Our goal was to determine whether binding-site motifs and modules could be used to predict tissue-specific differential expression and to identify motifs and modules that explain differential expression. Our hypothesis is that binding sites for activating (inhibiting) transcription factors would be significantly over-represented in promoters of transcripts with tissue-specific elevated (inhibited) expression. We construct predictive logic from motifs that best distinguish promoters of transcripts with elevated expression from promoters of transcripts with inhibited expression.

This document describes results that were omitted from the paper, and it describes methods in more detail than would fit in the body of the paper. Results given here include samples from the catalog of experimentally validated motifs and modules, *in silico*-identified motifs and modules, and predictive models based on motifs and modules (referred in the paper as predictors). All are used to predict tissue-specific differential expression.

1 Results

Only a limited selection of results could be presented in the body of the paper and in this supplementary information document. The remaining results are available directly from the authors.

1.1 Tissue differential promoter sets and transcript lists corresponding to predictor calls

Proximal promoters, in FASTA format, for each of the 56 tissues are available directly from the authors. Positive promoter sets (foreground) are denoted with an "fg" suffix, and negative promoter sets (background) are denoted with a "bg" suffix. Human sets have Hs before the suffix, and mouse sets have Mm before the suffix. For example, the positive promoter set for human trachea is called Trachea.Mm.fg, and the negative set for mouse ovary is called Ovary.Mm.bg. The promoter sets are compressed and will unpack to a directory called Promoters.

For each tissue, we provide (by request) the transcript lists predicted as true positive (corresponding promoters were in the positive set and were predicted to be positive by the predictive model for that tissue), false positive (negative set but predicted positive), true negative (negative set predicted negative), and false negative (positive set predicted negative).

1.2 Error Table

Table 2 is analogous to Table 1 and describes results for mouse tissues. It gives the number of terms selected by multivariate adaptive regression splines (MARS) for building the predictor to minimize prediction error, the classification error using this number of MARS terms, the prediction error according to 10-fold crossvalidation, and the corresponding Bonferroni corrected P value (corrected for MARS term selection).

1.3 Top TRANSFAC single motif predictors

We give the top 10 experimentally validated vertebrate motifs from TRANSFAC (1) for each tissue. Motifs are ranked by classification error, and similar lower ranking motifs are eliminated. The first column gives the motif accession from TRANSFAC. The second column gives the name of the associated factor or factor class, which may be edited for presentability. The third column gives the sequence logo built from the position weight matrix (PWM) for the motif. The fourth column indicates whether the motif was enriched in the foreground (positive set) or background (negative set). Foreground enrichment means that the motif

Table 2. DNA patterns in mouse proximal promoters predict tissue specific expression

Tissue	Terms	Err	PredctErr	P val
Pancreas	2	0.323	0.324	1.0e-28
Ovary	4	0.296	0.329	1.8e-27
Liver	2	0.354	0.363	6.3e-18
Adrenal gland	2	0.352	0.368	1.0e-16
Uterus	4	0.339	0.371	5.3e-16
Thyroid	3	0.343	0.374	2.6e-15
Bone marrow	3	0.371	0.382	1.5e-13
Adipose tissue	7	0.387	0.385	6.5e-13
Thymus	4	0.359	0.385	6.5e-13
Amygdala	2	0.403	0.391	1.1e-11
Testis	7	0.321	0.394	4.1e-11
Lymph node	5	0.342	0.405	4.1e-09
Olfactory bulb	4	0.352	0.405	4.1e-09
Salivary gland	2	0.374	0.409	2.0e-08
Lung	7	0.351	0.412	6.0e-08
Dorsal root ganglia	2	0.394	0.418	5.1e-07
CD4 T cells	2	0.407	0.426	7.1e-06
Kidney	3	0.389	0.434	7.6e-05
Cerebellum	7	0.379	0.437	1.7e-04
Placenta	2	0.438	0.441	6.4e-04
Hypothalamus	4	0.379	0.445	1.7e-03
Prostate	2	0.408	0.448	2.7e-03
CD8 T cells	2	0.427	0.456	1.5e-02
Pituitary	2	0.430	0.459	3.1e-02
Trachea	4	0.399	0.467	1.2e-01
Skeletal muscle	2	0.438	0.478	4.6e-01
Heart	5	0.386	0.485	9.8e-01
Trigeminal ganglion	5	0.405	0.492	1.0e+00

For each mouse tissue, we present the number of MARS terms selected for building the predictor to minimize prediction error, the classification error using this number of MARS terms, the prediction error according to 10-fold crossvalidation, and the corresponding Bonferroni corrected P value (corrected for MARS term selection). After correction, predictors for CD8 T cells, pituitary, trachea, skeletal muscle, heart, and trigeminal ganglion fail to predict significantly ($P > 0.01$). Err stands for the classification error and PredctErr is the prediction error.

Table 3. Ten TRANSFAC motifs with lowest classification error in human CD4 T cells

Acces.#	Factor	Logo	Enrich	Error	Sens	Spec
M00032	c-Ets-1(p54)		FG	0.409	0.346	0.836
M00017	ATF		FG	0.425	0.682	0.468
M00480	LUN-1		FG	0.437	0.768	0.358
M00793	YY1		FG	0.439	0.548	0.574
M00652	Nrf-1		FG	0.441	0.204	0.914
M00316	Imperfect		FG	0.445	0.348	0.762
M00175	AP-4		BG	0.445	0.510	0.600
M00431	E2F-1		FG	0.446	0.786	0.322
M00993	TAL1		BG	0.447	0.592	0.514
M00769	AML		FG	0.451	0.374	0.724

For each motif, we give the TRANSFAC accession (Acces.#), corresponding binding factor (Factor), sequence logo built from the PWM for the motif (Logo), indication whether the motif was enriched in the foreground or background (Enrich), and the classification error (Error) broken down to sensitivity (Sens) and specificity (Spec).

has optimal classification performance when assigning sequences with max-score greater than the optimal threshold to the foreground; similarly for background enrichment. Columns five through seven give the classification error, the sensitivity and the specificity, respectively.

The top 10 experimentally validated motifs for human CD4 T cells and human liver are given in Tables 3 and 4. Motifs for the remaining tissues are available from the authors.

1.4 Top TRANSFAC distinct motif pair predictors

We give the top five modules constructed from experimentally validated vertebrate motifs from TRANSFAC for each tissue. Modules are ranked by classification error, and modules that contain a motif similar to any motif in a higher-ranking module are eliminated. The set of columns is divided into three sections; the first two sections present each of the two motifs as described in Section 1.3, and the third section gives enrichment, classification error, sensitivity, and specificity.

The top 5 modules for human CD4 T cells and human liver are given in Tables 5 and 6. Modules for the remaining tissues are available directly from the authors.

