Supplemental Methods

cDNA/EST alignment and DBTSS mapping

We used BLAT (Kent 2002) version 33 to align 24,308 RefSeq (Pruitt et al. 2000) Release 15 and 30,284 H-Invitational (Imanishi et al. 2004) Release 2.0 human near-full-length protein-coding cDNAs against NCBI human genome (Lander et al. 2001) Build 35, and 18,613 RefSeq Release 15 and 40,078 FANTOM DB (Okazaki et al. 2002) 3.00 mouse cDNAs against mouse genome (Waterston et al. 2002) Build 35 to identify exon/intron structures. Transcripts that were in the same genomic orientation and had at least 50 bases of exonic overlap were considered to represent the same gene. We filtered out problematic cDNA alignments as in our previous study (Baek and Green 2005), obtaining 37,184 aligned human cDNAs (called "reference cDNAs" below) for 15,658 distinct genes, and 44,352 aligned mouse cDNAs for 14,763 genes.

We obtained 7.50 million alignments of human cDNAs and ESTs to the human genome from the University of California, Santa Cruz (UCSC), Genome Bioinformatics Site (Karolchik et al. 2003) (Feb 2, 2006, release), which reduced to 3.39 million alignments after filtering as described previously (Baek and Green 2005). The same procedures were applied to 4.69 million alignments (1.60 million after filtering) of mouse cDNAs and ESTs (Feb 2, 2006, release) to the mouse genome.

We obtained human and mouse genomic coordinates of transcription start sites (TSSs) from the data base of transcription start sites (Suzuki et al. 2002) (DBTSS). (Since genomic coordinates of mouse TSSs were based on an older version (Build 33), we transferred these to Build 35 by aligning the 1 kb segment of Build 33 upstream of each TSS to Build 35 using BLAT.) When the genomic coordinate of the 5' end of a cDNA or EST was close to a mapped DBTSS TSS (distance < 3 bases), we classified it as a "5'-complete" transcript. There were 0.60 million 5'-complete transcripts in human and 0.16 million in mouse.

cDNA clustering and identification of conserved promoters

For each human gene, aligned reference and 5'-complete cDNAs were clustered together if the genomic coordinates of their 5' ends differed by <500 bases. (Exons were not required to overlap). Within each cluster, we defined subclusters as sets of isoforms with identical first exon genomic coordinates, and used counts of matching 5'-complete ESTs as a rough measure of the expression level of each subcluster. A 'representative isoform' for the cluster was then arbitrarily selected from the isoforms in the most highly expressed subcluster. Mouse representative isoforms were identified in the same way. Each representative isoform was considered to derive from a distinct promoter, and if the first exons of human and mouse representative isoforms were orthologous (according to the UCSC BLASTZ (Schwartz et al. 2003) alignments of the mouse and human genomes), the promoter was considered to be conserved. We did not impose any alignment requirement on the promoter regions themselves, in order to avoid biasing the sequence analyses.

Conserved SPs were identified as follows. We first found all human genes having a single full-length cDNA cluster (as defined above) and meeting the additional conditions that the first exons of all cDNAs in the cluster overlapped, and no EST from the same gene had its first exon entirely upstream of the cDNAs or had its 5' end >500 bases upstream from the cDNA 5' ends. When mouse and human genes both meeting the above criteria had representative isoforms with orthologous first exons, the upstream region of this exon was considered to represent a conserved SP. See Supplemental Table S1 for AP identification criteria.

The dataset of conserved APs and SPs is Supplemental Table S6. Supplemental Table S8 gives a subclassification (in percentages) of those promoters that failed to meet our criteria for conserved SP or conserved AP, and that consequently were excluded from further analyses.

Detection of recently duplicated first exons

Each first exon of conserved APs and SPs was compared to the surrounding ~200-kb repeat-masked genomic sequence using the Smith-Waterman algorithm (Smith and Waterman 1981), scoring +2 for a match, -4 for a mismatch, -6 for gap initiation, and -5 for gap extension. First exons with alignment score ≥200 (equivalent to 100 perfectly matching bases) were considered putative recent duplicates. The corresponding promoters differ in a number of important characteristics from other APs (Supplemental Fig. S3 and ref. (Zhang et al. 2004)) and are removed from our subsequent analyses.

