
neuroectoderm
pns
salivary-gland
somatic-muscle
tracheal-system
ventral-ectoderm
visceral-mesoderm
ectoderm.2
eye.2
imaginal-disc.2
mesoderm.2
neuronal
reproductive-system
wing
HexDiff-rc
PAC-rc
Stubb-MDB-rc
HexYMF-s200-rc
HexMCD
D2z-cond-s100
D2z-cond-mo1-weights-rc
D2z-cond-weights
Fusion
B
Fusion
HexMCD
HexYMF-s200-rc
PAC-rc
HexDiff-rc
D2z-cond-mo1-weights-rc
Stubb-MDB-rc
D2z-cond-weights
D2z-cond-s100
Fusion 20 10 14 12 42 47 47 45
HexMCD 15 24 21 22 50 46 50 52
HexYMF-s200-rc 7 15 12 16 36 32 35 41
PAC-rc 12 13 14 22 41 29 40 33
HexDiff-rc 1 8 8 5 30 38 39 37
D2z-cond-mo1-weights-rc 3 8 5 5 8 10 16 15
Stubb-MDB-rc 14 16 18 9 21 23 17 12
D2z-cond-weights 10 16 9 5 17 20 10 8
D2z-cond-s100 14 14 10 9 19 20 4 9
C
Fusion
HexMCD
HexYMF-s200-rc
PAC-rc
HexDiff-rc
D2z-cond-mo1-weights-rc
D2z-cond-s100
D2z-cond-weights
Stubb-MDB-rc
Fusion 5 2 3 11 33 31 37 39
HexMCD 8 9 14 30 38 34 42
HexYMF-s200-rc 2 17 20 24 35 36
PAC-rc 8 14 31 26 31
HexDiff-rc 17 18 22 22
D2z-cond-mo1-weights-rc 8 7 13
D2z-cond-weights 1 5
D2z-cond-s100 4
Stubb-MDB-rc
D
blastoderm
cardiac-mesoderm
cns
eye.1
imaginal-disc.1
mesectoderm
mesoderm.1
neuroectoderm
pns
somatic-muscle
ventral-ectoderm
visceral-mesoderm
ectoderm.2
neuronal
wing
HexDiff-rc
PAC-rc
Stubb-MDB-rc
HexYMF-s200-rc
HexMCD
D2z-cond-s100
D2z-cond-mo1-weights-rc
D2z-cond-weights
Fusion
E
HexMCD
Multi-Species HexMCD
Nuc-Level Sensitivity
0.3
0.2
0.1
0
(18) (19)
(17) (17)
(15) (20)
(16) (19)
1
2
3
4
LOOCV Instantiation

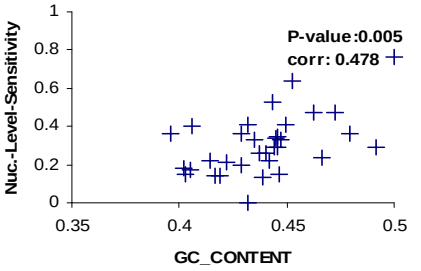
Figure S1. [previous page]. (A) LOOCV performance of each of nine different methods on 31 data sets in benchmark. Color accents represent Nucleotide-level sensitivity on a scale of 0 to 1, and cases with empirical p-value <= 0.05 are marked by asterisks. The top row shows the number of such data sets for each method. Numbers to the right are best Nucleotide-level sensitivity for each data set. One-on-one comparison of methods: For each pair of methods M1 (row) and M2 (column), (B) the "wins" of M1 versus M2 (i.e., the number of data sets on which nucleotide-level sensitivity of M1 was greater than that of M2 by at least 10% of data set size). (C) The difference between the wins of M1 versus M2 and the wins of M2 versus M1 in nucleotide-level sensitivity. (D) Sixteen data sets on which at least one method has nucleotide level sensitivity p-value of 0.05 or less in all four instantiations. Color indicates the number of instantiations (out of four) on which the performance was significant (p <= 0.05): white=4, yellow=3, orange=2, brown=1, black=0. (E) Comparison of single species and multi-species versions of HexMCD. For each each LOOCV instantiation, the Nucleotide-level sensitivity and number of amenable data sets at $p \leq 0.05$ (the value above each bar) are compared between the two methods.


CRM-Level-Sensitivity
P-value:0.009
corr: 0.446
FTT-Z rank

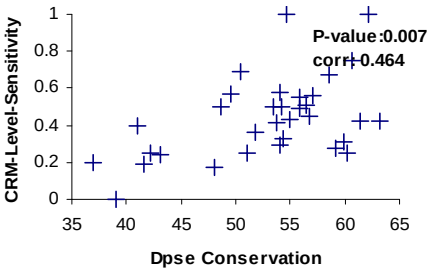

Nuc-Level-Sensitivity
P-value:0.011
corr: 0.435
FTT-Z rank

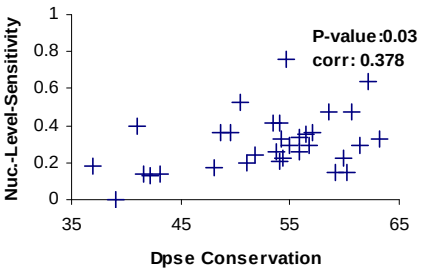
A


CRM-Level-Sensitivity
P-value:0.005
corr: 0.482
GC_CONTENT


Nuc.-Level-Sensitivity
P-value:0.005
corr: 0.478
GC_CONTENT

B


CRM-Level-Sensitivity
P-value:0.007
corr: 0.464
Dpse Conservation


Nuc.-Level-Sensitivity
P-value:0.03
corr: 0.378
Dpse Conservation

C

Figure S2. [previous page]. (A, B, C) The best CRM-level (left)/nucleotide-level(right) sensitivity on a data set is plotted against three high level characteristics of the data set: extent of homotypic clustering as measured by the "FTT-Z rank" (A), G/C content of training CRMs (B) and evolutionary conservation of training CRMs (C). The Pearson's correlation coefficient (corr) and its p-value are reported on each plot.

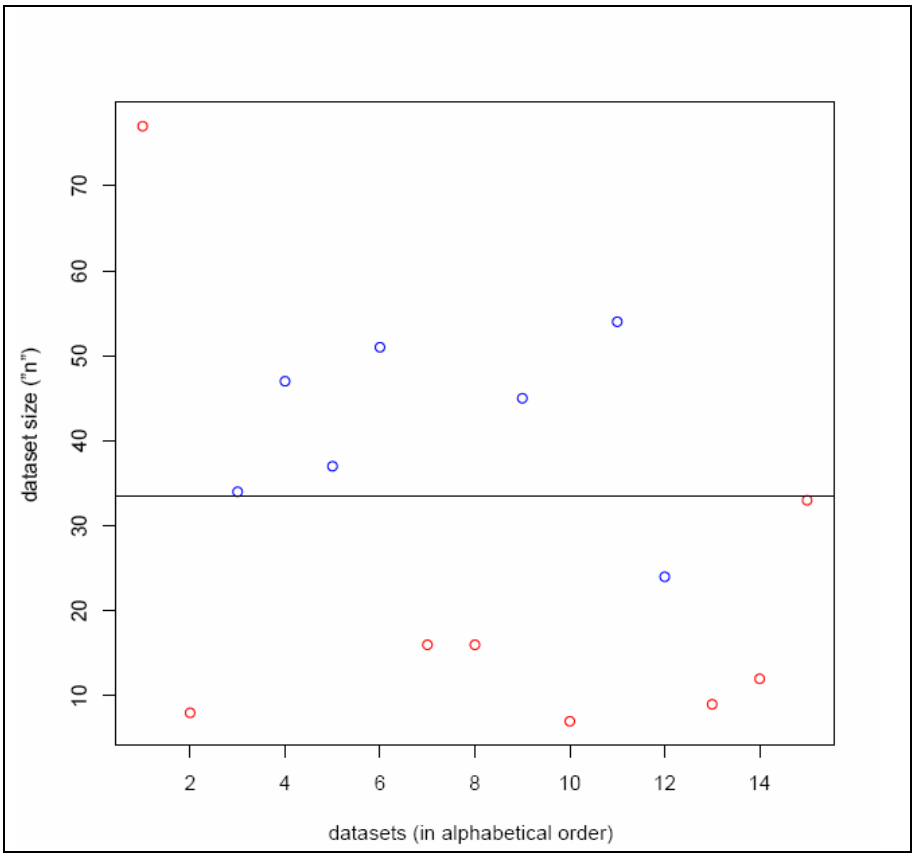

dataset size ("n")
10
20
30
40
50
60
70
2
4
6
8
10
12
14
datasets (in alphabetical order)

Figure S3. [previous page]. Dataset size and performance for amenable datasets in random training set evaluations. Red circles indicate datasets that performed no better with their appropriate training set than with a random training set; blue circles indicate datasets where random training sets led to worse performance. Near-perfect discrimination is achieved at n=33.5 (line).