1.5 Top motif predictors

As in Section 1.3, we give the top 10 motifs for each tissue, but we also include motifs identified *in silico* by DME (2) and DME-B; DME-B is a modified version of DME that considers only the best occurrences in each sequence. We name TRANSFAC motifs by their accession, and motifs identified *in silico* are given a name consisting of the prefix "Novel-" and an index. Some novel motifs resemble experimentally validated motifs (see Section 2.2.1). For these motifs, we assign a factor corresponding to the experimentally validated motif and describe the divergence between the motif and the experimentally validated binding site for the factor. We speculate that the assigned factor may bind to sites characterized by the novel motif, because it binds to the sites characterized by the similar experimentally validated motif. When several motifs have

Table 4. Ten TRANSFAC motifs with lowest classification error in human liver

Acces.#	Factor	Logo	Enrich	Error	Sens	Spec
M00918	E2F		BG	0.405	0.686	0.504
M00248	Oct-1		BG	0.406	0.664	0.524
M00071	E47		FG	0.415	0.584	0.586
M00646	LF-A1		FG	0.416	0.628	0.540
M00145	Brn-2		BG	0.417	0.586	0.580
M00103	Clox		BG	0.417	0.738	0.428
M00765	COUP		FG	0.421	0.492	0.666
M00318	Lentiviral		BG	0.421	0.648	0.510
M00129	HFH-1		BG	0.422	0.578	0.578
M00025	Elk-1		BG	0.422	0.622	0.534

For each motif, we give the TRANSFAC accession (Acces.#), corresponding binding factor (Factor), sequence logo built from the PWM for the motif (Logo), indication whether the motif was enriched in the foreground or background (Enrich), and the classification error (Error) broken down to sensitivity (Sens) and specificity (Spec).

Table 5. Five TRANSFAC motif pairs with lowest classification error in human CD4 T cells

Acces.#	Factor	Logo	Acces.#	Factor	Logo	Enrich	Error	Sens	Spec
M00017	ATF		M00793	YY1		FG	0.403	0.416	0.778
M00431	E2F-1		M00480	LUN-1		FG	0.403	0.602	0.592
M00032	Ets-1		M00423	FOXJ2		FG	0.406	0.328	0.860
M00257	RREB-1		M00993	TAL1		BG	0.425	0.472	0.678
M00147	HSF2		M00175	AP-4		BG	0.427	0.446	0.700

For each motif pair, we give the TRANSFAC accessions (Acces.#), corresponding binding factors (Factor), sequence logos built from the PWM for each motif (Logo), indication whether the motif was enriched in the foreground or background (Enrich), and the classification error (Error) broken down to sensitivity (Sens) and specificity (Spec).

Table 6. Five TRANSFAC motif pairs with lowest classification error in human liver

Acces.#	Factor	Logo	Acces.#	Factor	Logo	Enrich	Error	Sens	Spec
M00103	Clox		M00918	E2F		BG	0.364	0.554	0.718
M00465	POU6F1		M00695	ETF		BG	0.376	0.648	0.600
M00071	E47		M00646	LF-A1		FG	0.384	0.602	0.630
M00129	HFH-1		M00224	STAT1		BG	0.385	0.714	0.516
M00025	Elk-1		M00248	Oct-1		BG	0.389	0.496	0.726

For each motif pair, we give the TRANSFAC accessions (Acces.#), corresponding binding factors (Factor), sequence logos built from the PWM for each motif (Logo), indication whether the motif was enriched in the foreground or background (Enrich), and the classification error (Error) broken down to sensitivity (Sens) and specificity (Spec).

divergence lower than 1.0, we present the best match (lowest divergence). Motifs are available directly from the authors.

1.6 Top distinct motif pair predictors

As in Section 1.4, we give the top five modules for each tissue, but we include motifs identified *in silico* by DME and DME-B. Motifs are named as in Section 1.5, and eliminated as described in Section 1.4. Modules are available directly from the authors.

1.7 Predictive models and interpretation

The predictive model (called predictors in the main paper) for each tissue is described as a set of modules and the MARS function that models their interaction. For each predictive model, we describe the modules and features (see Section 2.3.1) and the motifs of which they are composed. The MARS function is given immediately below; it describes the interaction between the modules and is used to predict elevated and inhibited transcription. The function is broken into its terms (see Section 2.3.2), and each is given an interpretation. Here we describe more in depth analysis for human CD8 T cells and testis. Predictive models for all 56 human and mouse tissues are available directly from the authors.

1.7.1 Human testis predictive model

The human testis predictive model includes two modules that are composed of novel motifs. Differential expression in human testis is predicted with an error rate of 0.397 ($P < 1.5\text{E-}10$), compared to 0.441 using only experimentally verified motifs. The predictor includes two terms, *Te1* and *Te2*, each describing a synergistic relation between motifs pairs. *Te1* is the most significant contributor to prediction and it is overrepresented in the positive set. *Te2* is overrepresented in the negative set. Fig. 4 describes the contributions of *Te1* and *Te2* to final prediction. A predictor that uses *Te2* alone has high sensitivity and low specificity, and a predictor that uses *Te1* alone leads to higher specificity but lower sensitivity. The combination of *Te1* and *Te2* results in higher sensitivity than *Te1*-based prediction and higher specificity than either *Te1*- or *Te2*-based prediction. *Te1* includes C/G-rich motifs and *Te2* includes A/T-rich motifs. However, base composition alone is not enough to explain prediction quality; base composition has a classification error of 0.416 compared to 0.357 for the predictor. *Te1* motifs are compensatory: their combination increases sensitivity at the expense of decreasing specificity. Despite their similarity, they capture dependent but distinct signals. The overlap rate between the highest-likelihood binding sites of the two motifs is not enriched over what is expected by chance, and attempts to join the motifs resulted in weakened predictive ability, suggesting that an additive relation is the best way to capture their signal.

1.7.2 Predictive models for human CD8 T cells

As an example, consider the predictive model for human CD8 T cells give in Table 7 and Fig. 5. The predictor includes two modules $X1$ and $X2$. $X1$ is composed of two TRANSFAC motifs M00743 and M00793 using the max-score-sum feature. $X2$ is composed of two TRANSFAC motifs M00277 and M00691 using the max-score-product feature. The MARS function has three terms:

1. This positive constant term indicates that, by default, a promoter will be called positive.

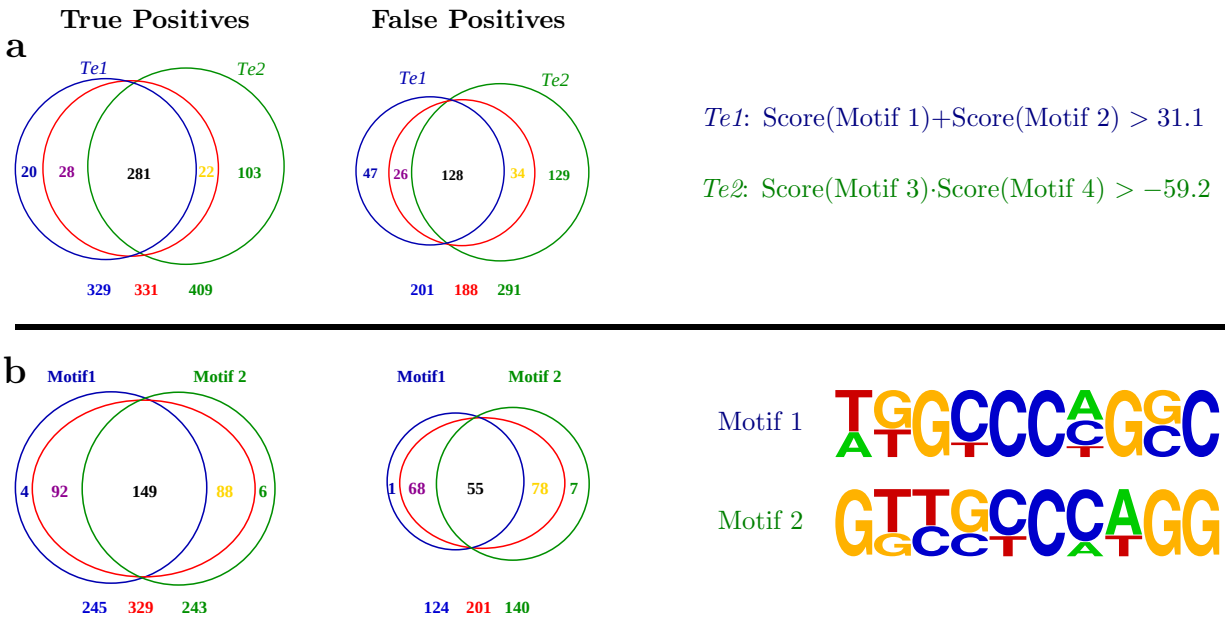


Fig. 4. (**a**) Human testis predictor constraints and the distribution of true positives and false positives across its two modules *Te1* and *Te2*. Motifs in *Te1* are overrepresented in the positive set and motifs in *Te2* are overrepresented in the negative set. Promoters that satisfy the conditions on *Te1* (*Te2*) are more likely to be predicted as positive (negative). Selections by *Te1*, the predictor, and *Te2* are indicated in blue, red, and green, and total set sizes are given immediately below. Compared with predictions based on *Te1* alone, predictions by the testis predictor include more true positives (331 vs. 329) and fewer false positives (188 vs. 201). Compared to *Te2* alone, predictions by the testis predictor include considerably fewer false positives (188 vs. 291). (**b**) Motifs included in *Te1* and prediction statistics for *Te1* across its two motif components. These motifs share a similar core and differ only in the left- and right-most flanking positions. Our experiments suggest that these two motifs function in a compensatory manner best modeled with an additive relationship.

2. This term is negative if the max-score-sum of M00743 and M00793 ($X1$) is less than 24.5, implying that if the sum of the scores of the highest-scoring substrings of a given promoter for M00743 and M00793 is lower than 24.5, this term will contribute to a negative call.

3. This term is negative if $X1$ is less than 24.5 and the max-score-product of M00277 and M00691 ($X2$) is less than 45.1, implying that this term will contribute to a negative call in the event that both modules score low. In our interpretation, this term will have an impact (it will further contribute to a negative call) only if the first term's negative contribution is smaller than the constant term, and $X2$ is less than 45.1.

It is important to remember that the interpretation in Fig. 5 describes the effect of each term on prediction, but terms do not act in isolation.

1.8 Human and mouse promoter set intersection

Some tissue-specific promoter sets include a large number of common promoters; for example positive sets for human CD4 and CD8 T cells include 70% common promoters. Common promoters correspond to common transcripts, and are a measure of tissue-specific expression similarity. We expect that tissues with foreground or background promoter sets that have a high similarity in composition will also have similar predictive models. Tables 8 and 9 give the proportion of similar promoters between each human tissue pair and mouse tissue pair. The upper triangle describes the similarity in composition of the positive sets (FG), and the lower describes the similarity in composition of the negative sets (BG).

1.9 Mouse heat map

The heat map in Fig. 6 is the mouse counterpart to the human heat map given in Fig. 3. It has a few changes in tissue order to address prediction similarity in the mouse tissues. The tissue set is also different, because a different set of predictors in mouse failed the significance test.

2 Methods

This section describes methods used to prepare the data, identify patterns (motifs and motif modules), and build predictive models based on those patterns. All programs named below are freely available under the GNU Public License as part of Comprehensive Regulatory Element Analysis and Discovery (CREAD) (http://cread.sf.net), with the exception of DME and DME-B (see Section 2.2.3), which are available from the authors.

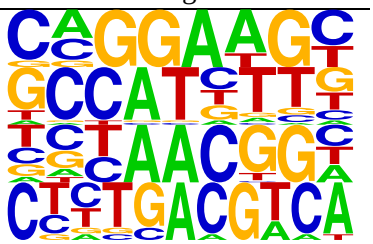
2.1 Data Preparation

This section describes our methodology for obtaining sets of promoter sequences for transcripts showing differential expression in tissue specific microarray experiments. The expression data discussed here are due to Su *et al.* (3). There are three major steps in obtaining sequence sets from the expression data:

1. Mapping probes to transcripts.

2. Mapping transcripts to promoters.

Table 7. TRANSFAC MARS classifiers for human CD8 T cells

Module	Feature	MotifName	Factor	Logo
X_1	SUM	M00743	c-Ets-1	
		M00793	YY1	
X_2	PRODUCT	M00227	v-Myb	
		M00691	ATF-1	

MARS function terms ($f(x) = \max(x, 0)$)**:**

Term	**Interpretation**
+ 0.495	FG by default
- 0.0788 f(24.5 - X_1)	BG if X_1 scores low
- 0.00355 f(24.5 - X_1) f(45.1 - X_2)	BG if X_1 scores low and X_2 scores low

Fig. 5. MARS predictor for human CD8 T cells.

3. Selecting sets of tissue-specific transcripts.

The objective is to obtain sets of sequences with unusually high and unusually low frequencies of binding-site patterns that are associated with the observed tissue-specific differential expression.

2.1.1 Mapping probes to transcripts

We map Affymetrix probes to RefSeq transcripts in an attempt to identify isoforms detected by the UniGene-centric arrays. This step is necessary, because many genes have tissue-specific isoforms (4–6), and because different isoforms often have different first exons and should be assigned different promoters. We focused exclusively on RefSeq transcripts because the vast majority of Su *et al.* (3) probes can be associated with RefSeq transcripts (7) (see Table 10), and because corresponding Cold Spring Harbor Laboratory Mammalian Promoter Database (CSHLmpd) (8) transcription start site (TSS) annotation can almost always be mapped to RefSeq transcripts.

We use the expression data to select sets of transcripts whose promoters are hypothesized to be rich in binding site patterns that determine tissue-specific expression, and to *a posteriori* evaluate tissue-specific motifs, modules, and predictive models. Transcript selection requires mapping probes to transcripts, and probe intensities to the corresponding transcript intensity. We classify transcripts as either enhanced with tissue specificity (positive) or inhibited with tissue specificity (negative) based on the motif sites in their promoters.

To associate probes with RefSeq transcripts, we mapped the probes back to the genomes (NCBI human genome assembly Hs33 and mouse genome assembly ver. 3C dating to February 2003) to identify the probe locations and exon targets. We used the resulting probe-to-exon map to identify the RefSeq transcripts targeted by each probe and to assign a probe set to each transcript. If in a particular tissue the probe set AP calls disagreed, we removed the transcript from further consideration. To obtain intensities for a transcript we simply took the mean intensity of its assigned probes.

