Tissue-specific regulatory elements in mammalian promoters:

Supplementary information

To identify mammalian tissue-specific elements we (1) identify transcripts with tissue-specific expression, (2) map transcripts to proximal promoters, and (3) analyze promoters to identify common binding site motifs and cis-regulatory modules. Our approach builds on previous work to identify motifs in promoters of co-expressed transcripts. Co-regulated transcripts (*e.g.* tissue-specific) are likely to be controlled by similar machinery. We identify common motifs and modules that distinguish promoters of tissue specific transcripts from other promoters.

1 Transcripts and promoters under tissue-specific regulation

We considered transcripts tissue-specific if they play a special role in a specific tissue or small set of tissues. Tissue-specificity is a function of regulation, and a particular transcript may be expressed or even specific to more than one tissue. We assumed that tissue-specific function is associated with tissue-specific regulation.

To build sets of tissue-specific transcripts, we combined information from GNF SymAtlas (Su et al., 2004) (for both human and mouse), the Hughes Toronto microarrays (Zhang et al., 2004b) (for mouse), EST data from dbEST (Boguski et al., 1993) and membership in specific GO categories (Ashburner et al., 2000). We used key-word searches in the NCBI Nucleotide database and the Eukaryotic Promoter Database (EPD) (Perier et al., 1998) for guidance only.

We used multiple data sources to circumvent problems associated with each individual source, and add robustness to tissue-specific transcript selection. For example, expression of certain transcripts in dbEST may only have been measured for a small number of tissues, and are not a robust indicator of specificity. The GNF microarray data includes two replicates. Ignoring highly variable observations within and between replicates leads to incomplete data; considering such observations is equally problematic. Multiple sources of information were combined using a voting system, and each source contributed at most one vote. We showed that orthologs of transcripts with multiple votes for tissue-specific regulation are significantly more likely to have evidence for tissue-specific regulation, even when Gene Ontology votes were discarded. Because ortholog information did not have a vote for tissue specificity, significantly more frequent verification by comparative genomics suggests that selection according to multiple evidence is more reliable.

The construction of tissue specific sets is a complex and error prone task. Voting systems need to minimize dependence between annotation sources, are sure to pass over tissue specific transcripts because of inconclusive evidence, and should treat paralog transcripts with special care. Because of varying data type and experiment quality, we varied thresholds used for assigning votes to evidence. As additional sources of expression information become available and experiment quality become more uniform, thresholds used to assign votes will become standard. One example for possibly erroneous tissue-specific annotation is evident in the mouse pancreas set, where 11 genes from the Klk family received a vote and were included in the tissue specific set. The votes originate from membership is a pancreas-related GO function, which was annotated for the human ortholog (the mouse genes share the same ortholog). We had no expression-based evidence for pancreas-specific regulation in our mouse data. The genes are highly similar and may be the product of a recent duplication. None of our mouse-pancreas-specific motifs or modules were identified in the Klk promoters, and their inclusion increased our error estimates.

1.1 microarray data

Because most of the probes called present in the experiments are associated with RefSeq transcripts, and because in general our characterization of promoters and first introns is better for RefSeqs than other annotations (almost 100% of experimentally verfied promoters have RefSeq annotations), we decided to focus exclusively on RefSeq transcripts (see Table 11). Moreover, there is evidence that expression of different transcripts in different tissues is widespread (Fuchs et al., 1999; Johnson et al., 2003; Zhang et al., 2004a), and that different transcripts can have different first exons, and hence different promoters (Yamashita et al., 2006).

To associate probes with RefSeq transcripts, we mapped the probes back to the genomes (NCBI human genome assembly Hs33 and mouse genome assembly v3C dating to February 2003 for GNF and Mm5 for the Hughes Toronto array) to identify the probe locations and exon targets. GNF and GeneNote probes (Affymetrix gene chips) are 25 bases long and probe-transcript mappings were required to be perfect matches. The Hughes Toronto array probes are 40-bases long and mappings were required to have 39 base matches. Probes that matched more than 3 DNA targets were discarded. We chose this simple criteria as a compromise between data loss and quality. Using multiple probes per transcript and repeated experiments on the same tissue samples allows for correcting imprecise expression intensity measurements for a particular transcript target. We used the resulting probe-to-exon map to identify the RefSeq transcripts targeted by each probe, and assign a probe set to each transcript.

Probes with detection ability that is not substantially different than the corresponding mismatch probe are suspect. To obtain A/P calls for the transcripts, we used the A/P calls of the corresponding probe sets. If in a particular tissue the probe set A/P calls disagreed, we removed the transcript from further consideration, otherwise we called the transcript *present* or *absent*. The number of excluded transcripts due to this pruning operation ranged from under 5% to 80% of potential tissue-specific transcripts. 285 transcripts had unusually high intensity level (above 3 standard deviations from mean) in the GNF data for human skeletal muscle. 227 (80%) of these were rejected because of A/P calls. 58 of the 227 were reinstated because of evidence from other experiments. Obtaining intensities and p-values for a transcript is more straight-forward: we took the mean intensity and geometric-mean p-value of its corresponding probes.

To identify transcripts with unusually high intensity in few tissues, we calculated the mean and standard deviation of of the normalized intensity across all experiments (for the GNF data mouse and human are evaluated separately). For the GNF and GeneNote data, we considered transcripts that are called present and have intensity greater than three standard deviations from the mean intensity to be specific. For the Hughes Toronto array, where variability between probe reads is higher, we called transcript specific if their normalized intensity was greater than 2 and greater than ten standard deviations from the mean, and they were called present. The Hughes Toronto array did not include experiments for mouse CD4 T-cells, and GeneNote did not include experiments for human CD4 T-cells and testis.

1.2 EST Data

dbEST is a depository for expressed sequence tags, maintained by NCBI (Boguski et al., 1993). When ESTs are derived from specific tissues, this information is specified in dbEST. Because of the nature of tissue sample collection, source variability, and because most ESTs are derived from tissue cocktails and not all experiments are focused on keeping tissue purity, dbEST tissue classification may be imprecise. However, when multiple EST libraries describe a particular gene as expressed in a particular tissue and this gene was detected in few other tissues, we considered the dbEST evidence strong enough to cast votes for tissue specificity. We considered evidence for tissue specificity of a gene as evidence for all alternative

transcripts of this gene. For a dbEST vote in a given tissue, at least 3 EST libraries had to describe expression observation in this tissue, and more than 50% of tissue-related observations had to be in this tissue.