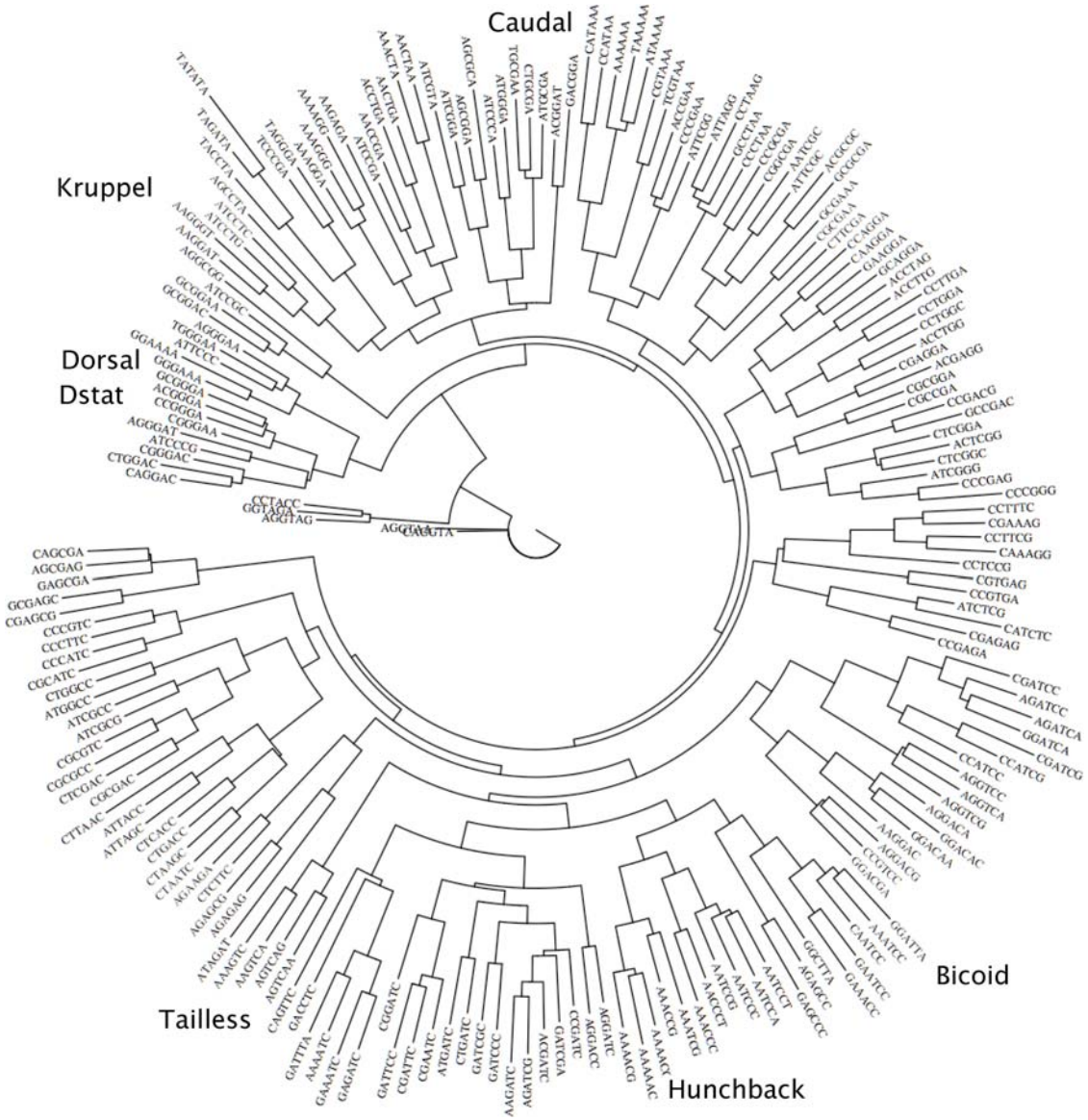

Caudal
Kruppel
Dorsal
Dstat
Tailless
Hunchback
Bicoid

Figure S4. [previous page]. Cladogram view of the 200 6-mers used by HexYMF for the blastoderm data set. We find k-mers matching the consensus string of six of the canonical motifs of the segmentation network (Bicoid, Caudal, Hunchback, Dstat, Kruppel, Tailless, matching the 6-mers GGATTA, CATAAA, AAAAAC, CCGGGA, AAGGGT, AGTCAA respectively) and Dorsal, also a blastoderm stage-active transcription factor, matching GGGAAA.

	#CRMs	HexDiff-rc	HexMCD	PAC-rc	HexYMF-s200-rc	D2z-cond-s100	D2z-cond-mo1-weights-rc	D2z-cond-weights
dorsal_root_ganglion	10	0.3	0.4	**0.7**	0.5	**0.7**	0.6	0.6
eye	16	0.19	0.25	0.25	0.38	0.31	0.56	0.25
forebrain	67	0.49	**0.8**	0.61	0.6	0.49	0.48	0.57
heart	19	0.26	0.32	0.32	0.26	0.37	0.37	0.26
hindbrain_rhombencephalon	32	0.56	**0.8**	**0.8**	0.56	**0.8**	0.69	0.66
limb	35	0.4	**0.6**	0.54	0.4	0.34	0.43	0.51
midbrain_mesencephalon	42	0.29	**0.6**	**0.6**	0.5	0.43	0.33	0.52
neural_tube	23	0.61	0.48	**0.8**	0.48	0.65	0.52	0.74

Figure S5. [previous page]. Cross-validation on mammalian enhancers. CRM-level sensitivity is shown for each method on each data set. Values with empirical p-value < 0.05 are shaded, and those at p < 0.01 are in darker shade. Best performance on each data set is shown in bold entries.

Table S1. **[next page].** Relationship between performance and methodological features. For each methodological feature, two methods were evaluated – one that incorporates the feature (top) and one that does not (bottom). (The two methods are identical otherwise). The average CRM-level sensitivity of each method and the net number of wins of the second method (i.e., one with the methodological feature) are shown. The "net number of wins" is the difference in numbers of wins of the two methods.

Method characteristic	*Method Name*	*CRM-Level Sennsitivity*	*# wins (Second method vs. First)*
Markov order effect (1-->2)	D2z-mo1	0.21	-1
	D2z-mo2	0.24	
effect of using markov background, in conditional setting	D2z-cond-s100-weights	0.24	10
	D2z-cond-mo1-s200-weights	0.25	
effect of markov background, in conditional setting and YMF weights	D2z-cond-weights	0.24	4
	D2z-cond-mo1-weights	0.25	
effect of using ymf weights in conditional setting with markov background	D2z-cond-mo1-s200-weights	0.25	4
	D2z-cond-mo1-weights	0.25	
effect of k=5 vs 6 in k-mers counting	D2z-cond-mo1-k5	0.23	6
	D2z-cond-mo1 (k=6)	0.24	
effect of reverse complement counting	D2z-cond-mo1-weights	0.25	8
	D2z-cond-mo1-weights-rc	0.26	
effect of conditional z-score	D2z-mo1	0.21	18
	D2z-cond-mo1	0.24	
effect of weights	D2-s100-rc	0.25	10
	D2-s100-weights-rc	0.24	
effect of using weights, in a z-score setting	D2z-cond-s100	0.24	-5
	D2z-cond-s100-weights	0.24	
effect of using subsets	D2z-cond	0.14	36
	D2z-cond-s100	0.24	
effect of subset size	D2-s100-weights-rc	0.24	-1
	D2-s200-weights-rc	0.25	
effect of not concatenating all training crms	D2z-mo1	0.21	20
	D2z-mo1-sepcrms	0.24	

Table S2[*]. In-depth analysis of CRM predictions for the blastoderm dataset. For each of the 113 predicted modules (Release 4.3 coordinates in Column A), the table shows: (i) nearest neighbor gene and location of module relative to gene (column B), (ii) whether or not it is a known module (column C), and if so, (iii) whether or not it is a "training CRM" (column D), (iv) whether or not the nearest gene has "pair-rule" expression pattern (column E), or "dorso-ventral (D/V)" expression pattern (column F), and (v) for the modules that do not overlap known modules (rows 62-115), the motifs associated with the module (columns G-AB). For modules that do overlap known modules, column AC names that known module, and for the subset of these that were not in the training set, column AD notes the tissue where the known module drives expression. The key word "long" in this column indicates that the reason for the module not being in the training set was its large length (modules longer than a threshold length were excluded from the training sets).

Table S3[*]. Cross-validation performance (complete data). The CRM and nucleotide level sensitivity and corresponding p-values of all methods in the four instantiations of cross-validation. Also included are the average sensitivity values over all instantiations.

Table S4[*]**.** Worksheets "HexYMF-s200-rc" and "PAC-rc" represent the genome-wide predictions in human for the early cardiovascular development data set, using these two methods respectively. Column A names the evolutionarily conserved region (ECR) in which the predicted module is located, with the start position of the window being given by column B. The score of the module is in column C, and information on adjacent gene(s) is

* These tables are included as separate Excel files: TableS2.xls, TableS3.xls, TableS4.xls, TableS5.xls

in column D. The fraction of the module's sequence that was identified as being short tandem repeats is shown in column E. Columns F and G indicate if the GATA motif or the Ets motif were found to be present in this module (and if so, which of the many variants of these motifs from TRANSFAC were present). The last two columns indicate the rank of the module by each method, based on the raw score (i.e., before modules were re-ranked based on other factors such as presence of at least one motif and the amount of short tandem repeat coverage).