Table 8. common promoter composition in human tissue-specific sets

	Adip	Adre	Amyg	Bone	CD4T	CD8T	Cere	Dors	Hear	Hypo	Kidn	Live	Lung	Lymp	Olfa	Ovar	Panc	Pitu	Plac	Pros	Sali	Skel	Test	Thym	Thyr	Trac	Trig	Uter
Adip	–	0.112	0.034	0.022	0.010	0.006	0.032	0.034	0.054	0.042	0.078	0.064	0.096	0.048	0.084	0.092	0.044	0.048	0.102	0.056	0.034	0.028	0.012	0.046	0.076	0.072	0.020	0.160
Adre	0.150	–	0.034	0.074	0.012	0.010	0.080	0.074	0.088	0.032	0.164	0.084	0.082	0.062	0.102	0.114	0.084	0.098	0.072	0.080	0.068	0.050	0.040	0.050	0.072	0.106	0.050	0.078
Amyg	0.098	0.080	–	0.014	0.006	0.008	0.226	0.020	0.008	0.340	0.010	0.006	0.016	0.008	0.044	0.020	0.014	0.096	0.022	0.034	0.016	0.004	0.008	0.014	0.066	0.012	0.020	0.042
Bone	0.080	0.176	0.052	–	0.014	0.012	0.054	0.084	0.044	0.014	0.070	0.038	0.022	0.106	0.086	0.080	0.052	0.070	0.026	0.054	0.088	0.062	0.038	0.090	0.022	0.108	0.066	0.032
CD4T	0.084	0.062	0.072	0.064	–	0.704	0.014	0.010	0.012	0.016	0.004	0.008	0.022	0.180	0.010	0.010	0.018	0.014	0.008	0.022	0.028	0.008	0.010	0.130	0.030	0.008	0.008	0.030
CD8T	0.058	0.064	0.074	0.048	0.414	–	0.016	0.008	0.002	0.012	0.008	0.002	0.010	0.178	0.008	0.018	0.016	0.020	0.004	0.024	0.028	0.008	0.010	0.122	0.036	0.006	0.002	0.036
Cere	0.102	0.186	0.132	0.154	0.034	0.050	–	0.072	0.052	0.138	0.090	0.026	0.024	0.046	0.080	0.122	0.074	0.078	0.024	0.020	0.026	0.040	0.024	0.062	0.034	0.070	0.024	0.044
Dors	0.084	0.068	0.050	0.118	0.036	0.040	0.072	–	0.044	0.026	0.072	0.008	0.016	0.072	0.172	0.224	0.120	0.054	0.022	0.046	0.182	0.096	0.014	0.044	0.030	0.076	0.218	0.072
Hear	0.084	0.174	0.076	0.202	0.032	0.028	0.158	0.092	–	0.016	0.070	0.092	0.128	0.062	0.052	0.038	0.048	0.024	0.038	0.062	0.044	0.126	0.038	0.040	0.062	0.042	0.042	0.042
Hypo	0.128	0.112	0.162	0.048	0.084	0.088	0.112	0.054	0.052	–	0.020	0.016	0.014	0.018	0.096	0.008	0.008	0.126	0.032	0.042	0.020	0.022	0.020	0.008	0.034	0.018	0.024	0.040
Kidn	0.120	0.248	0.064	0.204	0.044	0.050	0.200	0.090	0.230	0.084	–	0.170	0.084	0.056	0.056	0.144	0.086	0.048	0.058	0.072	0.094	0.064	0.018	0.054	0.084	0.116	0.036	0.044
Live	0.084	0.176	0.068	0.216	0.042	0.048	0.122	0.104	0.360	0.038	0.230	–	0.088	0.028	0.032	0.024	0.064	0.032	0.036	0.070	0.014	0.014	0.032	0.032	0.064	0.026	0.022	0.022
Lung	0.084	0.200	0.126	0.226	0.044	0.060	0.158	0.068	0.244	0.076	0.226	0.244	–	0.088	0.094	0.044	0.068	0.032	0.168	0.202	0.030	0.016	0.028	0.090	0.134	0.100	0.020	0.114
Lymp	0.118	0.098	0.092	0.108	0.126	0.112	0.120	0.086	0.066	0.112	0.100	0.042	0.118	–	0.076	0.108	0.110	0.064	0.058	0.052	0.072	0.042	0.018	0.118	0.052	0.086	0.058	0.108
Olfa	0.106	0.092	0.118	0.076	0.080	0.072	0.104	0.118	0.068	0.130	0.088	0.060	0.090	0.092	–	0.094	0.062	0.102	0.060	0.094	0.092	0.084	0.016	0.048	0.078	0.092	0.140	0.152
Ovar	0.118	0.096	0.066	0.106	0.042	0.054	0.082	0.302	0.114	0.050	0.108	0.090	0.094	0.114	0.116	–	0.182	0.038	0.050	0.058	0.178	0.044	0.020	0.090	0.022	0.140	0.160	0.118
Panc	0.104	0.104	0.088	0.100	0.080	0.104	0.064	0.116	0.092	0.084	0.096	0.100	0.120	0.120	0.088	0.172	–	0.038	0.062	0.092	0.096	0.056	0.026	0.096	0.076	0.152	0.084	0.064
Pitu	0.084	0.100	0.120	0.078	0.114	0.108	0.096	0.088	0.046	0.118	0.084	0.060	0.078	0.110	0.104	0.068	0.116	–	0.048	0.078	0.070	0.056	0.038	0.044	0.080	0.102	0.042	0.066
Plac	0.124	0.180	0.120	0.146	0.090	0.078	0.144	0.046	0.132	0.094	0.184	0.136	0.240	0.142	0.086	0.090	0.122	0.120	–	0.094	0.044	0.016	0.026	0.050	0.082	0.102	0.026	0.102
Pros	0.102	0.110	0.208	0.092	0.124	0.112	0.078	0.072	0.084	0.130	0.086	0.118	0.170	0.100	0.110	0.064	0.130	0.094	0.150	–	0.092	0.032	0.024	0.084	0.198	0.156	0.020	0.170
Sali	0.108	0.088	0.068	0.078	0.082	0.074	0.080	0.214	0.068	0.058	0.098	0.056	0.056	0.118	0.128	0.198	0.166	0.086	0.074	0.060	–	0.118	0.020	0.026	0.084	0.188	0.156	0.070
Skel	0.104	0.096	0.050	0.108	0.036	0.038	0.064	0.192	0.190	0.046	0.128	0.160	0.074	0.072	0.086	0.178	0.120	0.048	0.058	0.046	0.174	–	0.012	0.012	0.040	0.048	0.152	0.050
Test	0.100	0.166	0.100	0.146	0.068	0.058	0.128	0.072	0.168	0.090	0.166	0.160	0.194	0.130	0.094	0.082	0.094	0.102	0.154	0.132	0.088	0.084	–	0.032	0.018	0.028	0.020	0.016
Thym	0.118	0.152	0.120	0.122	0.096	0.098	0.130	0.048	0.130	0.106	0.144	0.098	0.236	0.154	0.114	0.072	0.120	0.100	0.234	0.174	0.060	0.058	0.170	–	0.040	0.090	0.030	0.040
Thyr	0.094	0.074	0.294	0.070	0.148	0.134	0.074	0.044	0.072	0.174	0.068	0.092	0.136	0.116	0.084	0.056	0.084	0.140	0.114	0.308	0.052	0.034	0.118	0.144	–	0.132	0.016	0.128
Trac	0.140	0.134	0.114	0.140	0.056	0.070	0.136	0.138	0.084	0.086	0.140	0.086	0.118	0.138	0.118	0.180	0.180	0.138	0.168	0.102	0.196	0.088	0.140	0.134	0.084	–	0.066	0.102
Trig	0.068	0.060	0.032	0.090	0.044	0.046	0.062	0.314	0.122	0.040	0.064	0.102	0.060	0.070	0.102	0.252	0.144	0.040	0.038	0.038	0.224	0.246	0.068	0.026	0.026	0.106	–	0.062
Uter	0.092	0.066	0.270	0.058	0.142	0.134	0.100	0.082	0.068	0.116	0.056	0.048	0.148	0.156	0.096	0.112	0.128	0.082	0.122	0.214	0.124	0.062	0.086	0.156	0.252	0.114	0.074	–

FG sets are given in the top portion and human BG sets in the bottom.