1.3 GO Terms

We associated a set of GO Terms with each tissue. This was done by compiling a set of keywords for each tissue (*e.g.* "renal" was associated with Kidney), and searching GO Term names and definitions for those keywords. For each tissue, we produced a set of GO Terms that were subsequently reviewed to ensure that the context of the keywords was appropriate. Each gene annotated with at least one GO Term associated with a tissue received one positive vote for specificity in that tissue. Similarly to EST evidence, we considered each point of evidence for a gene as evidence for all alternative transcripts of this gene.

1.4 Identifying factors expressed in tissues

To assist interpretation of results we assembled factors with evidence of expression in each tissue according SymAtlas, Hughes Lab, GeneNote and dbEST data. These lists are available in TCat. A transcript was considered expressed in a microarray experiment if it received a present call, regardless of intensity values. For dbEST, a transcript was considered expressed if any associated ESTs were observed at least once in that tissue.

1.5 Transcripts with strong evidence for tissue-specific regulation

In all tissues except CD4 T-cells, orthologs of transcripts with multiple votes for tissue-specific regulation were more likely to have evidence for specific regulation in that tissue. This suggests that the false-positive rate for calling a transcript tissue-specific is much lower when based on multiple votes. To compare the predictive power of orthologs, we assembled all pairs of transcripts and their orthologs, where at least one member had evidence for tissue-specific regulation. Every transcript appeared in only one pair, and transcripts with no known ortholog were eliminated. The pairs can be partitioned into 2 sets, those pairs that include a transcript with multiple votes and those who do not. We then used a hypergeometric fixed marginal contingency table test (Fisher exact) to compare the proportion of pairs (in each of the two sets) where each of the transcripts had evidence for tissue-specific regulation. The fixed marginal contingency table p-value follows the hypergeometric distribution (Agresti, 1992). The two-sided p-value for the table is the sum of the probabilities of all tables that are at least as extreme.

In the main body of the paper we gave the list of genes and orthologous transcripts with multiple votes for skeletal muscle-specific regulation in both human and mouse. Here, in Tables 5 to 10 we give the corresponding lists for the other 6 tissues. Note that each promoter set included 100 promoters, including some promoters that correspond to transcripts with only one vote for tissue-specific regulation. The distribution of votes per transcript in the set corresponding to analyzed promoter sets is given in Figure 3.

1.6 Obtaining promoter sequences

Regulatory elements can exist almost anywhere in the genome, but they are highly concentrated in proximal promoters. In past work (Smith et al., 2005b,a) we have been successful at identifying previously characterized motifs for factors known to play tissue-specific regulatory roles. Promoter quality (*i.e.* confidence in the transcription start site) has a large impact, and poor quality promoters may hurt motif discovery even more than poor quality sets of tissue specific transcripts (*e.g.* those containing ubiquitous or incorrectly assigned transcripts).

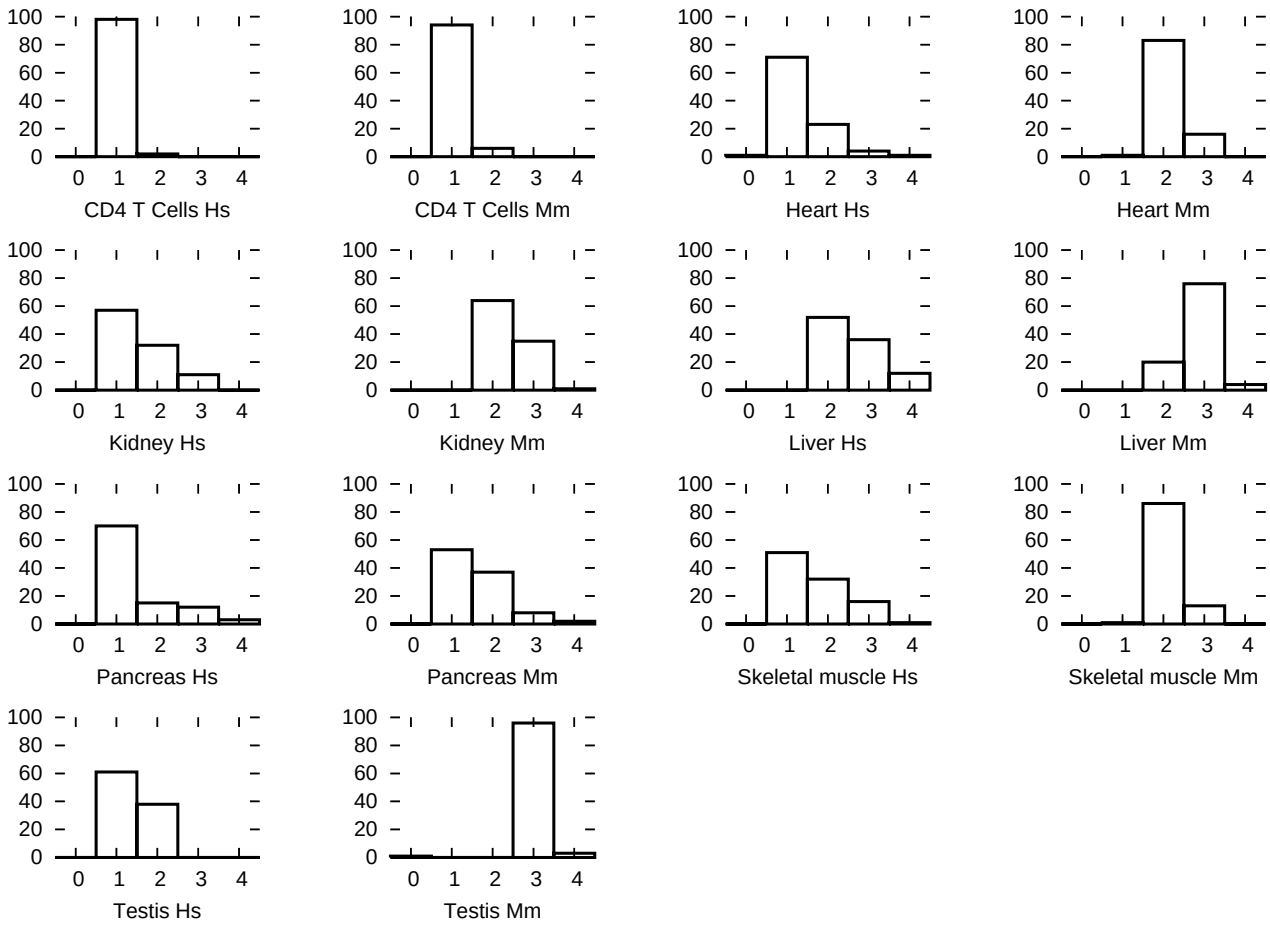

Figure 3. The distribution of votes per transcript corresponding to analyzed promoters sets. Each set included 100 promoters, and each promoter corresponded to a transcript with evidence for tissue specific regulation from possibly 0 to 4 sources. We display the number of transcripts (y-axis) with the given number of votes for tissue-specific regulation (x-axis). Distribution that are skewed to the left describe promoter sets with weaker evidence for tissue-specific regulation, and those that are skewed to the right describe promoter sets with stronger evidence.