Table S5[*]. CRM-level and nucleotide level sensitivity values for each of seven motif-blind methods on eight data sets of mouse enhancers.
Worksheet "localflanks.crm.sens": CRM-level sensitivity values. Worksheet "localflanks.crm.pval": Empirical p-values of CRM-level sensitivity. Worksheet "localflanks.nuc.sens": Nucleotide-level sensitivity values. Worksheet "localflanks.crm.pval": Empirical p-values of nucleotide-level sensitivity.

Table S6[*]. Cross-validation performance in mouse enhancers (complete data). The CRM and nucleotide level sensitivity and corresponding p-values of all methods in all 10 instantiations of cross-validation.

[*] These tables are included as separate Excel files: TableS5.xls, TableS6.xls

Table S7. Individual anatomy terms included in each data set of REDfly CRMs

amnioserosa (n=3):
amnioserosa
amnioserosa_anlage
amnioserosa_anlage_in_statu_nascendi

blastoderm (n=1):
blastoderm_embryo

cardiac (n=9):
cardiogenic_mesoderm
dorsal_vessel_specific_anlage
embryonic_dorsal_vessel
embryonic_heart
embryonic_pericardial_cell
larval_dorsal_vessel
larval_heart
larval_pericardial_cell
myocardial_cell

cns (n=21):
RP2_neuron
U_neuron
anterior_corner_cell
embryonic_brain
embryonic_central_brain
embryonic_central_brain_glia
embryonic_central_brain_mushroom_body
embryonic_central_nervous_system
embryonic_ganglion_mother_cell
embryonic_nervous_system
embryonic_neuroblast
embryonic_neuron
embryonic_ventral_nervous_system
interneuron
larval_central_nervous_system
larval_mushroom_body
larval_optic_lobe
neuroblast_NB6-4
neuron
thoracico-abdominal_ganglion
ventral_unpaired_cell

Imaginal-disc.1 (n=14):

ventral ectoderm (n=4):
ventral_ectoderm
ventral_ectoderm_anlage_in_statu_nascendi
ventral_ectodermal_primordium
ventral_epidermis_primordium

ectoderm.2 (n=16):
dorsal_ectoderm
dorsal_ectoderm_anlage
dorsal_ectoderm_anlage_in_statu_nascendi
dorsal_ectodermal_primordium
dorsal_epidermis_specific_anlage
ectoderm
ectoderm_anlage
ectoderm_anlage_in_statu_nascendi
embryonic_epidermis
procephalic_ectoderm_anlage
procephalic_ectoderm_anlage_in_statu_nascendi
procephalic_ectoderm_primordium
ventral_ectoderm
ventral_ectoderm_anlage_in_statu_nascendi
ventral_ectodermal_primordium
ventral_epidermis_primordium

mesoderm.2 (n=47):
abdominal_lateral_transverse_muscle
adepithelial_cell
adult_dorsal_vessel
adult_fat_body
adult_heart
adult_muscle_system
adult_pericardial_cell
adult_somatic_muscle
adult_visceral_muscle
cardiogenic_mesoderm
direct_wing_muscle
dorsal_internal_oblique_muscle_1
dorsal_vessel_specific_anlage
embryonic_dorsal_vessel
embryonic_fat_body
embryonic_heart
embryonic_muscle_system
embryonic_myoblast

antennal_disc
clypeo-labral_disc
eye_disc
eye-antennal_disc
eye-antennal_disc_specific_anlage
genital_disc
haltere_disc
imaginal_disk
labial_disc
male_genital_disc
ventral_thoracic_disc
ventral_thoracic_disc_specific_anlage
visual_primordium
wing_disc

ectoderm.1 (n=4):
ectoderm
ectoderm_anlage
ectoderm_anlage_in_statu_nascendi
embryonic_epidermis

endoderm (n=13):
anterior_endodermal_primordium
anterior_midgut_proper_primordium
embryonic_hindgut
embryonic_midgut
foregut_primordium
foregut_specific_anlage
larval_midgut
midgut_constriction_1
pHiGP2
posterior_embryonic_midgut
posterior_endoderm_anlage
posterior_endodermal_primordium
posterior_midgut_proper_primordium

mesoderm.1 (n=2):
mesoderm
mesoderm_anlage_in_statu_nascendi

neuroectoderm (n=2):
neuroectoderm
ventral_neuroectoderm

pns (n=15):
chemosensory_sensory_organ
embryonic_peripheral_nervous_system
mesothoracic_femoral_chordotonal_organ
metathoracic_femoral_chordotonal_organ
peripheral_nervous_system
embryonic_pericardial_cell
embryonic_somatic_muscle
embryonic_visceral_muscle
fat_body
fat_body/gonad_primordium
hemocyte_primordium
larval_dorsal_vessel
larval_fat_body
larval_heart
larval_muscle_system
larval_pericardial_cell
larval_somatic_muscle
larval_visceral_muscle
longitudinal_visceral_mesoderm
longitudinal_visceral_muscle_fibers
longitudinal_visceral_muscle_primordium
mesoderm
mesoderm_anlage_in_statu_nascendi
midgut_muscle
myocardial_cell
somatic_mesoderm
somatic_muscle_specific_anlage
tergotrochanteral_muscle
trunk_mesoderm_anlage
trunk_mesodermal_primordium
tubular_muscle
ventral_plate
visceral_mesoderm
visceral_muscle_primordium

neuronal (n=31):
RP2_neuron
U_neuron
anterior_corner_cell
chemosensory_sensory_organ
embryonic_brain
embryonic_central_brain
embryonic_central_brain_glia
embryonic_central_brain_mushroom_body
embryonic_central_nervous_system
embryonic_ganglion_mother_cell
embryonic_nervous_system
embryonic_neuroblast
embryonic_neuron
embryonic_peripheral_nervous_system
embryonic_ventral_nervous_system
interneuron
larval_central_nervous_system
larval_mushroom_body
larval_optic_lobe

prothoracic_femoral_chordotonal_organ
scolopophorous_organ
sensillum_campaniformium_of_anterior_crossvein
sensillum_campaniformium_of_dorsal_radius
sensillum_campaniformium_of_dorsal_radius_Sc1
sensory_organ_precursor_cell
twin_sensillum_of_margin_1
twin_sensillum_of_margin_2
ventral_cibarial_sense_organ
ventral_pit

tracheal-system (n=7):
embryonic_spiracle
embryonic_trachea
embryonic_tracheal_pit
embryonic_tracheal_system
filzkorper
posterior_spiracle_primordium
tracheal_primordium

mesothoracic_femoral_chordotonal_organ
metathoracic_femoral_chordotonal_organ
neuroblast_NB6-4
neuron
peripheral_nervous_system
prothoracic_femoral_chordotonal_organ
scolopophorous_organ
sensory_organ_precursor_cell
thoracico-abdominal_ganglion
ventral_cibarial_sense_organ
ventral_pit
ventral_unpaired_cell

wing (n=7):
sensillum_campaniformium_of_anterior_crossvein
sensillum_campaniformium_of_dorsal_radius
sensillum_campaniformium_of_dorsal_radius_Sc1
twin_sensillum_of_margin_1
twin_sensillum_of_margin_2
wing
wing_disc

Table S8. [next page]. [a] refers to percentage of CRMs that drove gene expression in the predicted pattern. na=not applicable to this method

[b] 5/7 (71%) gave enhancer activity, 1/7 was found to be a negative regulatory element

Note: Only studies in which at least 5 sequences were validated are included, and only those with true in vivo (not just cell culture) testing.

Study	Organism	Success Rate	Tissue Specificity[a]
(Halfon et al., 2002)	*Drosophila*	14% (1/7)	14%
(Berman et al., 2004)	*Drosophila*	33% (9/27)	22%
(Gaudet et al., 2004)	*C. elegans*	86% (6/7)[b]	100%
(Schroeder et al., 2004)	*Drosophila*	81% (13/16)	100%
(Ochoa-Espinosa et al., 2005)	*Drosophila*	79% (11/14)	100%
(Hallikas et al., 2006)	mouse	70% (7/10)	na
(Pennacchio et al., 2006)	mouse	45% (75/167)	na
(Philippakis et al., 2006)	*Drosophila*	50% (6/12)	33%
this study	*Drosophila* and mouse	100% (7/7)	57%

Table S9. Primer sequences used for Drosophila and mouse construct

Primer sequences used for *Drosophila* constructs

edl
F: CTATATTATTCCACGTTTTC
R: TGTCCAAGACGATTCTTAT

odd
F: GTTTTCAAATAAAATTACAATG
R: GCTAGGACGCAGAGCTG

srp
F: GGGAATTCCATTTATCTGA
R: CTTTTAATGCAACAATAAGC

SoxN
F: ATTACTTCGACTCCAGTC
R: CTTTCCTTTTAATCAAAGTG

cas
F: CATAAATATAATCAAATCTTAG
R: ATCGTACTCCGCCCCT

Primer sequences used for mouse constructs

C1orf164Fw	TAAGGATCCACAACCCCCTTATCCCTCAC
C1orf164Rv	TAAGTCGACCTGCTAGGACCCTGGAAGTG
EBF3Fw	TAAGGATCCTTTCAAAGAGCAACTGGGAC
EBF3Rv	TAAGATATCCGGTGGGCTATTGTTATAGGG