Table 9. Common promoter composition in mouse tissue-specific sets

	Adip	Adre	Amyg	Bone	CD4T	CD8T	Cere	Dors	Hear	Hypo	Kidn	Live	Lung	Lymp	Olfa	Ovar	Panc	Pitu	Plac	Pros	Sali	Skel	Test	Thym	Thyr	Trac	Trig	Uter
Adip	–	0.206	0.016	0.024	0.064	0.060	0.018	0.044	0.074	0.018	0.096	0.128	0.142	0.148	0.024	0.246	0.002	0.056	0.068	0.056	0.026	0.060	0.004	0.064	0.010	0.206	0.046	0.138
Adre	0.230	–	0.028	0.010	0.032	0.030	0.046	0.066	0.074	0.040	0.118	0.064	0.130	0.036	0.046	0.234	0.012	0.078	0.060	0.088	0.020	0.062	0.002	0.036	0.012	0.112	0.062	0.106
Amyg	0.052	0.068	–	0.010	0.004	0.004	0.256	0.178	0.022	0.352	0.008	0.002	0.024	0.006	0.378	0.030	0.014	0.060	0.006	0.026	0.014	0.018	0.000	0.022	0.006	0.010	0.146	0.028
Bone	0.082	0.050	0.052	–	0.024	0.038	0.018	0.012	0.054	0.010	0.028	0.034	0.030	0.094	0.024	0.008	0.042	0.034	0.040	0.026	0.070	0.042	0.006	0.074	0.088	0.036	0.006	0.034
CD4T	0.156	0.112	0.050	0.112	–	0.656	0.008	0.020	0.010	0.004	0.014	0.002	0.040	0.324	0.006	0.038	0.006	0.034	0.022	0.024	0.016	0.012	0.002	0.232	0.016	0.102	0.016	0.048
CD8T	0.094	0.078	0.050	0.120	0.472	–	0.010	0.038	0.014	0.012	0.010	0.008	0.048	0.318	0.014	0.024	0.018	0.038	0.026	0.038	0.020	0.014	0.004	0.240	0.030	0.096	0.016	0.034
Cere	0.104	0.124	0.192	0.076	0.054	0.044	–	0.186	0.014	0.244	0.020	0.006	0.024	0.010	0.310	0.026	0.022	0.074	0.020	0.048	0.016	0.040	0.002	0.022	0.034	0.016	0.180	0.022
Dors	0.114	0.090	0.148	0.056	0.116	0.098	0.122	–	0.026	0.158	0.024	0.014	0.050	0.018	0.126	0.034	0.030	0.064	0.030	0.042	0.010	0.044	0.002	0.016	0.012	0.012	0.574	0.026
Hear	0.094	0.108	0.056	0.138	0.072	0.048	0.100	0.058	–	0.022	0.100	0.026	0.050	0.020	0.030	0.034	0.028	0.038	0.022	0.032	0.038	0.346	0.004	0.018	0.056	0.036	0.030	0.024
Hypo	0.084	0.092	0.272	0.056	0.060	0.034	0.214	0.134	0.066	–	0.022	0.014	0.024	0.012	0.278	0.028	0.008	0.122	0.024	0.026	0.014	0.020	0.002	0.018	0.010	0.010	0.192	0.018
Kidn	0.130	0.134	0.058	0.138	0.082	0.058	0.090	0.090	0.188	0.052	–	0.242	0.056	0.010	0.016	0.050	0.008	0.042	0.038	0.060	0.032	0.052	0.004	0.010	0.026	0.044	0.024	0.034
Live	0.082	0.086	0.056	0.108	0.088	0.086	0.056	0.090	0.150	0.062	0.230	–	0.044	0.016	0.010	0.034	0.028	0.018	0.038	0.040	0.030	0.020	0.000	0.004	0.014	0.044	0.012	0.032
Lung	0.184	0.180	0.064	0.100	0.094	0.082	0.076	0.118	0.098	0.090	0.126	0.084	–	0.082	0.054	0.142	0.006	0.036	0.116	0.068	0.024	0.018	0.000	0.052	0.022	0.190	0.050	0.120
Lymp	0.096	0.102	0.038	0.150	0.194	0.230	0.078	0.054	0.114	0.056	0.070	0.084	0.118	–	0.018	0.038	0.028	0.022	0.050	0.046	0.022	0.010	0.000	0.132	0.044	0.178	0.018	0.070
Olfa	0.064	0.076	0.302	0.046	0.040	0.052	0.244	0.134	0.088	0.242	0.056	0.042	0.070	0.076	–	0.038	0.012	0.068	0.030	0.020	0.016	0.024	0.002	0.020	0.032	0.006	0.104	0.020
Ovar	0.244	0.234	0.058	0.062	0.136	0.070	0.138	0.136	0.084	0.102	0.100	0.058	0.178	0.092	0.070	–	0.006	0.092	0.102	0.100	0.016	0.038	0.006	0.064	0.014	0.106	0.048	0.336
Panc	0.044	0.042	0.064	0.176	0.042	0.062	0.048	0.062	0.098	0.082	0.086	0.182	0.054	0.106	0.060	0.026	–	0.052	0.056	0.022	0.112	0.030	0.004	0.010	0.084	0.010	0.016	0.016
Pitu	0.122	0.118	0.092	0.062	0.096	0.094	0.098	0.152	0.074	0.090	0.084	0.078	0.142	0.070	0.120	0.124	0.082	–	0.044	0.120	0.068	0.030	0.008	0.030	0.050	0.042	0.102	0.074
Plac	0.054	0.080	0.070	0.104	0.064	0.078	0.060	0.092	0.068	0.042	0.118	0.132	0.134	0.126	0.072	0.058	0.142	0.092	–	0.046	0.038	0.034	0.004	0.020	0.050	0.028	0.028	0.086
Pros	0.154	0.130	0.060	0.086	0.112	0.068	0.082	0.096	0.098	0.116	0.116	0.134	0.150	0.088	0.060	0.132	0.078	0.116	0.078	–	0.132	0.038	0.004	0.044	0.062	0.068	0.030	0.098
Sali	0.088	0.072	0.044	0.116	0.064	0.078	0.062	0.070	0.112	0.052	0.114	0.176	0.088	0.096	0.058	0.064	0.206	0.100	0.148	0.144	–	0.024	0.002	0.018	0.170	0.016	0.020	0.028
Skel	0.096	0.078	0.054	0.122	0.072	0.092	0.074	0.080	0.226	0.052	0.154	0.166	0.070	0.070	0.060	0.078	0.168	0.068	0.126	0.100	0.134	–	0.010	0.022	0.028	0.034	0.044	0.022
Test	0.098	0.088	0.062	0.114	0.092	0.068	0.096	0.088	0.118	0.102	0.098	0.134	0.068	0.090	0.076	0.104	0.190	0.088	0.130	0.066	0.120	0.138	–	0.004	0.002	0.006	0.000	0.004
Thym	0.158	0.172	0.066	0.112	0.228	0.238	0.106	0.104	0.082	0.070	0.082	0.066	0.180	0.154	0.068	0.190	0.030	0.092	0.050	0.148	0.066	0.086	0.080	–	0.022	0.124	0.020	0.040
Thyr	0.074	0.048	0.054	0.142	0.080	0.088	0.066	0.088	0.120	0.052	0.094	0.162	0.078	0.078	0.066	0.060	0.260	0.112	0.152	0.108	0.230	0.154	0.150	0.078	–	0.018	0.026	0.022
Trac	0.222	0.188	0.058	0.100	0.116	0.074	0.088	0.106	0.104	0.072	0.094	0.066	0.200	0.082	0.046	0.226	0.058	0.120	0.064	0.136	0.078	0.104	0.098	0.200	0.086	–	0.026	0.110
Trig	0.084	0.086	0.206	0.038	0.052	0.066	0.136	0.226	0.106	0.176	0.112	0.098	0.098	0.040	0.162	0.072	0.068	0.118	0.068	0.098	0.072	0.082	0.072	0.058	0.046	0.090	–	0.036
Uter	0.168	0.176	0.056	0.076	0.140	0.074	0.126	0.100	0.122	0.124	0.092	0.048	0.152	0.136	0.072	0.258	0.042	0.122	0.066	0.158	0.096	0.086	0.102	0.168	0.066	0.162	0.070	–

FG sets are given in the top portion and human BG sets in the bottom.