Symbol	Name	Human RefSeq	Mouse RefSeq	Votes
HMGCS2	3-hydroxy-3-methylglutaryl-Coenzyme A synthase 2 (mitochondrial)	NM_005518	NM_008256	8
VTN	vitronectin (serum spreading factor, somatomedin B, complement S-protein)	NM_000638	NM_011707	7
SERPINC1	serpin peptidase inhibitor, clade C (antithrombin), member 1	NM_000488	NM_080844	7
KNG1	kininogen 1	NM_000893	NM_023125	7
HRG	histidine-rich glycoprotein	NM_000412	NM_053176	7
BAAT	bile acid Coenzyme A: amino acid N-acyltransferase (glycine N-choloyltransferase)	NM_001701	NM_007519	7
APOH	apolipoprotein H (beta-2-glycoprotein I)	NM_000042	NM_013475	7
ADH1A	alcohol dehydrogenase 1A (class I), alpha polypeptide	NM_000667	NM_007409	7
UGT2B4	UDP glucuronosyltransferase 2 family, polypeptide B4	NM_021139	NM_152811	6
TDO2	tryptophan 2,3-dioxygenase	NM_005651	NM_019911	6
PROC	protein C (inactivator of coagulation factors Va and VIIIa)	NM_000312	NM_008934	6
PLG	plasminogen	NM_000301	NM_008877	6
LIPC	lipase, hepatic	NM_000236	NM_008280	6
LCAT	lecithin-cholesterol acyltransferase	NM_000229	NM_008490	6
ITIH3	inter-alpha (globulin) inhibitor H3	NM_002217	NM_008407	6
FABP1	fatty acid binding protein 1, liver	NM_001443	NM_017399	6
CPB2	carboxypeptidase B2 (plasma, carboxypeptidase U)	NM_016413	NM_019775	6
ARG1	arginase, liver	NM_000045	NM_007482	6
AKR1C4	aldo-keto reductase family 1, member C4	NM_001818	NM_030611	6
SLC2A2	solute carrier family 2 (facilitated glucose transporter), member 2	NM_000340	NM_031197	5
PAH	phenylalanine hydroxylase	NM_000277	NM_008777	5
NR1H4	nuclear receptor subfamily 1, group H, member 4	NM_005123	NM_009108	5
MST1	macrophage stimulating 1 (hepatocyte growth factor-like)	NM_020998	NM_008243	5
MAT1A	methionine adenosyltransferase I, alpha	NM_000429	NM_133653	5
ITIH4	inter-alpha (globulin) inhibitor H4 (plasma Kallikrein-sensitive glycoprotein)	NM_002218	NM_018746	5
ITIH2	inter-alpha (globulin) inhibitor H2	NM_002216	NM_010582	5
HPX	hemopexin	NM_000613	NM_017371	5
HAO1	hydroxyacid oxidase (glycolate oxidase) 1	NM_017545	NM_010403	5
GNMT	glycine N-methyltransferase	NM_018960	NM_010321	5
FETUB	fetuin B	NM_014375	NM_021564	5
F13B	coagulation factor XIII, B polypeptide	NM_001994	NM_031164	5
F10	coagulation factor X	NM_000504	NM_007972	5
EBP	emopamil binding protein (sterol isomerase)	NM_006579	NM_007898	5
CYP2E1	cytochrome P450, family 2, subfamily E, polypeptide 1	NM_000773	NM_021282	5
CYP2C9	cytochrome P450, family 2, subfamily C, polypeptide 9	NM_000771	NM_007815	5
CRP	C-reactive protein, pentraxin-related	NM_000567	NM_007768	5
CPN1	carboxypeptidase N, polypeptide 1, 50kD	NM_001308	NM_030703	5
C8A	complement component 8, alpha polypeptide	NM_000562	NM_146148	5
C6	complement component 6	NM_000065	NM_016704	5
ASS	argininosuccinate synthetase	NM_000050	NM_007494	5
ASGR2	asialoglycoprotein receptor 2	NM_080914	NM_007493	5
APOM	apolipoprotein M	NM_019101	NM_018816	5
APOC3	apolipoprotein C-III	NM_000040	NM_023114	5
APOA2	apolipoprotein A-II	NM_001643	NM_013474	5
APCS	amyloid P component, serum	NM_001639	NM_011318	5
ALDOB	aldolase B, fructose-bisphosphate	NM_000035	NM_144903	5
AGXT	alanine-glyoxylate aminotransferase	NM_000030	NM_016702	5
ADH6	alcohol dehydrogenase 6 (class V)	NM_000672	NM_007409	5
ADH1B	alcohol dehydrogenase IB (class I), beta polypeptide	NM_000668	NM_007409	5
SLCO1B1	solute carrier organic anion transporter family, member 1B1	NM_006446	NM_020495	4
SLC22A1	solute carrier family 22 , member 1	NM_153187	NM_009202	4
SERPING1	serpin peptidase inhibitor, clade G	NM_000062	NM_009776	4
SEC14L2	SEC14-like 2 (S. cerevisiae)	NM_012429	NM_144520	4
SDS	serine dehydratase	NM_006843	NM_145565	4
RGN	regucalcin (senescence marker protein-30)	NM_152869	NM_009060	4
POR	P450 (cytochrome) oxidoreductase	NM_000941	NM_008898	4
PMVK	phosphomevalonate kinase	NM_006556	NM_026784	4
PEMT	phosphatidylethanolamine N-methyltransferase	NM_148173	NM_008819	4
ORM1	orosomucoid 1	NM_000607	NM_008768	4
ITIH1	inter-alpha (globulin) inhibitor H1	NM_002215	NM_008406	4
IGFALS	insulin-like growth factor binding protein, acid labile subunit	NM_004970	NM_008340	4
HP	haptoglobin	NM_005143	NM_017370	4
GGCX	gamma-glutamyl carboxylase	NM_000821	NM_019802	4
CYP27A1	cytochrome P450, family 27, subfamily A, polypeptide 1	NM_000784	NM_024264	4
C4BPA	complement component 4 binding protein, alpha	NM_000715	NM_007576	4
ASL	argininosuccinate lyase	NM_000048	NM_133768	4
AHSG	alpha-2-HS-glycoprotein	NM_001622	NM_013465	4
ACOX2	acyl-Coenzyme A oxidase 2, branched chain	NM_003500	NM_053115	4
AADAC	arylacetamide deacetylase (esterase)	NM_001086	NM_023383	4

Table 5. Transcripts with multiple votes for liver-specificity in both human and mouse. The “Votes” column gives the total number of votes for liver-specificity in both human and mouse.