Table S10. Search Criteria used for defining expression gene sets

Blastoderm (BDGP)

select stages 1-6 only genes with blastoderm pattern as used in Halfon et. al (2002).

cardiac_mesoderm (BDGP)

cardiac mesoderm primordium cardiac mesoderm primordium dorsal vessel specific anlage embryonic/larval circulatory system embryonic/larval dorsal vessel	embryonic/larval pericardial cell lymph gland lymph gland specific anlage pericardial cell specific anlage

cns (BDGP)

embryonic brain embryonic central brain embryonic central brain embryonic central brain glia embryonic central brain glia embryonic central brain mushroom body embryonic central brain neuron embryonic central brain neuron embryonic central brain pars intercerebralis embryonic central brain surface glia embryonic central brain surface glia embryonic central nervous system embryonic ganglion mother cell	embryonic ganglion mother cell embryonic inner optic lobe embryonic inner optic lobe primordium embryonic optic lobe embryonic optic lobe primordium embryonic outer optic lobe embryonic outer optic lobe primordium mushroom body primordium neuroblasts of ventral nervous system pars intercerebralis primordium procephalic neuroblasts protocerebrum primordium ventral nerve cord ventral nerve cord primordium

eye.1 (FlyBase)

FBbt: 00004508 (eye)

imaginal_disc.1 (FlyBase)

FBbt: 00001761 (imaginal disc)

mesectoderm (BDGP)

mesectoderm anlage mesectoderm anlage in statu nascendi	mesectoderm primordium midline primordium

mesoderm.1 (BDGP)

head mesoderm P2 primordium, head mesoderm anlage,	trunk mesoderm anlage, trunk mesoderm anlage in statu nascendi,

head mesoderm in statu nascendi, mesoderm anlage in statu nascendi,	trunk mesoderm primordium

neuroectoderm (FlyBase)

a) FBbt: 00001061 (ventral neurogenic region) b) FBbt: 00005554 (ventral_nerve_cord_primordium)

pns (BDGP)

sensory nervous system primordium, sensory nervous system specific anlage

somatic muscle (FlyBase)

a) FBbt: 00000464 (embryonic/larval somatic muscle) b) FBbt: 00000129 (somatic mesoderm)

ventral_ectoderm (BDGP)

ventral ectoderm anlage ventral ectoderm anlage in statu nascendi	ventral ectoderm primordium ventral epidermis primordium ventral epidermis primordium

visceral_mesoderm (BDGP)

circular visceral muscle fibers embryonic/larval visceral muscle longitudinal visceral mesoderm primordium	longitudinal visceral muscle fibers visceral muscle primordium

ectoderm.2 (BDGP)

dorsal ectoderm anlage dorsal ectoderm anlage in statu nascendi dorsal ectoderm primordium dorsal epidermis primordium ectoderm anlage in statu nascendi embryonic dorsal epidermis procephalic ectoderm anlage	procephalic ectoderm anlage in statu nascendi procephalic ectoderm primordium ventral ectoderm anlage ventral ectoderm anlage in statu nascendi ventral ectoderm primordium ventral epidermis primordium

neuronal (BDGP)

embryonic central brain neuron, lateral cord neuron, ventral midline neuron

Supplementary Note S1. General-level characterization of various scoring schemes evaluated here.

Score category	**Scores in category**	**Novel?**	**What is new?**	**Based on all words or subset of words?**
Motif compendium-based	Clover-ClusterBuster	No	-	N/A
	Stubb-MDB	**Yes**	Motif selection procedure	
Markov chain based discrimination	HexMCD	No	-	All
	DiHexMCD	**Yes**	Discriminative training of Markov chains	
Dot product-based	D2z variants	**Yes**	Calculation of statistical significance	Either all or subset
Word selection, followed by Poisson statistics	PAC	**Yes**	Definition of score is new	Subset
Word selection, followed by weighted sum of counts	HexDiff	No	-	Subset
	HexYMF	**Yes**	Selection and weights of words	

Score category	**Scores in category**	**Statistical significance of score calculated?**	**CRM Model?**	**Background model for test sequence**
Motif compendium-based	Clover-ClusterBuster	Yes (use of LLR)	Yes (HMM)	Markov chain (low order)
	Stubb-MDB			
Markov chain based discrimination	HexMCD	Yes (use of LLR)	Yes (MC-5)	Markov chain (5^{th} order)
	DiHexMCD			
Dot product-based	D2z variants	Yes (z-score)	No	MC (low order)
Word selection, followed by Poisson statistics	PAC	Partially (Poisson p-values are combined heuristically)	No	MC (low order)
Word selection, followed by weighted sum of counts	HexDiff	No	No	No
	HexYMF			

Supplementary Note S2. Discussion on the choice of p≤0.05 as the threshold for statistically significant sensitivity

Note that Figure 1A marks a cell with a "*" whenever the sensitivity of the method on the data set has p-value ≤ 0.05. Given that we are testing 9 methods for each data set, there is an issue similar to "multiple hypotheses correction", and a concern whether 0.05 is "too liberal" a choice.

- Firstly, most (i.e., 90/118) of the p-values corresponding to the "*" cells are actually ≤ 0.01. (See Supplementary Table S3, worksheet "crm_pval.1"; the "<0.05" and "<0.01" cases are highlighted in darker and lighter shades respectively.).
- Secondly, we believe that 0.05 is a reasonable threshold for assessing significant performance of a given method on a given data set. The problem arises when we make global claims about the number of significant cells in Figure 1A; such claims must correct for "multiple hypotheses testing". The first such "global" claim we have made is that "… the best motif-blind methods … succeed on close to half of the data sets." We have checked that this is true if we use Q-values of Storey & Tibshirani (PNAS 2003) for multiple hypotheses correction. The q-value is, loosely speaking, the expected false discovery proportion corresponding to a p-value. For the three methods mentioned in context, the p-value threshold of 0.05 corresponds to a q-value threshold of 0.02, 0.10 and 0.07 respectively, thus validating our claim even after multiple hypotheses correction. The only other global claim we have made about the matrix of p-values (Figure 1A) is that 15 of the 31 data sets are amenable to at least one method (among the eight listed). For this claim,

we have required that the p-value threshold (of 0.05) be met by the method on each of four independent instantiations of cross-validation, thus effectively setting a much stricter threshold on each (method, data set) combination.

Finally, as a “back-of-the-envelope” calculation, we point out that there are 118 cells marked with a “*” in Figure 1(A), while at $p \leq 0.05$, one would expect $0.05 \times 9 \times 31 = 14$ such cells. Similarly, there are 90 cells with $p \leq 0.01$ (Supplementary Table S3), while only 3 are expected.

Supplementary Note S3. Motif-based prediction using an externally developed suite of algorithms

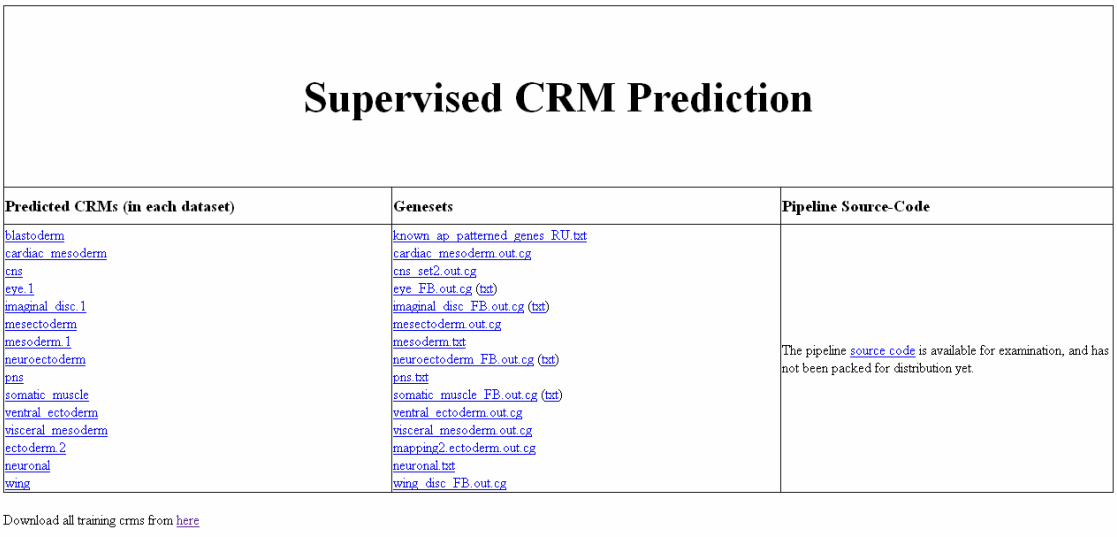
We have deployed our new method called "Stubb-MDB" as a point of comparison to the motif-blind approaches that are the main topic of this paper. Results showing that Stubb-MDB performs relatively poorly compared to the motif-blind methods tested here, may be found in Figure 1. In order to have great confidence in this comparison between motif-based and motif-blind approaches, we evaluated an alternative motif-based pipeline – that of using "Clover" for motif selection and "ClusterBuster" for scanning with selected motifs. Both programs are developed by Zhiping Weng and colleagues and run with default. The motif compendium used was the same as that for Stubb-MDB. ClusterBuster was run with default parameters. We found success levels to be comparable to that of Stubb-MDB. (See Supplementary Table S3, worksheets crm_sens.1 and crm_pval.1, showing the CRM-level sensitivity and p-values respectively with each method for each data set, on one of the cross-validation instantiations.) Here is a brief summary of the results of Stubb-MDB and Clover-ClusterBuster:

	Stubb-MDB	Clover-ClusterBuster
# data sets tested on	31	31
Average CRM-level sensitivity	28%	23%
Average CRM-level sensitivity over 15 amenable data sets	39%	35%
# data sets with p <= 0.05	10	9
Average nucleotide-level sensitivity	14%	17%
Average nucleotide-level sensitivity over 15 amenable data sets	25%	20%
# data sets with p <= 0.05	12	12

Supplementary Note S4. Difference between our benchmark and that of (Ivan et al., 2008).