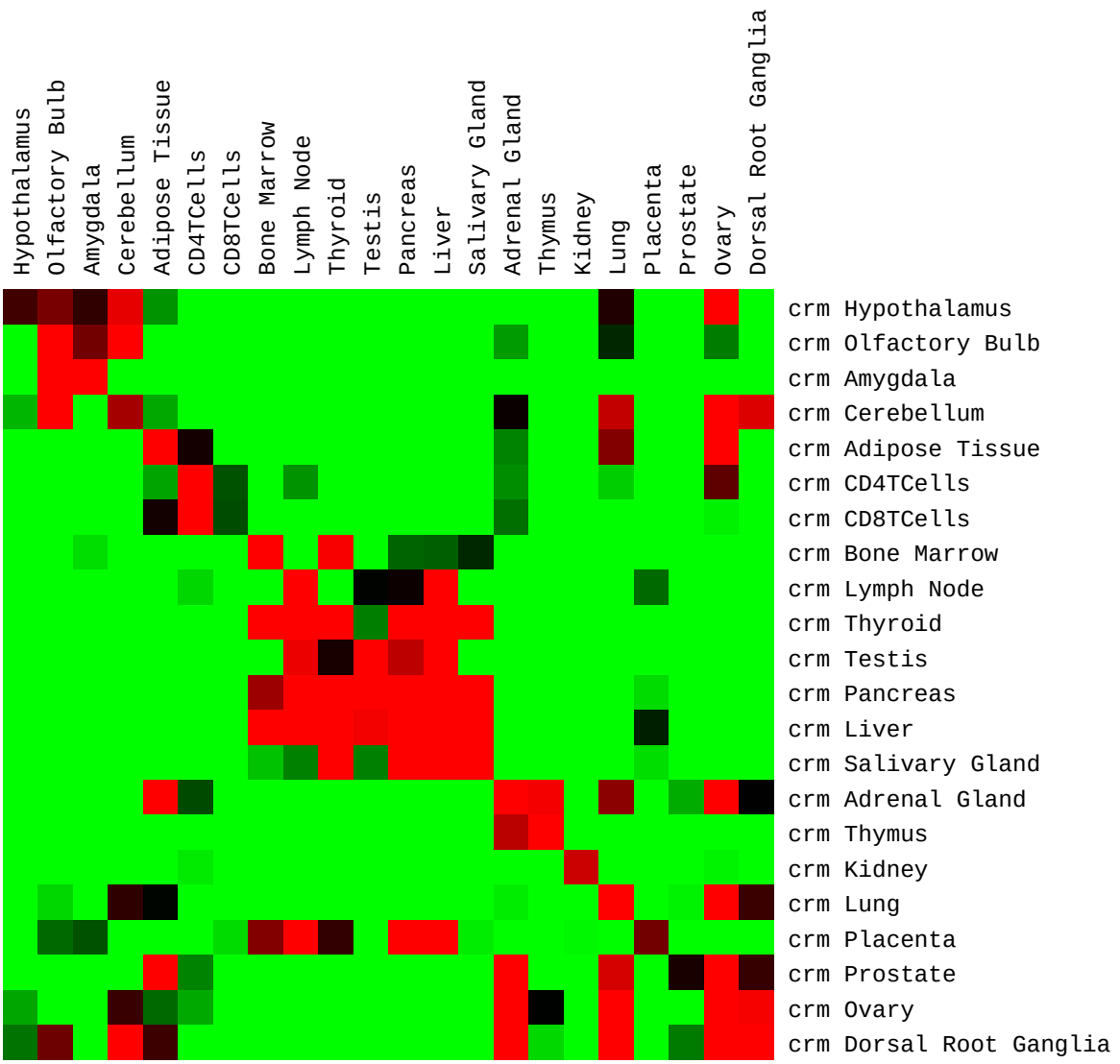


Fig. 6. Prediction error of *cis*-regulatory module (CRM)-based predictors (right) trained on specific mouse tissues and tested on all mouse tissues (top). Errors below, at, and above 45% are displayed in red, black, and green. The diagonal, corresponding to predictors trained and tested on the same tissue, gives prediction error under 10-fold crossvalidation. When applying a predictor to a tissue other than that on which it was trained, promoters common to both tissues are excluded. Tissues and corresponding predictors with crossvalidation error P value above the 0.01 significance cutoff are omitted.

Table 10. Summary of probe data

	Human	Mouse
Probes on the chip	33,674	36,182
Probes called present at least once	20,041	22,728
Probes associated with a RefSeq transcript	15,967	10,950
Probes with more than one RefSeq transcript	2,035	308
Mean intensity for probe present calls	2,886.4	750.4

Number of probes on the chips includes Affymetrix probes and GNF probes, but not the Celera probes.

Table 11. Summary of transcript data

	Human	Mouse
UniGene sets represented on chip	33,954	31,479
RefSeq transcripts represented on chip	12,651	9,851
RefSeq transcripts called present at least once	8,450	7,398
RefSeq transcript ortholog pairs	22,209	22,209
RefSeq transcripts with at least one ortholog	11,793	10,189
RefSeq transcripts with unique first exon	8,157	9,884
RefSeq transcripts with more than one probe	3,757	1,322
Mean number of probes per RefSeq transcript	1.3	1.03
Most probes associated with a transcript	25	23
Percent of calls for which probes disagree (*i.e.* discarded calls)	7.8%	5.1%
Mean intensity for RefSeq transcript present calls	2,970.1	973.1
Proportion of RefSeq transcript calls discarded	0.264	0.383

2.1.2 Mapping transcripts to promoters

Regulatory elements can exist almost anywhere in the genome, but they are known to have a high concentration in proximal promoters. Smith *et al.* (2) and Sumazin *et al.* (9) were successful in identifying experimentally validated motifs for factors known to play tissue-specific regulatory roles. Promoter quality (*i.e.*, confidence in the TSS) has a large impact, and poor-quality promoters may hurt motif discovery as much as poor-quality sets of tissue-specific transcripts (for example, those containing ubiquitous or incorrectly assigned transcripts).

To map transcripts to promoters, we used CSHLmpd, which includes annotations for human, mouse, and rat (8). We exclude isoforms that share first exons and have inconsistent calls in any of the tissues. We prohibit multiple representations of promoters in our positive and negative sets (see below for definition). CSHLmpd includes 51,506 and 46,475 promoter annotations for human and mouse, of which 16,433 and 15,061 are assigned to RefSeq transcripts. We extracted promoter sequences for each transcript with a promoter in CSHLmpd, using a sequence from -1,000 to +100 relative to the TSS. We note that mouse TSS prediction is thought to be substantially less accurate than human TSS prediction at present.