Symbol	Name	Human RefSeq	Mouse RefSeq	Votes
CD3D	CD3D antigen, delta polypeptide (TiT3 complex)	NM_000732	NM_013487	4

Table 6. Transcripts with multiple votes for CD4 T cells-specificity in both human and mouse. The “Votes” column gives the total number of votes for CD4 T cells-specificity in both human and mouse.

Symbol	Name	Human RefSeq	Mouse RefSeq	Votes
MYL7	myosin, light polypeptide 7, regulatory	NM_021223	NM_022879	6
CSRP3	cysteine and glycine-rich protein 3 (cardiac LIM protein)	NM_003476	NM_013808	6
CASQ2	calsequestrin 2 (cardiac muscle)	NM_001232	NM_009814	6
TNNT2	troponin T type 2 (cardiac)	NM_000364	NM_011619	5
NKX2-5	NK2 transcription factor related, locus 5 (Drosophila)	NM_004387	NM_008700	5
MYBPC3	myosin binding protein C, cardiac	NM_000256	NM_008653	5
ACTC	actin, alpha, cardiac muscle	NM_005159	NM_009608	5
S100A1	S100 calcium binding protein A1	NM_006271	NM_011309	4
RYR2	ryanodine receptor 2 (cardiac)	NM_001035	NM_023868	4
MYOZ2	myozenin 2	NM_016599	NM_021503	4
MFN2	mitofusin 2	NM_014874	NM_133201	4

Table 7. Transcripts with multiple votes for heart-specificity in both human and mouse. The "Votes" column gives the total number of votes for heart-specificity in both human and mouse.

Symbol	Name	Human RefSeq	Mouse RefSeq	Votes
SLC34A1	solute carrier family 34 (sodium phosphate), member 1	NM_003052	NM_011392	6
UMOD	uromodulin (uromucoid, Tamm-Horsfall glycoprotein)	NM_003361	NM_009470	5
PTHR1	parathyroid hormone receptor 1	NM_000316	NM_011199	5
NAT8	N-acetyltransferase 8 (camello like)	NM_003960	NM_023455	5
FMO1	flavin containing monooxygenase 1	NM_002021	NM_010231	5
DAO	D-amino-acid oxidase	NM_001917	NM_010018	5
CDH16	cadherin 16, KSP-cadherin	NM_004062	NM_007663	5
SLC6A13	solute carrier family 6 (neurotransmitter transporter, GABA), member 13	NM_016615	NM_144512	4
SLC27A2	solute carrier family 27 (fatty acid transporter), member 2	NM_003645	NM_011978	4
SLC17A1	solute carrier family 17 (sodium phosphate), member 1	NM_005074	NM_009198	4
PCK1	phosphoenolpyruvate carboxykinase 1 (soluble)	NM_002591	NM_011044	4
FXYD2	FXYD domain containing ion transport regulator 2	NM_021603	NM_052824	4
FOLR1	folate receptor 1 (adult)	NM_016731	NM_008034	4
ASS	argininosuccinate synthetase	NM_054012	NM_007494	4
ALDRL6	myo-inositol oxygenase	NM_017584	NM_019977	4

Table 8. Transcripts with multiple votes for kidney-specificity in both human and mouse. The "Votes" column gives the total number of votes for kidney-specificity in both human and mouse.

Symbol	Name	Human RefSeq	Mouse RefSeq	Votes
ELA3A	elastase 3A, pancreatic	NM_005747	NM_026419	7
ELA2A	elastase 2A	NM_033440	NM_007919	7
PRSS3	protease, serine, 3 (mesotrypsin)	NM_002771	NM_009430	6
PNLIPRP2	pancreatic lipase-related protein 2	NM_005396	NM_011128	6
ELA3B	elastase 3B, pancreatic	NM_007352	NM_026419	6
CTRL	chymotrypsin-like	NM_001907	NM_023182	6
PNLIPRP1	pancreatic lipase-related protein 1	NM_006229	NM_018874	5
PNLIP	pancreatic lipase	NM_000936	NM_026925	5
KLK1	kallikrein 1, renal/pancreas/salivary	NM_002257	NM_010639	5
CTRB1	chymotrypsinogen B1	NM_001906	NM_025583	5
CEL	carboxyl ester lipase (bile salt-stimulated lipase)	NM_001807	NM_009885	5
REG1B	regenerating islet-derived 1 beta (pancreatic stone protein, pancreatic thread protein)	NM_006507	NM_009042	4
REG1A	regenerating islet-derived 1 alpha (pancreatic stone protein, pancreatic thread protein)	NM_002909	NM_009043	4
RBPSUHL	recombining binding protein suppressor of hairless (Drosophila)-like	NM_014276	NM_009036	4
INS	insulin	NM_000207	NM_008387	4
GP2	glycoprotein 2 (zymogen granule membrane)	NM_001502	NM_025989	4
CLPS	colipase, pancreatic	NM_001832	NM_025469	4

Table 9. Transcripts with multiple votes for pancreas-specificity in both human and mouse. The "Votes" column gives the total number of votes for pancreas-specificity in both human and mouse.

Symbol	Name	Human RefSeq	Mouse RefSeq	Votes
ODF1	outer dense fiber of sperm tails 1	NM_024410	NM_008757	6
TNP1	transition protein 1 (during histone to protamine replacement)	NM_003284	NM_009407	5
SPINLW1	serine peptidase inhibitor-like, with Kunitz and WAP domains 1 (eppin)	NM_020398	NM_029325	5
PHF7	PHD finger protein 7	NM_173341	NM_027949	5
MCSP	sperm mitochondria-associated cysteine-rich protein	NM_030663	NM_008574	5
DPEP3	dipeptidase 3	NM_022357	NM_027960	5
ACTL7A	actin-like 7A	NM_006687	NM_009611	5
ZPBP	zona pellucida binding protein	NM_007009	NM_015785	4
PRM1	protamine 1	NM_002761	NM_013637	4
LDHC	lactate dehydrogenase C	NM_017448	NM_013580	4
GAPDS	glyceraldehyde-3-phosphate dehydrogenase, spermatogenic	NM_014364	NM_008085	4
CRISP2	cysteine-rich secretory protein 2	NM_003296	NM_009420	4
ACTL7B	actin-like 7B	NM_006686	NM_025271	4

Table 10. Transcripts with multiple votes for testis-specificity in both human and mouse. The "Votes" column gives the total number of votes for testis-specificity in both human and mouse.

	Human	Mouse
Promoters in CSHLmpd	51506	46475
RefSeq transcripts with a promoter in CSHLmpd	16433	15061
RefSeq transcripts with known TSS (EPD, DBTSS and GenBank)	7212	4742

Table 11. Summary of promoter data.