In our new benchmark each test CRM is planted in a genomic sequence of length 10 Kbp, with G/C content similar to the native flank of the CRM (called target sequence). Therefore the search space for CRM is 10kb+CRM length. This is harder benchmark than our previous benchmark introduced in Ivan et. al (2008) in which each CRM is embedded in a genomic sequence of length 10 times CRM length. Since all of our datasets except cns, imaginal-disc-2, tracheal system, and wing have average CRM length of less than 1kb, the search space is usually shorter than the search space in our new benchmark.

Supplementary Note S5. Web Interface for the results, training datasets and sources



Supervised CRM Prediction

Predicted CRMs (in each dataset)	Genesets	Pipeline Source-Code
blastoderm cardiac_mesoderm cns eye.1 imaginal_disc.1 mesectoderm mesoderm.1 neuroectoderm pns somatic_muscle ventral_ectoderm visceral_mesoderm ectoderm.2 neuronal wing	known_ap_patterned_genes_RU.txt cardiac_mesoderm.out.cg cns_set2.out.cg eye_FB.out.cg (txt) imaginal_disc_FB.out.cg (txt) mesectoderm.out.cg mesoderm.txt neuroectoderm_FB.out.cg (txt) pns.txt somatic_muscle_FB.out.cg (txt) ventral_ectoderm.out.cg visceral_mesoderm.out.cg mapping2.ectoderm.out.cg neuronal.txt wing_disc_FB.out.cg	The pipeline source code is available for examination, and has not been packed for distribution yet.

Download all training crms from here

The web interface[2] includes (i) a link to predicted CRMs for each dataset which contains some information about the CRM (e.g. location, nearby gene, chromosome number, corresponding link to genome surveyor, flybase, BDGP and flyexp), the number of presented motifs, and link to presented motifs' logo. (ii) The genesets that have been used in this report (the search criteria used for defining the genesets are itemized in the next section) (iii) the pipeline source code that includes the implementation of all methods described in the paper and (iv) The training crms.

[2] http://veda.cs.uiuc.edu/scrm/index.htm

Supplementary Note S6. Novel CRMs in blastoderm data set

We report here on the overlap between the 113 modules predicted for the blastoderm data set by our pipeline and the A/P patterning modules predicted by a previous motif-based method (Sinha et al., 2004) and a non motif-based method (Grad et al., 2004). Note that the latter was based on two-species comparisons, while the predictions (considered here) from (Sinha et al., 2004) as well as our current work are based on D. melanogaster alone.

1) We have noted in text that the 5 tested fly CRMs were not reported by Ahab.
2) We considered the list of predictions made using Stubb (single species) in (Sinha et al., 2004).
3) We considered the list of predictions made by (Grad et al., 2004).

Our "novel" set of 54 predictions had very little overlap with the predictions of (Grad et al., 2004). There was a more substantial overlap (23/54) when considering all 2964 module predictions made by Stubb-SS in (Sinha et al., 2004). However, when we consider only the top 113 modules predicted by Stubb (i.e., the same number of predictions as our method), there is an overlap of only two modules. In other words, our novel set largely consists of unreported modules, although many of these were predicted at relatively low confidence levels in (Sinha et al., 2004) (Stubb). In any case, we designate these 54 modules as novel since none of them have been experimentally characterized previously.

	Our predictions with literature support (total 59)	**Our novel predictions (total 54)**
Overlap with Stubb (all 2964)	27	23
Overlap with Stubb (top 113)	14	2
Overlap with Grad et al (all 412)	24	5
Overlap with Grad et al (top 113)	10	3

Supplementary Note S7. Relationship between performance and methodological features

We performed pair-wise comparison of different variants of the D2-z score that differed in only one aspect (such as Markov order, or use of subsets of motifs, etc.), by counting the number of data sets (over all four instantiations) on which one method outperforms the other. The results are shown in Table S1 (Supplementary Materials). The most interesting observations from this table are that we get significantly better performance overall by (1) using a subset of words instead of all possible k-mers, (2) calculating conditional z-scores (of D2) rather than using unconditional z-scores as done in (cite), (3) keeping the CRMs in the training set separate, rather than concatenating them into one sequence, (4) using a first-order Markov order for the background model rather than an iid model, (5) counting words on both strands. Note that these features needed to be investigated only in the context of the D2-z score; the other scores such as HexDiff, HexYMF and PAC, do not face issues (1) – (4), and HexDiff and HexYMF do count words on both strands. On the other hand, the published D2-z score does not handle issues (1), (2), (3) and (5), and allows both iid and Markov backgrounds (issue 4) without prescribing which is better for our application.

Supplementary Note S8. The ten enhancers known to function in the developing blood and vasculature were taken from: (Chan et al., 2007; Chapman et al., 2003; De Val et al., 2004; Donaldson et al., 2005; Gottgens et al., 2002; Landry et al., 2008; Pimanda et al., 2007a; Pimanda et al., 2007b).

REFERENCES

Berman, B. P., Pfeiffer, B. D., Laverty, T. R., Salzberg, S. L., Rubin, G. M., Eisen, M. B. and Celniker, S. E. (2004). Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in Drosophila melanogaster and Drosophila pseudoobscura. *Genome Biol* **5**, R61.

Chan, W. Y., Follows, G. A., Lacaud, G., Pimanda, J. E., Landry, J. R., Kinston, S., Knezevic, K., Piltz, S., Donaldson, I. J., Gambardella, L. et al. (2007). The paralogous hematopoietic regulators Lyl1 and Scl are coregulated by Ets and GATA factors, but Lyl1 cannot rescue the early Scl-/- phenotype. *Blood* **109**, 1908-16.

Chapman, M. A., Charchar, F. J., Kinston, S., Bird, C. P., Grafham, D., Rogers, J., Grutzner, F., Graves, J. A., Green, A. R. and Gottgens, B. (2003). Comparative and functional analyses of LYL1 loci establish marsupial sequences as a model for phylogenetic footprinting. *Genomics* **81**, 249-59.

De Val, S., Anderson, J. P., Heidt, A. B., Khiem, D., Xu, S. M. and Black, B. L. (2004). Mef2c is activated directly by Ets transcription factors through an evolutionarily conserved endothelial cell-specific enhancer. *Dev Biol* **275**, 424-34.

Donaldson, I. J., Chapman, M. and Gottgens, B. (2005). TFBScluster: a resource for the characterization of transcriptional regulatory networks. *Bioinformatics* **21**, 3058-9.

Gaudet, J., Muttumu, S., Horner, M. and Mango, S. E. (2004). Whole-genome analysis of temporal gene expression during foregut development. *PLoS Biol* **2**, e352.

Gottgens, B., Nastos, A., Kinston, S., Piltz, S., Delabesse, E. C., Stanley, M., Sanchez, M. J., Ciau-Uitz, A., Patient, R. and Green, A. R. (2002). Establishing the transcriptional programme for blood: the SCL stem cell enhancer is regulated by a multiprotein complex containing Ets and GATA factors. *Embo J* **21**, 3039-50.

Grad, Y. H., Roth, F. P., Halfon, M. S. and Church, G. M. (2004). Prediction of similarly acting cis-regulatory modules by subsequence profiling and comparative genomics in Drosophila melanogaster and D.pseudoobscura. *Bioinformatics* **20**, 2738-50.

Halfon, M. S., Grad, Y., Church, G. M. and Michelson, A. M. (2002). Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated combinatorial model. *Genome Res* **12**, 1019-28.

Hallikas, O., Palin, K., Sinjushina, N., Rautiainen, R., Partanen, J., Ukkonen, E. and Taipale, J. (2006). Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. *Cell* **124**, 47-59.

Ivan, A., Halfon, M. S. and Sinha, S. (2008). Computational discovery of cis-regulatory modules in Drosophila without prior knowledge of motifs. *Genome Biol* **9**, R22.

Landry, J. R., Kinston, S., Knezevic, K., de Bruijn, M. F., Wilson, N., Nottingham, W. T., Peitz, M., Edenhofer, F., Pimanda, J. E., Ottersbach, K. et al. (2008). Runx genes are direct targets of Scl/Tal1 in the yolk sac and fetal liver. *Blood* **111**, 3005-14.