2.1.3 Transcripts differentially expressed in a tissue

Using the intensities we assigned to transcripts and the information in CSHLmpd, we constructed sets of promoters for genes with tissue-specific differential expression. The procedure we used is as follows:

Table 12. Summary of tissue data

	Human	Mouse
Total number of tissues	79	61
Number of tissues with experiments in human and mouse		28
Number of tissue-differential transcripts	1,243/2,517	1,113/2,175
Mean intensity for a present call	3,550.5/3,560	2,317.2/2,123.2
Number of tissue-specific orthologs	756/1162	744/1,123
Mean intensity of tissue-specific orthologs	4,653.9/4,282.7	2,561.3/2,207

1. For each transcript and each tissue, we calculate the number of standard deviations by which the intensity of that transcript in the tissue differed from the mean intensity for that transcript across all tissues. We call this the standard intensity.

2. If the standard intensity is positive, the transcript is said to be enhanced with tissue specificity, and if the standard intensity is negative, the transcript is said to be inhibited with tissue specificity.

3. The sets of enhanced and inhibited transcripts are ranked according to their standard intensity, from highest to lowest in each tissue.

4. Transcripts with no associated promoter in CSHLmpd are removed from consideration.

5. Of the remaining overrepresented transcripts, the top-ranking 500 (according to standard intensity) are selected in each tissue, and their promoters form the positive set (foreground) for this tissue. The 500 promoters corresponding to remaining transcripts with lowest standard intensity form the negative set (background).

The resulting positive and negative sets are called the tissue-differential sets. These uniform-size sets include promoters of transcripts with intensities near the mean in some tissues and exclude promoters of transcripts with intensities far from the mean in other tissues. However, experimentation suggests that this size (500) is a good compromise between tissue specificity and statistical power. The benefits of using large and uniform size sets include consistent prediction estimates that are comparable across tissues and robustness to outliers and features that are shared by few promoters. The drawbacks include higher estimates of prediction-error and elevated noise level.

2.2 Obtaining motifs and modules

Motifs are abstract characterizations of the DNA sequence elements to which transcription factors bind. Modules are sets of motifs whose corresponding sites are thought to interact synergistically.

2.2.1 PWMs

We represent motifs with PWMs; PWMs are described in detail by Stormo (10). A rigorous statistical model based on PWMs was developed by Liu *et al.* (11). The statistical interpretation of PWMs allows them to be converted into scoring matrices. Because PWMs describe distributions over the substrings of a sequence, we can estimate the likelihood that the substring was generated from the distribution described by a PWM. The score given by the scoring matrix for a substring is the log of the likelihood that the substring was generated by the distribution described by the PWM, divided by the likelihood that the substring was generated by the

base composition (which can be the genomic base composition, or the base composition of some other set of sequences). Negative scores indicate that the substring was more likely to have been generated by the base composition than the PWM. Positive scores indicate a greater likelihood for generation by the PWM.

When using experimentally verified or *in silico*-identified motifs, we rank each motif as a predictor and eliminate motifs that are similar to higher-ranked motifs. We also match *in silico*-identified motifs to transcription factors based on similarity of identified motifs and experimentally verified binding-site motifs for these factors. We compare motif pairs using matcompare (12), which aligns matrices and compares the aligned column; matcompare implements various measures of distance or similarity between columns. We use the Kullback-Leibler divergence method to compare columns (this is a general distance between divergence). In comparing matrices, we require that the set of aligned columns be contiguous, and for the smaller of the two matrices being compared, we require that, at most, a specified number of positions not be aligned (we call this the number of overhang columns). To indicate when two matrices are to be considered as representing the same motif, we require that the divergence per aligned column be at most 1.0, and we restrict the number of overhang columns to 1. These values are suitable for matrices of widths roughly in the range 8-12. In our experience, when aligning longer matrices, it is best to allow a greater overhang value, and when aligning shorter matrices, it is best to reduce the threshold for divergence and not allow any overhang.

2.2.2 Motifs from the TRANSFAC database

TRANSFAC (1) is the largest database of experimentally validated transcription factor-binding sites and corresponding binding-site models. We used motifs from the vertebrate subset of TRANSFAC Professional version 8.4 PWMs. This subset includes 546 vertebrate matrices of varying qualities, of which fewer than 120 are distinct according to matcompare.

2.2.3 Motif discovery using DME and DME-B

We used DME (2) and a variant DME-B for *de novo* motif discovery. DME enumerates PWM-based motifs composed of columns from discrete sets of column types. The DME algorithm defines motif occurrences, and these are used to measure the quality of a motif. The score for a motif in DME is the score of all of the motif's foreground occurrences, minus the score of all the motif's background occurrences. DME is described in detail by Smith *et al.* (2).

DME-B uses the same enumerative strategy, but the score for a motif in DME-B is the number of foreground sequences with a score greater than 0 minus the number of background sequences with a score greater than 0. DME-B keeps track of the sequences that contain occurrences and is more expensive to compute in terms of time and space.

For both DME and DME-B, the enumerative strategy uses several parameters including motif width (w), granularity (g), refinement limit (r), and information content measured in bits/column (b). See Smith *et. al* (2) for detailed definitions. We use the tuple (w, g, r, b) to describe sets of parameters used by DME and DME-B. These parameters indicate the degeneracy of motifs and are related to the expected number of occurrences of the motifs in random sequences.

For each tissue, DME was used to obtain three sets of motifs, each set containing 30 motifs. The parameter combinations used were $(12, 0.5, 0.25, 1.45)$, $(10, 0.5, 0.125, 1.6)$, and $(8, DN, 0.125, 1.8)$. DME-B was used to obtain two sets of motifs, each set containing 30 motifs, with parameter combinations $(10, 1, 0.25, 1.6)$ and $(8, 1, 0.125, 1.8)$. All of the motifs were pooled to obtain a single set for each tissue.

2.2.4 Selecting motifs from a set of candidates

The initial set of candidate motifs may include experimentally validated motifs, computationally identified motifs, or both. We used motifclass (this program is available from CREAD) to assign scores to motifs based on their ability to discriminate between positive and negative promoters sets. For a given tuple consisting of a promoter and a motif of length w, motifclass produces a score that is equal to the score of the highest-scoring subsequence of length w in the promoter and its reverse complement. Subsequences are scored using the log-likelihood score, as described in Section 2.2.1. Having assigned a score to each promoter in the positive and negative sets using the motif, motifclass evaluates the motif's ability to discriminate between positive and negative promoters based on error rate. For a given score threshold, motifclass predicts that promoters with scores above the threshold belong to the positive set, and scores below belong to the negative set. For each motif, motifclass selects a score threshold that minimizes the error corresponding to this prediction, where the error is simply the rate of incorrect predictions.

After scoring each motif using motifclass, we used the uniqmotifs program to eliminate motifs from the set if they are sufficiently similar (based on Kullback-Leibler divergence) to a higher-scoring motif. Finally, we retained the top 100 remaining motifs.

2.2.5 Building modules from motifs

We constructed modules consisting of pairs and triples of motifs from the set of retained individual motifs. Pairs were assembled using the modclass program, which combines motifs into modules and evaluates them in a manner similar to motifclass, using max-scoring subsequences and thresholds. For a module to classify a sequence as positive, the sequence must have substrings with scores above threshold for each motif in the module. To construct modules, modclass simultaneously finds optimal threshold values for each motif in a candidate module, which is time consuming and done using a branch and bound algorithm.