The main resource for mapping transcripts to promoters is the CSHL mammalian promoter database (CSHLmpd) (Xuan et al., 2005), which includes human, mouse and rat. CSHLmpd includes experimentally confirmed promoters that are annotated in EPD (Perier et al., 1998), DBTSS (Suzuki et al., 2002) and GenBank, as well as computationally predicted promoters. All but 27 transcripts with an experimentally verified transcription start site (TSS) correspond to RefSeqs (see Table 11 for CSHLmpd statistics). may be associated with the same TSS.

Repetitive regions in promoters. Human and mouse proximal promoters contain a high proportion of LINEs (Human), SINEs and simple repeats. Repeats were not masked for most of our analysis. If tissue-specific promoters have no special relation with any particular kind of repeat, the repeats will have no effect on our analysis because we measured enrichment of motifs and modules relative to background sets composed of real promoters, and not simplified statistical models. Repeat regions make a considerable portion of the nucleotide content of our foreground-set promoters. Using RepeatMasker (Bedell et al., 2000) to mask primate and rodent repeats in the human and mouse tissue-specific promoters indicates that 12% to 26% of the nucleotides are within repeat regions. The breakdown per tissue is given in Table 12.

CpG-related promoters. Vertebrate promoter sequences are enriched with CpG islands and these are often used to help identify promoters. CpG islands tend to occur with higher frequency in promoters of housekeeping genes than tissue-specific genes (Gardiner-Garden and Frommer, 1987). Because CpG island frequency is related to tissue-specific transcription, we calculated CpG frequency and GC-content in our promoter sets. The Gardiner-Garden method (Gardiner-Garden and Frommer, 1987) requires CpG islands to have GC-content at least 0.50, length greater than 200bp, and number of CpG dinucleotides at least 0.6 times the number expected based on the GC-content. The Takai method (Takai and Jones, 2002) refines the Gardiner-Garden method with attention to detection of human CpG islands. We say that a promoter is CpG

Tissue	Human	Mouse
CD4 T-cells	20.22%	16.55%
Heart	15.96%	10.71%
Kidney	17.21%	15.47%
Liver	22.77%	16.09%
Pancreas	22.23%	18.24%
Skeletal muscle	11.73%	12.61%
Testis	25.76%	17.40%

Table 12. Proportions of RepeatMasker-masked nucleotides from total promoter length.

Tissue	Human	Mouse
CD4 T-cells	0.68	0.29
Testis	0.73	0.42
Heart	0.42	0.36
Skeletal muscle	0.35	0.46
Liver	0.32	0.21
Kidney	0.19	0.29
Pancreas	0.19	0.16
Ubiq	0.91	0.83

Table 13. Proportion of CpG related promoters in our tissue specific sets and in a set corresponding to ubiquitous transcripts (house keeping).

related if a CpG island is detected less than 2000 bases upstream or 500 bases downstream from the TSS. To compare the proportion of CpG related promoters in our tissue-specific sets to the proportion of of CpG related promoters of house keeping genes we selected transcripts with EST evidence in multiple tissues. In both mouse and human we ranked transcripts to minimize the proportion of EST evidence in any given tissue, and selected the top 1000. Table 13 gives a comparison between CpG related promoters in these sets and CpG related promoters in tissue specific sets.

1.7 Tissue-expressed transcription factors

We attempted to identify transcription factors that are expressed in a given tissue using the same data used to identify tissue specific sets. This data includes GNF SymAtlas, the Hughes Toronto array, and EST data from dbEST. We were interested in expression evidence and were not concerned with tissue-specificity, however attempts to identify expression of a single transcript are fundamentally different than attempts to construct transcript sets. The former require each transcript to have strong evidence, while the later are robust to outliers and are evaluated as a set. For this reason we elected not use the GO annotation for calling tissue-specificity of single transcripts. Due to data and analysis constraints, we did not require multiple confirmations for factor expression and simply described the evidence for each factor in TCat. The highest confidence data had multiple votes, few absent calls from the microarray expression data, and high proportion in the given tissue according to dbEST.

2 Motifs and redundancy elimination

Previously characterized motifs were taken from the vertebrate subset of the matrix table of TRANSFAC version 9.1 (Matys et al., 2003). We used UNIQMOTIFS (Smith et al., 2005c) to eliminate redundancies in the presentation. A motif was considered redundant if it was similar to a higher ranked motif, and similarity was measured using the Kullback-Leibler divergence between position frequency matrices (Smith et al., 2005b). We associated redundant motifs with the highest ranking similar non-redundant motif.

2.1 Significance of ranks for known motifs

The identities of known tissue-specific regulators along with a characterization of their binding sites, can be used to verify our ranking method. Skeletal muscle and liver are well studied (Odom et al., 2004; Krivan and Wasserman, 2001; Wasserman and Fickett, 1998; Johnson et al., 2005), and the transcription factors known to have large functional roles in these tissues are well characterized. In liver, the most important factors are known to be HNF-1, HNF-3, HNF-4, C/EBP, VDR and CDP (Schrem et al., 2002, 2004). In skeletal muscle, the most important factors are known to be MEF-2, SRF, PAX, and members of the Myogenin family (Duprey and Lesens, 1994). We used the Wilcoxon signed ranks test to determine if the ranks of the matrices associated with these factors received significantly high ranks when the entire set of 554 vertebrate matrices from TRANSFAC (release 9.1) were ranked according to importance.

There were a total of 37 matrices for liver, the sum of their ranks was 2574, and the associated p-value was 1.38×10^{-15}. There was also a total of 37 matrices for skeletal muscle, with a rank sum of 5115.5 and an associated p-value of 6.07×10^{-8}.

2.2 Models and significance of motifs and modules.

We used position-weight matrices to model transcription factor binding sites, and top-scoring sites in each promoter to measure motif enrichment, but these are not always appropriate. For example, the POU3 family of factors bind to a pair of sites that can be separated by up to three nucleotides (Li et al., 1993). Such motifs cannot be accurately described using position-weight matrices, and we expect that large classes of factors will be identified with binding specificity that is poorly characterized by position-weight matrices. Recent results show that true binding sites for certain factors are often accompanied by nearby lower-affinity sites (Zhang et al., 2006), and for such factors this information should be considered when evaluating motif enrichment.