Ochoa-Espinosa, A., Yucel, G., Kaplan, L., Pare, A., Pura, N., Oberstein, A., Papatsenko, D. and Small, S. (2005). The role of binding site cluster strength in Bicoid-dependent patterning in Drosophila. *Proc Natl Acad Sci U S A* **102**, 4960-5.

Pennacchio, L. A., Ahituv, N., Moses, A. M., Prabhakar, S., Nobrega, M. A., Shoukry, M., Minovitsky, S., Dubchak, I., Holt, A., Lewis, K. D. et al. (2006). In vivo enhancer analysis of human conserved non-coding sequences. *Nature* **444**, 499-502.

Philippakis, A. A., Busser, B. W., Gisselbrecht, S. S., He, F. S., Estrada, B., Michelson, A. M. and Bulyk, M. L. (2006). Expression-guided in silico evaluation of

candidate cis regulatory codes for Drosophila muscle founder cells. *PLoS Comput Biol* **2**, e53.

Pimanda, J. E., Donaldson, I. J., de Bruijn, M. F., Kinston, S., Knezevic, K., Huckle, L., Piltz, S., Landry, J. R., Green, A. R., Tannahill, D. et al. (2007a). The SCL transcriptional network and BMP signaling pathway interact to regulate RUNX1 activity. *Proc Natl Acad Sci U S A* **104**, 840-5.

Pimanda, J. E., Ottersbach, K., Knezevic, K., Kinston, S., Chan, W. Y., Wilson, N. K., Landry, J. R., Wood, A. D., Kolb-Kokocinski, A., Green, A. R. et al. (2007b). Gata2, Fli1, and Scl form a recursively wired gene-regulatory circuit during early hematopoietic development. *Proc Natl Acad Sci U S A* **104**, 17692-7.

Schroeder, M. D., Pearce, M., Fak, J., Fan, H., Unnerstall, U., Emberly, E., Rajewsky, N., Siggia, E. D. and Gaul, U. (2004). Transcriptional control in the segmentation gene network of Drosophila. *PLoS Biol* **2**, E271.

Sinha, S., Schroeder, M. D., Unnerstall, U., Gaul, U. and Siggia, E. D. (2004). Cross-species comparison significantly improves genome-wide prediction of cis-regulatory modules in Drosophila. *BMC Bioinformatics* **5**, 129.

Supplementary Methods

March 10, 2009

1 Alignment free similarity measures

In this section we describe PAC and compute the z-score for the various modifications of the $D2$ statistic used in the text.

1.1 A Poisson based similarity measure, PAC.

This measure is based on [6]. Let A be a random DNA sequence. Fix a word length k. For a k-word w, let N_w^A be the count of the word w in the sequence A, including overlaps. A simple model for N_w^A is the binomial distribution, which is approximately a Poisson distribution in our applications. This approach assumes an iid background model. It does not take into account the dependency of occurrences of the word w in overlapping positions. Also, because of this independence assumption, the count of occurrences of a word w is done on one strand only.

The parameters for the Poisson distributions can be computed as follows. Let f_w^A be the prior probability of seeing the word w at any position of the sequence A. The expected number of occurrences is then $m_w^A = f_w(|A| - k + 1)$, where $|A|$ is the length of the sequence A. (Note: in the main text we used the notation $\lambda(w)$ instead of m_w.)

Now suppose A and B are two DNA sequences. Define

$$C_w^{AB} = \min(N_w^A, N_w^B).$$

Then, assuming the sequences A and B are generated independently,

$$\Pr(C_w^{AB} \geq c) = \begin{cases} (1 - F(m_w^A, c-1))(1 - F(m_w^B, c-1)) & \text{if } c > 0, \\ 1 & \text{otherwise,} \end{cases}$$

where $F(m, -)$ is the Poisson distribution function with mean m.

Now define a similarity measure as follows. Let

$$s_w^{AB} = 1 - \Pr(C_w^{AB} \geq c_w^{AB}),$$

where c_w^{AB} is the observed value of C_w^{AB}.

For a set W of k-words, the *additive Poisson* similarity measure between the sequences A and B is defined in [6] to be

$$S_{add}^{AB} = 1/|W| \sum_{w \in W} s_w^{AB}. \tag{1}$$

In our setting of LOOCV we have a dataset of n known CRMs planted in n random sequences. For each fold, we fix the sequence A to be a concatenation of $n-1$ CRMs. Sequence B is then a sliding window along the test sequence, where the window length is the average length of the CRMs in the given dataset. In this setting, the length of sequence A is so much longer than the length of B that in effect, $\min(N_w^A, N_w^B) = N_w^B$. Hence, the above Poisson measure was simplified to

$$PAC(B|A) = 1/|W| \sum_{w \in W} F(m_w^B, n_w^B - 1), \tag{2}$$

where n_w^B is the observed value of N_w^B. (Note: in the main text, we used the notation $n_B(w)$ instead of n_w^B).

The abbreviation PAC stands for "Poisson Additive Conditional", since the score is computed given that the sequence A is known. Note that the input from sequence A in PAC is only captured during the process of choosing the set of k-words W. For this we used the YMF motif finding program [5] and took the top 100 scoring k-words.

1.2 $D2z$ score

The $D2z$ score was developed in [1]. It is a z-score for the $D2$ statistic [3]

$$D2(A,B) =< \mathbf{N}^A, \mathbf{N}^B >, \tag{3}$$

where $\mathbf{N}^A$ ($\mathbf{N}^B$) is the count vector for the k-words in A (B), and $< -, - >$ is the inner product operator. The z-score is then

$$D2z(A,B) = \frac{D2(A,B) - E(D2(A,B))}{\sigma(D2(A,B))}, \tag{4}$$

where the (theoretical) expectation, $E(D2(A,B))$, and standard deviation, $\sigma(D2(A,B))$, are computed under the assumption that the random sequences A and B are independent and were generated by two, possibly different, Markov processes of any order (order zero means iid). The parameters of the Markov model are learned from the data.

1.3 Conditional $D2z$ score, $cD2z$.

When the sequence A is fixed (e.g. concatenated known CRMs), the formulas for the expectation and variance given in [1] are not valid. Here we compute the mean and the variance for $D2(A,B)$ where A is *fixed* and B is a random sequence generated by a Markov process. We call the corresponding z-score, $cD2z(B|A)$, a "conditional D2z" score. (Note: in the main text the $cD2z$ score is denoted by $D2z$-cond).

In what follows we use the following notation.

Notation. Let $n_1 = |A|$ be the length of sequence A and let $n_2 = |B|$. Write $\bar{n_1} = n_1 - k + 1$. and $\bar{n_2} = n_2 - k + 1$. Let I be the index set

$$I = \{(i,j) \in \mathbb{N} \times \mathbb{N} \,:\, 1 \le i \le \bar{n_1}, \text{ and } 1 \le j \le \bar{n_2}\}.$$

For indices $i \le c$, use the notation $A[i,c]$ for the subsequence of length $c - i + 1$, $A_i A_{i+1} \dots A_c$.

For a k-word w and an index $m < k$, define $suf_m(w)$ to be the last m letters in w,

$$suf_m(w) = w_{k-m+1} \dots w_k.$$

Define $pre_m(w)$ to be the first m letters in w,

$$pre_m(w) = w_1 \dots w_m.$$

IID case.

Proposition 1.1. *Under the IID model,*

$$E(cD2(B|A)) = \bar{n_2} \sum_{|w|=k} N_w^A \Pr^B(w), \tag{5}$$

where $\Pr^B(w)$ *is the probability of seeing the word* w *at any position in the sequence* B.

Proof. The $D2$ statistic in (3) can be computed via

$$D2(A,B) = \sum_{i=1}^{\bar{n_1}} \sum_{j=1}^{\bar{n_2}} Y_{(i,j)}, \tag{6}$$

where $Y_{(i,j)}$ is the k-word match indicator at position i in A and j in B,

$$Y_{(i,j)} = \begin{cases} 1 & \text{if } A[i, i+k-1] = B[j, j+k-1], \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$\begin{aligned} E(cD2(B|A)) &= \sum_{i=1}^{\bar{n_1}} \sum_{j=1}^{\bar{n_2}} E(Y_{ij}|A) = \sum_{i=1}^{\bar{n_1}} \sum_{j=1}^{\bar{n_2}} \Pr^B(A[i, i+k-1]) \\ &= \bar{n_2} \sum_{i=1}^{\bar{n_1}} \Pr^B(A[i, i+k-1]) = \bar{n_2} \sum_{|w|=k} N_w^A \Pr^B(w). \end{aligned}$$

□

Theorem 1.2. *Under the IID model,*

$$\begin{aligned} &\mathrm{Var}(cD2(B|A)) = \\ &2 \sum_{m=1}^{k-1} (n_2 - 2k + m + 1) \sum_{|v|=m} \Big(\sum_{\substack{|w|=k \\ suf_m(w)=v}} N_w^A \Pr^B(w) \Big) \Big(\sum_{\substack{|u|=k \\ pre_m(u)=v}} N_u^A \Pr^B(suf_{k-m}(u)) \Big) \\ &+ \bar{n_2} \sum_{|w|=k} (N_w^A)^2 \Pr^B(w) \\ &- 2 \sum_{m=1}^{k-1} (n_2 - 2k + m + 1) \left(\sum_{|w|=k} N_w^A \Pr^B(w) \right)^2 \\ &- \bar{n_2} \left(\sum_{|w|=k} N_w^A \Pr^B(w) \right)^2 \end{aligned}$$

Note that the above formula can be computed in $O(4^k)$ (the count vectors for the fixed sequence A are precomputed).