Donor splice site scores

Donor splice site scores were computed as in our previous study (Baek and Green 2005), except that foreground frequencies were computed from a non-redundant set of first exon donor sites of the aligned reference cDNAs, and background frequencies from 2-kb genomic sequences flanking each donor site.

Identification and scoring of CpG islands

Candidate CpG islands were determined within each unmasked chromosome sequence by a modification of the high-scoring segment approach (Lander et al. 2001). Briefly, each C nucleotide followed by a G was assigned a score of +17, and all other nucleotides (including C's not followed by a G) were assigned scores of -1. The maximal scoring segment in the chromosome was found by dynamic programming; it was then masked, and the search repeated on the masked sequence. These steps were iterated until the score of the best segment found dropped below 100. The CpG island score for a promoter was then computed as the mouse-human average total number of CpG island bases lying within a 3-kb window centered on the transcription start site. A promoter was considered CpG-rich when the CpG island score is ≥50, and CpG-poor otherwise.

Gene ontology and tissue/development associations

We obtained gene ontology (GO) terms from the GO data base (Harris et al. 2004) and crosslinks of GO terms with RefSeq sequences from the NCBI Entrez Gene data base (Maglott et al. 2005), and then mapped GO terms to genomic coordinates using our reference cDNA alignments. If a mouse or human representative isoform had exons overlapping a mapped GO term, all promoters for the corresponding gene were considered to be associated with the GO term. For genes with multiple promoters a single promoter was chosen at random for use in the association analyses.

To detect associations, a 2×2 Fisher's exact contingency test was used with one column representing a particular promoter type (CpG-rich AP, CpG-poor AP, CpG-rich or CpG-poor SP) and the other column representing the other three promoter types combined, and with rows representing presence vs. absence of the searched term. Since most genes are linked to multiple GO terms, and a single set of overrepresented genes can therefore cause multiple terms to appear significant, we adopted the following iterative process to ensure that reported associations are independent of each other. We first identified the term having the highest enrichment value relative to the other three promoter types, and for which the Bonferroni-corrected $P < 0.05$. All genes or promoters associated with the given term were then removed, and the search repeated with the reduced set of promoters. Iteration continued until no Bonferroni-corrected *P*'s were less than 0.05.

To investigate tissue and development-stage associations, a similar approach was used except that we counted the number of cDNAs and ESTs aligned on the first exons of representative isoforms of conserved promoters. (Although such data give a rough measure of expression, for purposes of discriminating among different promoters for the same gene they are preferable to most current microarray datasets, which often lack probes in the informative exons.) We excluded ESTs in the opposite genomic orientation to the representative isoforms, whose 5' ends were more than 500 bases from 5' ends of all representative isoforms, or that lacked information on tissue type or developmental stage.

We obtained tissue type information for each EST from the UCSC Genome Bioinformatics Site and identified a subset of tissue types having at least 1000 filtered ESTs. Mixtures of multiple tissue types and disease tissues were removed, and related tissue types (e.g. brain and hypothalamus) were combined, resulting in 47 tissue clusters. Tissue types that included "cancer", "tumor", "metastasis", "carcino-", "-oma", and "-ima" in the name were combined into a single 'cancer-related' cluster.

Developmental stage terms were processed in a similar manner. We classified each developmental stage term as prenatal (including egg, embryo, and fetus) or postnatal (neonate to adult).

Putative housekeeping promoters

We obtained mouse microarray expression data from the Genomics Institute of the Novartis Research Foundation (Su et al. 2002), and genomic coordinates of the microarray probes from the UCSC Genome Bioinformatics Site. Probes not uniquely placed were disregarded, leaving 29,092 that were. Probes in which expression levels in 61 tissue types lay within a window $[(1-\alpha)m, (1+\alpha)m]$ where m is the median expression level of the probe and $\alpha = 0.9$, were considered putative housekeeping probes; 4,266 probes satisfied this criterion.

For genes with multiple promoters a single promoter was chosen at random for use in analyses. We considered a conserved predicted promoter to be a putative housekeeping promoter if the mouse representative isoform had at least 50 bases of exonic overlap with putative housekeeping probes. This resulted in 357 putative housekeeping promoters.