Because module enumeration and construction are computationally expensive, we built modules of size 2 using exhaustive enumeration, and modules of size 3 were generated from modules of size 2. When three pairs (size-2 modules) are composed of only three distinct motifs, those three motifs are combined to form a module of size 3. All such triples were retained and along with the top-scoring 100 pairs produced by modclass, form the set of retained modules.

2.3 Building predictive models

Using the retained motifs and modules, we built models to predict whether the transcripts associated with promoters would be overexpressed or underexpressed in a specific tissue. Because we wanted to determine whether proximal promoters contained sufficient information to make statistically significant predictions, we evaluated the accuracy of predictions made by these models. Our method consisted of first constructing sets of features using motifs and modules, then using MARS (a machine-learning algorithm) to build a model from those features; the quality of the models was determined by crossvalidation.

2.3.1 Pattern-based features

First, we used the sets of retained motifs and modules to construct features from which to build the models. Features (in this context) are functions describing properties of sequences in terms of motifs or modules. We initially tested many types of features to use for classification. These include functions based on counts and strengths of motif and module occurrences in sequences, as well as relative positions of occurrences

for motifs in a module and positions of motif occurrences relative to the transcription start site. Although positions of occurrences are known to be important for binding sites of certain factors, the features we tested that were based on position performed poorly relative to substring-score features.

The feature types selected to be used for classification included the max-score feature (score of highest scoring substring, see Section 2.2.4) for individual motifs; for modules, we used the sums (max-score-sum) and products (max-score-product) of the max-score feature for motifs comprising the module. Each type of feature was evaluated for each motif or module, as appropriate, on each positive and negative promoter. The final sets of features for the tissue-differential sets corresponded to the max-score features and the max-score-sum and -product features (one for each retained module).

2.3.2 Classification using MARS

We used the MARS algorithm (13) to build predictive models (classifiers) from the sets of features. MARS is a nonparametric and adaptive regression method that builds a set of models using stepwise forward selection and backward elimination and in terms of basis functions and their products. Each basis function has the general form

$$\max(0, x - k) \text{ or } \max(0, k - x),$$

where x is an input variable (a feature), and k is a constant called the knot value. Each term is selected to minimize reduction in variance. Let y_i be the response variable for the i_{th} observation, and let $\bar{y}_i$ be the corresponding predicted outcome, then reduction in variance is defined as

$$\text{RIV} = 1 - (\textstyle\sum_{i=1}^{m}(\Gamma_i - \bar{\Gamma})^2)/(\sum_{i=1}^{m}(y_i - \bar{y})^2),$$

where $\Gamma_i = y_i - \hat{y}_i$, and $\bar{y}$ and $\bar{\Gamma}$ are the corresponding means.

We build a model up to a maximum number of terms, and then remove terms iteratively to generate a set of models of different sizes. At each stage, one term is removed so that the performance of the model consisting of the remaining terms is maximized. MARS produces either (i) all resulting models, or (ii) the model with size maximizing the error under crossvalidation (14). We adapted MARS to function as a classifier by having it regress against response variables restricted to 1 for a positive promoter (observation) and -1 for a negative observation.

The models constructed by MARS can reveal complex interactions between features. The sign of each term indicates whether that term will contribute to assigning observations to the positive or the negative set. The knots appearing in the basis functions act as a cutoff value and impose criteria that eliminate the influence of any term containing a basis function not meeting a particular value. As classifiers, these models behave like Boolean formulas with weighted terms.

2.3.3 Evaluating predictive ability

We use crossvalidation to evaluate how well predictive models predict differential expression in a tissue. To correctly perform crossvalidation, each testing set must be excluded from motif discovery, optimization, module construction, and predictor construction to ensure accurate estimation of the prediction error.

We used 10-fold crossvalidation, which randomly partitions the data into 10 equal size parts. For part k, a model is trained on the combined other 9 parts and then tested on part k. In this way, each observation is involved in exactly one testing set. The testing results for each observation are used to calculate the sensitivity, specificity, and error rate of the predictions. These statistics correspond to 10 distinct trained

models and provide an indication of how well a model trained on the entire data set would do when making predictions about data not yet observed.

Under the null hypothesis that no special signals are common to a given set of tissue-differential promoters, we expect our predictive models to have a predictive error of 0.5. Therefore, the probability of observing a predictive error lower than α for a particular tissue is distributed according to a binomial distribution, and we can obtain a P value to use in determining statistical significance.

As described in Section 2.3.2, our implementation of MARS produces 6 different models for each training set. As the appropriate model size for each tissue, we select the size with the smallest predictive error. Because this represents selecting the optimal from among a set of size 6, we use a Bonferroni correction for multiple testing, which amounts to multiplying the resulting P value by 6.

To demonstrate the effect of motif identification and motif and module optimization before test-set exclusion, we generated 100 pairs of positive and negative promoter set for both human and mouse by randomly selecting promoters from CSHLmpd. We identified and optimized motifs in the entire set, used MARS to generate a predictive model on each training set and estimated the prediction error on the test set. Average prediction errors for human and mouse were 0.389 ± 0.013 and 0.378 ± 0.012, which after multiple sampling correction, are statistically significant. When motif identification and optimization were done after test-set exclusion (the correct way), the average prediction error for human and mouse was 0.491 ± 0.018 and 0.491 ± 0.017, which is not statistically significant.

References

[1] Matys, V., Fricke, E., Geffers, R., Gossling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A. E., Kel-Margoulis, O. V. *et al.* (2003) *Nucleic Acids Res.* **31**, 374–378.

[2] Smith, A. D., Sumazin, P., & Zhang, M. Q. (2005) *Proc. Natl. Acad. Sci. USA* **102**, 1560–1565.

[3] Su, A. I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K. A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G. *et al.* (2004) *Proc. Natl. Acad. Sci. USA* **101**, 6062–6067.

[4] Fuchs, P., Zorer, M., Rezniczek, G., Spazierer, D., Oehler, S., Castañón, M., Hauptmann, R., & Wiche, G. (1999) *Hum. Mol. Genet.* **8**, 2461–2472.

[5] Johnson, J. M., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P. M., Armour, C. D., Santos, R., Schadt, E. E., Stoughton, R., & Shoemaker, D. D. (2003) *Science* **302**, 2141–2144.

[6] Zhang, T., Haws, P., & Wu, Q. (2004) *Genome Res.* **14**, 79–89.

[7] Pruitt, K. D., Tatusova, T., & Maglott, D. R. (2005) *Nucl. Acids Res.* **33**, D501–504.

[8] Xuan, Z., Zhao, F., Wang, J. H., Chen, G. X., & Zhang, M. Q. (2005) *Genome Biol.* **6**, R72.

[9] Sumazin, P., Chen, G., Hata, N., Smith, A. D., Zhang, T., & Zhang, M. Q. (2005) *Bioinformatics* **21**, 31–38.

[10] Stormo, G. D. (2000) *Bioinformatics* **16**, 16–23.

[11] Liu, J. S., Lawrence, C. E., & Neuwald, A. (1995) *J. Am. Stat. Assoc.* **90**, 1156–1170.

[12] Schones, D., Sumazin, P., & Zhang, M. Q. (2005) *Bioinformatics* **21**, 307–313.

[13] Hastie, T., Tibshirani, R., & Friedman, J. H. (2001) *The Elements of Statistical Learning.* (Springer, New York).

[14] Das, D., Banerjee, N., & Zhang, M. Q. (2004) *Proc. Natl. Acad. Sci. USA* **101**, 16234 –9.