Proof.

$$\mathrm{Var}(D_2) = \mathrm{Var}(\sum_{(i,j)\in I} Y_{(i,j)}) = \sum_{(i,j),(s,t)\in I} \mathrm{Cov}(Y_{(i,j)}, Y_{(s,t)}).$$

case 1: no overlap in B.
In this case, since we are in the iid case, $Y_{(i,j)}$ and $Y_{(s,t)}$ are independent and hence $\mathrm{Cov}(Y_{(i,j)}, Y_{(s,t)}) = 0$.
case 2: overlap in B.
Let m be the size of the overlap. Then

$$E(Y_{(i,j)}Y_{(s,t)}) = \begin{cases} \mathrm{Pr}^B(A[i,i+k-1])\mathrm{Pr}^B(A[s+m,s+k-1]) \\ \qquad\qquad \text{if } A[i+k-m,i+k-1] = A[s,s+m-1], \\ 0 \qquad\qquad \text{otherwise.} \end{cases} \tag{7}$$

Hence, for a fixed j and $t = j+k-m$ (i.e. have overlap of size m in B),

$$\sum_{i=1}^{\bar{n_1}}\sum_{s=1}^{\bar{n_1}} E(Y_{(i,j)}Y_{(s,t)}) =$$
$$\sum_{|v|=m}(\sum_{\substack{|w|=k \\ suf_m(w)=v}} N_w^A \mathrm{Pr}^B(w))(\sum_{\substack{|u|=k \\ pre_m(u)=v}} N_u^A \mathrm{Pr}^B(suf_{k-m}(u)))$$

Add them all up:

$$\sum_{m=1}^{k}\sum_{\substack{j,t \\ |t-j|=k-m}}\sum_{i=1}^{\bar{n_1}}\sum_{s=1}^{\bar{n_1}} E(Y_{(i,j)}Y_{(s,t)}) =$$
$$2\sum_{m=1}^{k-1}(n_2-2k+m+1)\sum_{|v|=m}(\sum_{\substack{|w|=k \\ suf_m(w)=v}} N_w^A \mathrm{Pr}^B(w))(\sum_{\substack{|u|=k \\ pre_m(u)=v}} N_u^A \mathrm{Pr}^B(suf_{k-m}(u)))$$
$$+\bar{n_2}\sum_{|w|=k}(N_w^A)^2\mathrm{Pr}^B(w).$$

Finally subtract all the terms $E(Y_{(i,j)})E(Y_{(s,t)})$ with overlap in B:

$$\sum_{m=1}^{k} \sum_{\substack{j,t \\ |t-j|=k-m}} \sum_{i=1}^{\bar{n_1}} \sum_{s=1}^{\bar{n_1}} E(Y_{(i,j)}E(Y_{(s,t)})) =$$

$$\sum_{m=1}^{k} \sum_{\substack{j,t \\ |t-j|=k-m}} \left(\sum_{s=1}^{\bar{n_1}} \Pr^B(A[s, s+k-1]) \right) \left(\sum_{i=1}^{\bar{n_1}} \Pr^B(A[i, i+k-1]) \right) =$$

$$\sum_{m=1}^{k} \sum_{\substack{j,t \\ |t-j|=k-m}} \left(\sum_{|w|=k} N_w^A \Pr^B(w) \right)^2 =$$

$$2 \sum_{m=1}^{k-1} (n_2 - 2k + m + 1) \left(\sum_{|w|=k} N_w^A \Pr^B(w) \right)^2 + \bar{n_2} \left(\sum_{|w|=k} N_w^A \Pr^B(w) \right)^2 .$$

□

Higher order Markov model case. Here we state and prove the formulas for the mean and variance for $cD2z$ under Markov model (MM) of order 1. Extension to higher order MM is clear. We follow notation from [2] and [5]. For a DNA sequence B, and a nucleotide b let $p_j^B(b)$ be the probability of a b in position j. As justified in [2], we may assume p_j^B is independent of j and denote it by p^B. The probabilities $\{p^B(b)\}$ are calculated as the steady state probabilities of the background Markov model. For a k-word $w = w_1 w_2 \dots w_k$, write $p_*^B(w)$ for $\Pr(w|w_1)$, i.e., the probability of a k-word being w, given that its first letter is w_1. With this notation,

$$\Pr^B(w) = p^B(w_1) p_*^B(w) \tag{8}$$

Then the expectation of $cD2(-|A)$ can be expressed as follows.

Proposition 1.3. *Assume the sequence B is generated by a Markov model of order one. Then*

$$E(cD2(B|A)) = \bar{n_2} \sum_{|w|=k} N_w^A \Pr^B(w) = \bar{n_2} \sum_{|w|=k} N_w^A p^B(w_1) p_*^B(w.). \tag{9}$$

Proof. Same as Proposition 1.1, and use equation (8). □

In the next Theorem, we give a formula for the variance of $cD2$. We will use the following notation.

Notation. For two nucleotides, b and c, and an index $l > 0$ let $p_l^B(c|b)$ be the probability of seeing c, l steps in the Markov chain after b. Let

$$S_{b,c}^B = \sum_{j=1}^{q_2-1} \sum_{t=j+k}^{\bar{n_2}} p_l^B(c|b).$$

where $l = s - (i + k - 1)$ and $q_2 = n_2 - 2k + 2$.

Theorem 1.4. *Under Markov model of order one,*

$$\begin{aligned}
&\mathrm{Var}(cD2(B|A)) = \\
&2\sum_{w_k,u_1} S^B_{w_k,u_1} \sum_{w_1,\ldots,w_{k-1}} \mathrm{Pr}^B(w_1w_2\ldots w_k)N^A_{w_1w_2\ldots w_k} \sum_{u_2,\ldots,u_k} {p_*}^B(u_1u_2\ldots u_k)N^A_{u_1u_2\ldots u_k} \\
&+2\sum_{m=1}^{k-1}(n_2-2k+m+1)\sum_{|v|=m}\sum_{\substack{|w|=k\\ suf_m(w)=v}} N^A_w\mathrm{Pr}^B(w)\sum_{\substack{|u|=k\\ pre_m(u)=v}} N^A_u p^B(u_{m+1}|v_m){p_*}^B(suf_{k-m}(u)) \\
&+\bar{n_2}\sum_{|w|=k}(N^A_w)^2\mathrm{Pr}^B(w) \\
&-{\bar{n_2}}^2\left(\sum_{|w|=k} N^A_w\mathrm{Pr}^B(w)\right)^2
\end{aligned} \tag{10}$$

Proof.

$$\mathrm{Var}(D_2) = \mathrm{Var}(\sum_{(i,j)\in I} Y_{(i,j)}) = \sum_{(i,j),(s,t)\in I} \mathrm{Cov}(Y_{(i,j)}, Y_{(s,t)}).$$

case 1: no overlap in B.
Unlike the iid case, $Y_{(i,j)}$ and $Y_{(s,t)}$ may be dependent under the Markov Model. To compute $\mathrm{Cov}(Y_{(i,j)}, Y_{(s,t)})$, we consider the following. For a fixed $j < t$, with $t = j+k-1+l$ and $l > 0$ (i.e. no overlap in B and l is the "gap" from the last letter of the first word in B to the the first letter of the second word in B),

$$\begin{aligned}
&\sum_{i=1}^{\bar{n_1}}\sum_{s=1}^{\bar{n_1}} E(Y_{(i,j)}Y_{(s,t)}) = \\
&\sum_{i=1}^{\bar{n_1}}\sum_{s=1}^{\bar{n_1}} \mathrm{Pr}(B[j,j+k-1] = A[i,i+k-1], B[t,t+k-1] = A[s,s+k-1]) = \\
&\sum_{i=1}^{\bar{n_1}}\sum_{s=1}^{\bar{n_1}} p^B(A_i){p_*}^B(A[i,i+k-1]){p_*}^B(A[s,s+k-1])p^B_l(A_s|A_{i+k-1}) = \\
&\sum_{|w|=k}\sum_{|u|=k} N^A_w N^A_u p^B(w_1){p_*}^B(w){p_*}^B(u)p^B_l(u_1|w_k).
\end{aligned}$$

Sum up over all such j's and t's:

$$\begin{aligned}
&\sum_{\substack{j,t\\|t-j|\geq k}} \sum_{i=1}^{\bar{n_1}}\sum_{s=1}^{\bar{n_1}} E(Y_{(i,j)}Y_{(s,t)}) = \\
&2\sum_{j=1}^{q_2-1}\sum_{t=j+k}^{\bar{n_2}}\sum_{|w|=k}\sum_{|u|=k} N_w^A N_u^A p^B(w_1){p_*}^B(w){p_*}^B(u)p_l^B(u_1|w_k) = \\
&2\sum_{|w|=k}\sum_{|u|=k} N_w^A N_u^A p^B(w_1){p_*}^B(w){p_*}^B(u)\sum_{j=1}^{q_2-1}\sum_{l=1}^{q_2-j} p_l^B(u_1|w_k) = \\
&2\sum_{|w|=k}\sum_{|u|=k} N_w^A N_u^A p^B(w_1){p_*}^B(w){p_*}^B(u)S^B_{w_k,u_1}.
\end{aligned} \tag{11}$$

The computational complexity of the formula above, as a function of k, is $O(4^{2k})$. To reduce this to $O(4^k)$ we split the sum as follows.

$$\begin{aligned}
&2\sum_{|w|=k}\sum_{|u|=k} N_w^A N_u^A p^B(w_1){p_*}^B(w){p_*}^B(u)S^B_{w_k,u_1} = \\
&2\sum_{w_k,u_1} S^B_{w_k,u_1}\sum_{w_1,\ldots,w_{k-1}} \Pr^B(w_1w_2\ldots w_k)N^A_{w_1w_2\ldots w_k}\sum_{u_2,\ldots,u_k}{p_*}^B(u_1u_2\ldots u_k)N^A_{u_1u_2\ldots u_k}.
\end{aligned}$$

This gives the first term of equation (10).

case 2: overlap in B.