Primer design

Primers were picked using Consed's (Gordon et al. 1998) automated, command-line PCR primer picking function AUTOPCRAMPLIFY (also see Consed documentation at http://bozeman.mbt.washington.edu/consed/consed.html). AUTOPCRAMPLIFY performs additional checks to those of Consed's sequencing primer picker, and in particular avoids primer pairs that could amplify other locations in the target sequence or form primer dimers, or that have melting temperatures differing by more than 3 degrees.

In the GeneRacer protocol, two universal 5' primers match different locations within the GeneRacer RNA primer. AUTOPCRAMPLIFY was used to find a compatible gene-specific 3' PCR primer for each universal primer, with melting temperatures between 64°C and 66°C or between 68°C and 70°C depending on the universal primer.

PCR amplification

Platinum Taq (Catalog# 10966-034, Invitrogen) was used to PCR amplify the RACE-ready cDNA pools. The first round of amplification was performed using a 5'GeneRacer primer (5' cgactggagcacgaggacactga 3') paired with a gene-specific 3' primer. The PCR reactions consisted of .6 µM 5' primer, .2 µM 3' primer, .2 mM dNTPs, 1.5 mM $MgCl_2$ and .5 units of polymerase in a 10 µL reaction. The reactions were run on an MJ Research Tetrad Thermal Cycler using a hot start and touchdown PCR cycling—samples were heated for 2 minutes at 94°C prior to cycling as follows: 5 cycles of 94°C 30 seconds, 70°C 30 seconds and 72°C 60 seconds; 5 cycles of 94°C 30 seconds, 68°C 30 seconds and 72°C 60 seconds; and 25 cycles of 94°C 30 seconds, 65°C 30 seconds and 72°C 60 seconds.

The resulting PCR product was diluted 10-fold and used as template in a second round of PCR using a nested 5' GeneRacer primer (5'ggacactgacatggactgaaggagta 3') and a nested 3' gene-specific primer. The reaction conditions were the same as for the initial round of PCR except that the primers were both at a final concentration of .2 µM and the final annealing temperature was 66°C for 15 cycles.

Sequencing

PCR products were diluted 10-fold with distilled water prior to sequencing. Sequencing reactions were performed using ABI BigDye Terminator Kit, version 3.1 (Applied Biosystems) in 96-well plates. The standard protocol was modified to 1/20th reactions and the nested 3' gene-specific primers were used as sequencing primers. Sequencing reactions were ethanol precipitated to remove unincorporated dyes. Reactions were resuspended in 10 µL HI-DI formamide (Applied Biosystems) and run on an ABI Prism 3130XL Genetic Analyzer using POP7 polymer and a 50 cm capillary array.

References

Baek, D. and P. Green. 2005. Sequence conservation, relative isoform frequencies, and nonsense-mediated decay in evolutionarily conserved alternative splicing. *Proc Natl Acad Sci U S A* **102:** 12813-12818.

Gordon, D., C. Abajian, and P. Green. 1998. Consed: a graphical tool for sequence finishing. *Genome Res* **8:** 195-202.

Harris, M.A., J. Clark, A. Ireland, J. Lomax, M. Ashburner, R. Foulger, K. Eilbeck, S. Lewis, B. Marshall, C. Mungall et al. 2004. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* **32:** D258-261.

Imanishi, T. T. Itoh Y. Suzuki C. O'Donovan S. Fukuchi K.O. Koyanagi R.A. Barrero T. Tamura Y. Yamaguchi-Kabata M. Tanino et al. 2004. Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS Biol* **2:** e162.

Karolchik, D., R. Baertsch, M. Diekhans, T.S. Furey, A. Hinrichs, Y.T. Lu, K.M. Roskin, M. Schwartz, C.W. Sugnet, D.J. Thomas et al. 2003. The UCSC Genome Browser Database. *Nucleic Acids Res* **31:** 51-54.

Kent, W.J. 2002. BLAT--the BLAST-like alignment tool. *Genome Res* **12:** 656-664.

Lander, E.S. L.M. Linton B. Birren C. Nusbaum M.C. Zody J. Baldwin K. Devon K. Dewar M. Doyle W. FitzHugh et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921.

Maglott, D., J. Ostell, K.D. Pruitt, and T. Tatusova. 2005. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res* **33:** D54-58.

Okazaki, Y. M. Furuno T. Kasukawa J. Adachi H. Bono S. Kondo I. Nikaido N. Osato R. Saito H. Suzuki et al. 2002. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420:** 563-573.