Modeling the organization of regulatory modules is a formidable challenge. Some experimental evidence suggest that conserved relative spacing and orientation between module sites is important, while other evidence suggest the opposite; it is likely that no single organization model can adequately describe all modules (Erives and Levine, 2004). We did not use conservation of order, spacing or orientation to reverse engineer modules, but believe that in many cases co-linear order is conserved on the module level. Our method for identifying modules required that each component (*i.e.* motif) in the module contribute significantly to the quality of the module, which provides statistical evidence for functional interaction between all motifs in the module. We found that the pairing of myogenin family members with SRF or MEF-2 produces significant modules, but pairing of MEF-2 and SRF is not significantly enriched in skeletal-muscle specific promoters. This finding suggests that MEF-2 and SRF regulate transcription in skeletal muscle independently. We note that computational identification of module component sites is especially challenging and depends on interactions between the components. In present work, when several possible sites of variable affinity were predicted, we chose to annotate only the top-scoring sites.

Program	Width	Bits	Gran.	Refine	N
DME	12	1.55	0.5	0.25	25
	10	1.6	D	0.125	50
	8	1.8	D	0.125	50
DME-B	10	1.6	0.5	0.25	25
	8	1.8	D	0.125	50

Table 14. Parameters used for runs of DME and DME-B to discover motifs *de novo* that are enriched in the tissue-specific promoter sets relative to the background sets. The *Bits* value refers to bits per column, the *Gran.* value refers to the granularity, the *Refine* value describes the parameter of the refinement procedure, and the N value is the number of motifs requested. More information on these parameters can be found in (Smith et al., 2005b)

Binding specificity for some transcription factors may be poorly characterized. Examples include characterizations based on too few sites, and *in vitro* verification in cell-lines where a factor has a different conformation or affinity for a particular site. An example of a factor with possibly incomplete characterization of binding specificity is HNF-6. Our data did not include evidence for its expression in liver but HNF-6 is a known liver regulator (Samadani and Costa, 1996), whose TRANSFAC motif was not found to be enriched in our liver promoter sets. The HNF-6 binding motif was derived from a small set of sites identified in a restricted context (Samadani and Costa, 1996; Lannoy et al., 1998), and other evidence suggests that HNF-6 has two modes of binding to DNA (Lannoy et al., 1998).

2.3 Motifs identified *de novo*

Motifs discovered *de novo* were obtained using the DME (Smith et al., 2005b) and DME-B (Smith et al., 2006) algorithms. We used the parameter sets in Table 14. Redundant motifs were combined with the non-redundant one to which they were associated, and the combination was optimized greedily with respect to the balanced error rate. The program used to combine the motifs and optimize with respect to motif importance is available from the authors. When a *de novo* identified motif M was similar to an experimentally verified motif M', we annotated M with the factor that is known to bind sequences matching M'. Similarity was measured using Kullback-Leibler divergence (Kullback and Leibler, 1951), using MATCOMPARE (Schones et al., 2005) (available in CREAD), and motifs M and M' of length $m \leq m'$ were called similar if the motifs could be aligned in at least $m - 1$ positions without gaps with K-L divergence below 1.0.

2.4 Motif rank and order correlation between human and mouse

To measure correlation between the tissue-specific regulatory apparatus in human and mouse, we compared single motif ranks. This comparison method is largely independent of sequence similarity and ortholog information, and is used to compare over- and under-representation of motifs in our foreground sets. Motif ranking summarizes known information in each foreground set, and is therefore ideal for measuring regulatory correlation. We used the Spearman rank correlation test (Altman, 1991) to compare vertebrate motif ranks in human and mouse. We found significant correlation in all but CD4 T-cells (Table 15).

To test whether the correlation can be detected using standard sequence alignment methods, we identified the highest likelihood sites for the vertebrate motif set in each proximal promoter and its ortholog, and compared the order of the sites in the pair using a Wilcoxon signed-ranks test. When the top score

Tissue	CD4 T-cells	Pancreas	Testis	Kidney	Skeletal muscle	Liver	Heart
p-val	0.0618	1.49E-07	5.63E-12	3.87E-12	3.19E-19	5.10E-29	6.46E-57

Table 15. Correlation between vertebrate motif ranks in human and mouse tissues.

was shared by multiple sites we chose the site that best matched the ortholog order. We noticed that promoters with significant co-linear site conservation also had high sequence similarity. We measured similarity as the proportion of nucleotides matched by a clustal w alignment. Our data included 1, 17, 6, 12, 15, 40, and 11 ortholog promoters for pancreas, testis, kidney, skeletal muscle, liver, and heart, respectively. Sequence similarity over the 102 promoter pairs was 0.51 ± 0.079, which is significantly higher than random promoter pairs (0.411 ± 0.022), and 9 pairs were found to have significant co-linear site conservation. In testis, the proximal promoter for PHD finger protein 7 (PHF7) (NM_027949 and NM_016483) had significant co-linear conservation; the similarity between the promoters was 0.65. In kidney, solute carrier family 27 (SLC27A2) (NM_003645 and NM_011978) with 0.49 similarity. In skeletal muscle, troponin C type 2 (TNNC2) (NM_003279 and NM_009394) with 0.61 similarity, and myogenic factor 6 (MYF6) (NM_002469 and NM_008657) with 0.76 similarity. In liver, argininosuccinate synthetase (ASS) (NM_000050 and NM_007494) with 0.56 similarity, and lecithin-cholesterol acyltransferase (LCAT) (NM_008490 and NM_000229) with 0.70 similarity. In heart, mitofusin 2 (MFN2) (NM_133201 and NM_014874) with 0.42 similarity, ryanodine receptor 2 (RYR2) (NM_001035 and NM_023868) with 0.47 similarity, and actin alpha cardiac muscle (ACTC) (NM_009608 and NM_005159) with 0.68 similarity. The result suggests that only a small proportion of the highest-likelihood ortholog sites can be recovered using traditional multiple sequence alignment.

References

A. Agresti. A survey of exact inference for contingency tables. *Statistical Science*, 7:131–177, 1992.

D. G. Altman. *Practical Statistics for Medical Research*. Chapman & Hall, London, 1991.

M. Ashburner et al. Gene ontology: tool for the unification of biology. *Nat Genet.*, 25(1):25–29, 2000.

J. A. Bedell, I. Korf, and W. Gish. Maskeraid: a performance enhancement to RepeatMasker. *Bioinformatics*, 16(11):1040–1041, 2000.

M. S. Boguski, T. M. Lowe, and C. M. Tolstoshev. dbEST – database for expressed sequence tags. *Nat Genet.*, 4(4):332–333, 1993.

L. Boxer, T. Miwa, T. Gustafson, and L. Kedes. Identification and characterization of a factor that binds to two human sarcomeric actin promoters. *J. Biol. Chem.*, 264(2):1284–1292, 1989.

P. Carninci et al. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet.*, 38 (6):626–635, 2006.

A. Charbonneau and V. Luu-The. Assignment of steroid 5beta-reductase (SRD5B1) and its pseudogene (SRD5BP1) to human chromosome bands 7q32–>q33 and 1q23–>q25, respectively, by in situ hybridization. *Cytogenet Cell Genet*, 84(1-2):105–106, 1999.