The next two terms in formula (10) come from $E(Y_{(i,j)}Y_{(s,t)})$ where the k-words in sequence B overlap. The proof follows the proof of the IID case (Theorem 1.2) with the obvious changes for the Markov model.

Finally subtract all the terms $E(Y_{(i,j)}E(Y_{(s,t)}))$:

$$\begin{aligned}
\sum_{i=1}^{\bar{n_1}}\sum_{j=1}^{\bar{n_2}}\sum_{s=1}^{\bar{n_1}}\sum_{t=1}^{\bar{n_2}} E(Y_{(i,j)}E(Y_{(s,t)}) &= \left(\sum_{i=1}^{\bar{n_1}}\sum_{j=1}^{\bar{n_2}} E(Y_{(i,j)})\right)\left(\sum_{s=1}^{\bar{n_1}}\sum_{t=1}^{\bar{n_2}} E(Y_{(s,t)})\right) \\
&= \left(\sum_{i=1}^{\bar{n_1}}\sum_{j=1}^{\bar{n_2}} E(Y_{(i,j)})\right)^2 \\
&= \left(\bar{n_2}\sum_{|w|=k} N_w^A \Pr^B(w)\right)^2.
\end{aligned}$$

This is the last term in (10). □

Remark 1.5. The formula for the variance in equation (10) has computational complexity of $O(n_2^2 4^k)$ due to the $O(n_2^2)$ computation of $S^B_{b,c}$. To reduce complexity we use the following approximation. By [2], when the length of the sequence is large enough,

$$S^B_{b,c} \approx \frac{q_2(q_2+1)}{2}p^B(c) - (q_2-1)\mathbf{e}_b'\mathbf{Q}\mathbf{P}\mathbf{e}_c - \mathbf{e}_b'\mathbf{Q}\mathbf{P}^2\mathbf{Q}\mathbf{e}_c, \tag{12}$$

where $\mathbf{P}$ is the $d \times d$ matrix of transition probabilities, where d is the size of our alphabet (i.e. $d = 4$) $\mathbf{Q} = (\mathbf{P} - \mathbf{I} + \mathbf{1p}')^{-1}$, with $\mathbf{1}' = (1, \ldots, 1)$ and $\mathbf{p}$ is the vector of steady state probabilities. The vector $\mathbf{e}_c$ is the unit vector that picks up the column corresponding to c.

With this approximation, formula (10) can be computed in $O(4^k)$ time.

1.4 Conditional $D2z$ with weights, $wD2z$.

In $cD2(B|A)$ of section 1.3, we give every k-word the same weight. In a more realistic setting, we would like to give more weight to some words and less to others. For this, we define a weighted conditional $D2$ measure, $wcD2(B|A)$ or $wD2(B|A)$ for short, and compute its z-score. We assume the weights are learned from the sequence A. Hence we make the following definition. (Note: in the main text $wD2z$ is called $D2z$-cond-weights).

Definition 1.6. Let A and B be two DNA sequences. For each k-word w, let a_w be a weight associated with w. Define a weighted D2 score by

$$wD2(B|A) = \sum_{|w|=k} a_w N_w^A N_w^B.$$

In terms of the match indicator random variables,

$$wD2(B|A) = \sum_{i=1}^{\bar{n_1}} \sum_{j=1}^{\bar{n_2}} a_{ij} Y_{(i,j)}, \tag{13}$$

where a_ij is the weight of the k-word $A[i, i+k-1]$.

Next we compute the z-score for $wD2(-|A)$.

IID case.

Theorem 1.7. *Under the IID model,*

$$E(wD2(B|A)) = \bar{n_2} \sum_{|w|=k} a_w N_w^A \Pr^B(w), \tag{14}$$

and

$$\begin{aligned}
&\mathrm{Var}(wD2(B|A)) = \\
&2\sum_{m=1}^{k-1}(n_2 - 2k + m + 1) \sum_{|v|=m} \Big(\sum_{\substack{|w|=k \\ suf_m(w)=v}} a_w N_w^A \Pr^B(w) \Big) \Big(\sum_{\substack{|u|=k \\ pre_m(u)=v}} a_u N_u^A \Pr^B(suf_{k-m}(u)) \Big) \\
&+ \bar{n_2} \sum_{|w|=k} a_w^2 (N_w^A)^2 \Pr^B(w) \\
&- 2\sum_{m=1}^{k-1}(n_2 - 2k + m + 1) \left(\sum_{|w|=k} a_w N_w^A \Pr^B(w) \right)^2 \\
&- \bar{n_2} \left(\sum_{|w|=k} a_w N_w^A \Pr^B(w) \right)^2
\end{aligned} \tag{15}$$

Note that the above formula can be computed in $O(4^k)$.

Proof. Same as the proof of Proposition 1.1 and Theorem 1.2, while keeping track of the weights. □

Markov Model case.

Theorem 1.8. *Under Markov model of order one and with notation as in Theorem 1.4,*

$$E(wD2(B|A)) = \bar{n_2} \sum_{|w|=k} a_w N_w^A \Pr^B(w) = \bar{n_2} \sum_{|w|=k} a_w N_w^A p^B(w_1) {p_*}^B(w.). \tag{16}$$

For the variance formula, to simplify notation, write a_w for $a_{w_1 \dots w_k}$ and a_u for $a_{u_1 \dots u_k}$. Then

$$\begin{aligned}
&\mathrm{Var}(wD2(B|A)) = \\
&2 \sum_{w_k, u_1} S^B_{w_k, u_1} \sum_{w_1, \dots, w_{k-1}} \Pr^B(w_1 \dots w_k) a_w N^A_{w_1 \dots w_k} \sum_{u_2, \dots, u_k} {p_*}^B(u_1 \dots u_k) a_u N^A_{u_1 \dots u_k} \\
&+ 2 \sum_{m=1}^{k-1} (n_2 - 2k + m + 1) \sum_{|v|=m} \left(\sum_{\substack{|w|=k \\ suf_m(w)=v}} a_w N_w^A \Pr^B(w) \right) \\
&\qquad \left(\sum_{\substack{|u|=k \\ pre_m(u)=v}} a_u N_u^A p^B(u_{m+1}|v_m) {p_*}^B(suf_{k-m}(u)) \right) \\
&+ \bar{n_2} \sum_{|w|=k} a_w^2 (N_w^A)^2 \Pr^B(w) \\
&- {\bar{n_2}}^2 \left(\sum_{|w|=k} a_w N_w^A \Pr^B(w) \right)^2
\end{aligned} \tag{17}$$

Proof. Same as the proof of Proposition 1.3 and Theorem 1.4, while keeping track of the weights. □

As in Remark 1.5, an approximation of formula (17) for the variance can be computed in $O(4^k)$.

References

[1] Kantorovitz, M.R, Robinson, G.E. and Sinha S. (2007) A statistical method for alignment-free comparison of regulatory sequences. *Bioinformatics* **23** No. 13, pp. i249-255.

[2] Kleffe, J. and Borodovsky, M. (1992). *Comput. Appl. Biosci.,* **8**, 433–441.

[3] Lippert, R.A., Huang, H. and Waterman, M.S. (2002). Distributional regimes for the number of k-word matches between two random sequences, *Proc. Natl. Acad. Sci. USA* **99** 13980–13989.

[4] Sinha, S. and Tompa, M. (2000). A statistical method for finding transcription factor binding sites. *Proc Int Conf Intell Syst Mol Biol.* **8**:344-54.

[5] Sinha, S., Schroeder, M.D., Unnerstall, U., Gaul, U., and Siggia, E.D.(2004). Cross-species comparison significantly improves genome-wide prediction of cis-regulatory modules in Drosophila. *BMC Bioinformatics.* **5:129**.

[6] van Helden, J. (2004) Metrics for comparing regulatory sequences on the basis of pattern counts. *Bioinformatics* **20**, 399–406.