M. J. Clarkson and V. R. Harley. Sex with two SOX on: SRY and SOX9 in testis development. *Trends in Endocrinology and Metabolism*, 13(3):106–111, 2002.

M. K. Connor, I. Irrcher, and D. A. Hood. Contractile Activity-induced Transcriptional Activation of Cytochrome c Involves Sp1 and Is Proportional to Mitochondrial ATP Synthesis in C2C12 Muscle Cells. *J. Biol. Chem.*, 276(19):15898–15904, 2001.

S. J. Cooper, N. D. Trinklein, E. D. Anton, L. Nguyen, and R. M. Myers. Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome. *Genome Res.*, 16(1):1–10, 2006.

P. Duprey and C. Lesens. Control of skeletal muscle-specific transcription: involvement of paired homeodomain and mads domain transcription factors. *Int. J. Dev. Biol.*, 38:591–604, 1994.

A. Erives and M. Levine. Coordinate enhancers share common organizational features in the *Drosophila* genome. *Proc. Natl. Acad. Sci. USA*, 101(11):3851–3856, 2004.

R. M. Evans. The Nuclear Receptor Superfamily: A Rosetta Stone for Physiology. *Mol Endocrinol*, 19(6): 1429–1438, 2005.

M. Frain, E. Hardon, G. Ciliberto, and J. M. Sala-Trepat. Binding of a liver-specific factor to the human albumin gene promoter and enhancer. *Mol Cell Biol.*, 10(3):991–999, 1990.

M. C. Frith, J. Ponjavic, D. Fredman, C. Kai, J. Kawai, P. Carninci, Y. Hayashizaki, and A. Sandelin. Evolutionary turnover of mammalian transcription start sites. *Genome Res.*, 16(6):713–722, 2006.

P. Fuchs, M. Zorer, G. Rezniczek, D. Spazierer, S. Oehler, M. Castanon, R. Hauptmann, and G. Wiche. Unusual 5' transcript complexity of plectin isoforms: novel tissue-specific exons modulate actin binding activity. *Hum. Mol. Genet.*, 8(13):2461 – 2472, 1999.

M. Gardiner-Garden and M. Frommer. CpG islands in vertebrate genomes. *J. Mol. Biol.*, 196(2):261–282, 1987.

M. Gupta and J. S. Liu. De novo cis-regulatory module elicitation for eukaryotic genomes. *Proc. Natl. Acad. Sci. USA*, 102(20):7079–7084, 2005.

J. Jensen. Gene regulatory factors in pancreatic development. *Developmental Dynamics*, 229(1):176–200, 2004.

D. S. Johnson, Q. Zhou, K. Yagi, N. Satoh, W. Wong, and A. Sidow. De novo discovery of a tissue-specific gene regulatory module in a chordate. *Genome Res.*, page gr.4062605, 2005.

J. M. Johnson, J. Castle, P. Garrett-Engele, Z. Kan, P. M. Loerch, C. D. Armour, R. Santos, E. E. Schadt, R. Stoughton, and D. D. Shoemaker. Genome-Wide Survey of Human Alternative Pre-mRNA Splicing with Exon Junction Microarrays. *Science*, 302(5653):2141–2144, 2003.

T. H. Kim, L. O. Barrera, M. Zheng, C. Qu, M. A. Singer, T. A. Richmond, Y. Wu, R. D. Green, and B. Ren. A high-resolution map of active promoters in the human genome. *Nature*, 436(7052):876–880, 2005.

S. Kimmins, N. Kotaja, I. Davidson, and P. Sassone-Corsi. Testis-specific transcription mechanisms promoting male germ-cell differentiation. *Reproduction*, 128(1):5–12, 2004.

W. Krivan and W. W. Wasserman. A predictive model for regulatory sequences directing liver-specific transcription. *Genome Res.*, 11:1559–1966, 2001.

S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:76–86, 1951.

V. J. Lannoy, T. R. Burglin, G. G. Rousseau, and F. P. Lemaigre. Isoforms of Hepatocyte Nuclear Factor-6 Differ in DNA-binding Properties, Contain a Bifunctional Homeodomain, and Define the New ONECUT Class of Homeodomain Proteins. *J. Biol. Chem.*, 273(22):13552–13562, 1998.

P. Li, X. He, M. Gerrero, M. Mok, A. Aggarwal, and M. Rosenfeld. Spacing and orientation of bipartite DNA-binding motifs as potential functional determinants for POU domain factors. *Genes Dev.*, 7(12B): 2483–4, 1993.

X. Li, J. H. Huang, H. Y. Rienhoff jr, and W. S.-L. Liao. Two adjacent c/ebp-binding sequences that participate in the cell-specific expression of the mouse serum amyloid a3 gene. *Mol. Cell. Biol.*, 10: 6624–6631, 1990.

W. MacLellan, T. Lee, R. Schwartz, and M. Schneider. Transforming growth factor-beta response elements of the skeletal alpha- actin gene. Combinatorial action of serum response factor, YY1, and the SV40 enhancer-binding protein, TEF-1. *J. Biol. Chem.*, 269(24):16754–16760, 1994.

V. Matys et al. TRANSFAC(R): transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, 31 (1):374–378, 2003.

K. Morotomi-Yano, K.-i. Yano, H. Saito, Z. Sun, A. Iwama, and Y. Miki. Human Regulatory Factor X 4 (RFX4) Is a Testis-specific Dimeric DNA-binding Protein That Cooperates with Other Human RFX Members. *J. Biol. Chem.*, 277(1):836–842, 2002.

G. E. Muscat, T. A. Gustafson, and L. Kedes. A common factor regulates skeletal and cardiac alpha-actin gene transcription in muscle. *Mol Cell Biol.*, 8(10):4120–4133, 1988.

S. Nelander, E. Larsson, E. Kristiansson, R. Mansson, O. Nerman, M. Sigvardsson, P. Mostad, , and P. Lindahl. Predictive screening for regulators of conserved functional gene modules (gene batteries) in mammals. *BMC Genomics*, 6(68), 2005.

D. T. Odom, N. Zizlsperger, D. B. Gordon, G. W. Bell, N. J. Rinaldi, H. L. Murray, T. L. Volkert, J. Schreiber, P. A. Rolfe, D. K. Gifford, E. Fraenkel, G. I. Bell, and R. A. Young. Control of Pancreas and Liver Gene Expression by HNF Transcription Factors. *Science*, 303(5662):1378–1381, 2004.

G. Paonessa, F. Gounari, R. Frank, and R. Cortese. Purification of a NF1-like DNA-binding protein from rat liver and cloning of the corresponding cDNA. *EMBO J.*, 7:3115–3123, 1988.

R. C. Perier, T. Junier, and P. Bucher. The eukaryotic promoter database EPD. *Nucleic Acids Res.*, 26(1): 353–357, 1998.

B. Ren and B. D. Dynlacht. Use of chromatin immunoprecipitation assays in genome-wide location analysis of mammalian transcription factors. *Methods Enzymol*, 376:304–315, 2004.

G. Robertson, M. Bilenky, K. Lin, A. He, W. Yuen, M. Dagpinar, R. Varhol, K. Teague, O. L. Griffith, X. Zhang, Y. Pan, M. Hassel, M. C. Sleumer, W. Pan, E. D. Pleasance, M. Chuang, H. Hao, Y. Y. Li, N. Robertson, C. Fjell, B. Li, S. B. Montgomery, T. Astakhova, J. Zhou, J. Sander, A. S. Siddiqui, and S. J. M. Jones. cisRED: a database system for genome-scale computational discovery of regulatory elements. *Nucl. Acids Res.*, 34(Suppl 1):D68–73, 2006.

U. Samadani and R. Costa. The transcriptional activator hepatocyte nuclear factor 6 regulates liver gene expression. *Mol. Cell. Biol.*, 16(11):6273–6284, 1996.

A. Sandelin, W. Alkema, P. Engström, W. W. Wasserman, and B. Lenhard. JASPAR: an open access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, 32(1):D91–D94, 2004.

K. Sawadaishi, T. Morinaga, and T. Tamaoki. Interaction of a hepatoma-specific nuclear factor with transcription-regulatory sequences of the human alpha-fetoprotein and albumin genes. *Mol Cell Biol.*, 8(12):5179–5187, 1988.

D. Schones, P. Sumazin, and M. Q. Zhang. Similarity of position frequency matrices for transcription factor binding sites. *Bioinformatics*, 21(3):307–313, 2005.

H. Schrem, J. Klempnauer, and J. Borlak. Liver-Enriched Transcription Factors in Liver Function and Development. Part I: The Hepatocyte Nuclear Factor Network and Liver-Specific Gene Expression. *Pharmacol Rev*, 54(1):129–158, 2002.

H. Schrem, J. Klempnauer, and J. Borlak. Liver-Enriched Transcription Factors in Liver Function and Development. Part II: the C/EBPs and D Site-Binding Protein in Cell Cycle Control, Carcinogenesis, Circadian Gene Regulation, Liver Regeneration, Apoptosis, and Liver-Specific Gene Regulation. *Pharmacol Rev*, 56(2):291–330, 2004.

O. Shmueli, S. Horn-Saban, V. Chalifa-Caspi, M. Shmoish, R. Ophir, H. Benjamin-Rodrig, M. Safran, E. Domany, and D. Lancet. GeneNote: whole genome expression profiles in normal human tissues. *C. R. Biologies*, 326, 2003.

A. D. Smith, P. Sumazin, D. Das, and M. Q. Zhang. Mining ChIP-chip data for transcription factor and cofactor binding sites. *Bioinformatics*, 21(Suppl 1):i403–i412, 2005a.

A. D. Smith, P. Sumazin, and M. Q. Zhang. Identifying tissue-selective transcription factor binding sites in vertebrate promoters. *Proc. Natl. Acad. Sci. USA*, 102(5):1560–1565, 2005b.

A. D. Smith, P. Sumazin, and M. Q. Zhang. CREAD: Comprehensive regulatory element analysis and discovery. World Wide Web (http://cread.sf.net/), 2005c.

A. D. Smith, P. Sumazin, Z. Xuan, and M. Q. Zhang. DNA motifs in human and mouse proximal promoters predict tissue-specific expression. *Proc. Natl. Acad. Sci. USA*, 104(16):6275–6280, 2006.

J. D. Storey and R. Tibshirani. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA*, 100(16):9440–9445, 2003.

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

A. I. Su, T. Wiltshire, S. Batalov, H. Lapp, K. A. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, G. Kreiman, M. P. Cooke, J. R. Walker, and J. B. Hogenesch. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. USA*, 101(16):6062–6067, 2004.

Y. Suzuki, R. Yamashita, K. Nakai, and S. Sugano. DBTSS: DataBase of human Transcriptional Start Sites and full-length cDNAs. *Nucleic Acids Res.*, 30(1):328–331, 2002.

D. Takai and P. A. Jones. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc. Natl. Acad. Sci. USA*, 99(6):3740–3745, 2002.

W. Thompson, M. J. Palumbo, W. W. Wasserman, J. S. Liu, and C. E. Lawrence. Decoding human regulatory circuits. *Genome Res.*, 14(10A):1967–1974, 2004.

Y. Urano, K. Watanabe, M. Sakai, and T. Tamaoki. The human albumin gene. Characterization of the 5' and 3' flanking regions and the polymorphic gene transcripts. *J. Biol. Chem.*, 261(7):3244–3251, 1986.

W. W. Wasserman and J. W. Fickett. Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.*, 278:167–181, 1998.

X. Xie, J. Lu, E. J. Kulbokas, T. R. Golub, V. Mootha, K. Lindblad-Toh, E. S. Lander, and M. Kellis. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature*, 434(7031):338–345, 2005.

Z. Xuan, F. Zhao, J. H. Wang, G. X. Chen, and M. Q. Zhang. Genome-wide promoter extraction and analysis in human, mouse, and rat. *Genome Biology*, 6:R72, 2005.

R. Yamashita, Y. Suzuki, H. Wakaguri, K. Tsuritani, K. Nakai, and S. Sugano. DBTSS: DataBase of human transcription start sites, progress report 2006. *Nucleic Acids Res.*, 34(Database issue):D86–D89, 2006.

C. Zhang, Z. Xuan, S. Otto, J. R. Hover, S. R. McCorkle, G. Mandel, and M. Q. Zhang. A clustering property of highly-degenerate transcription factor binding sites in the mammalian genome. *Nucleic Acids Res.*, 34(8):2238–2246, 2006.

T. Zhang, P. Haws, and Q. Wu. Multiple Variable First Exons: A Mechanism for Cell- and Tissue-Specific Gene Regulation. *Genome Res.*, 14(1):79–89, 2004.

W. Zhang et al. The functional landscape of mouse gene expression. *J. Biol.*, 3(5):21–21, 2004.

Q. Zhou and W. H. Wong. CisModule: De novo discovery of cis-regulatory modules by hierarchical mixture modeling. *Proc. Natl. Acad. Sci. USA*, 101(33):12114–12119, 2004.

Z. Zhu, J. Shendure, and G. M. Church. Discovering functional transcription-factor combinations in the human cell cycle. *Genome Res.*, 15(6):848–855, 2005.