Pruitt, K.D., K.S. Katz, H. Sicotte, and D.R. Maglott. 2000. Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. *Trends Genet* **16:** 44-47.

Schwartz, S., W.J. Kent, A. Smit, Z. Zhang, R. Baertsch, R.C. Hardison, D. Haussler, and W. Miller. 2003. Human-mouse alignments with BLASTZ. *Genome Res* **13:** 103-107.

Smith, T.F. and M.S. Waterman. 1981. Identification of common molecular subsequences. *J Mol Biol* **147:** 195-197.

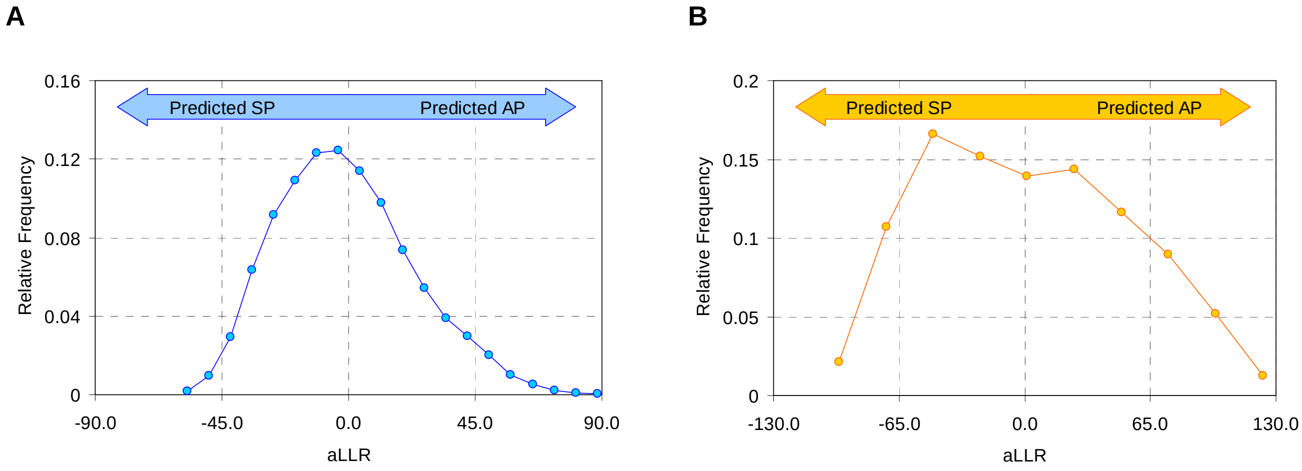
Su, A.I., M.P. Cooke, K.A. Ching, Y. Hakak, J.R. Walker, T. Wiltshire, A.P. Orth, R.G. Vega, L.M. Sapinoso, A. Moqrich et al. 2002. Large-scale analysis of the human and mouse transcriptomes. *Proc Natl Acad Sci U S A* **99:** 4465-4470.

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

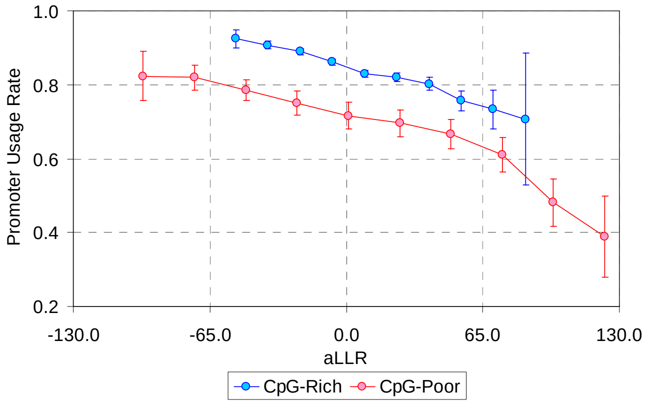
Waterston, R.H. K. Lindblad-Toh E. Birney J. Rogers J.F. Abril P. Agarwal R. Agarwala R. Ainscough M. Alexandersson P. An et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520-562.

Zhang, T., P. Haws, and Q. Wu. 2004. Multiple variable first exons: a mechanism for cell- and tissue-specific gene regulation. *Genome Res* **14:** 79-89.

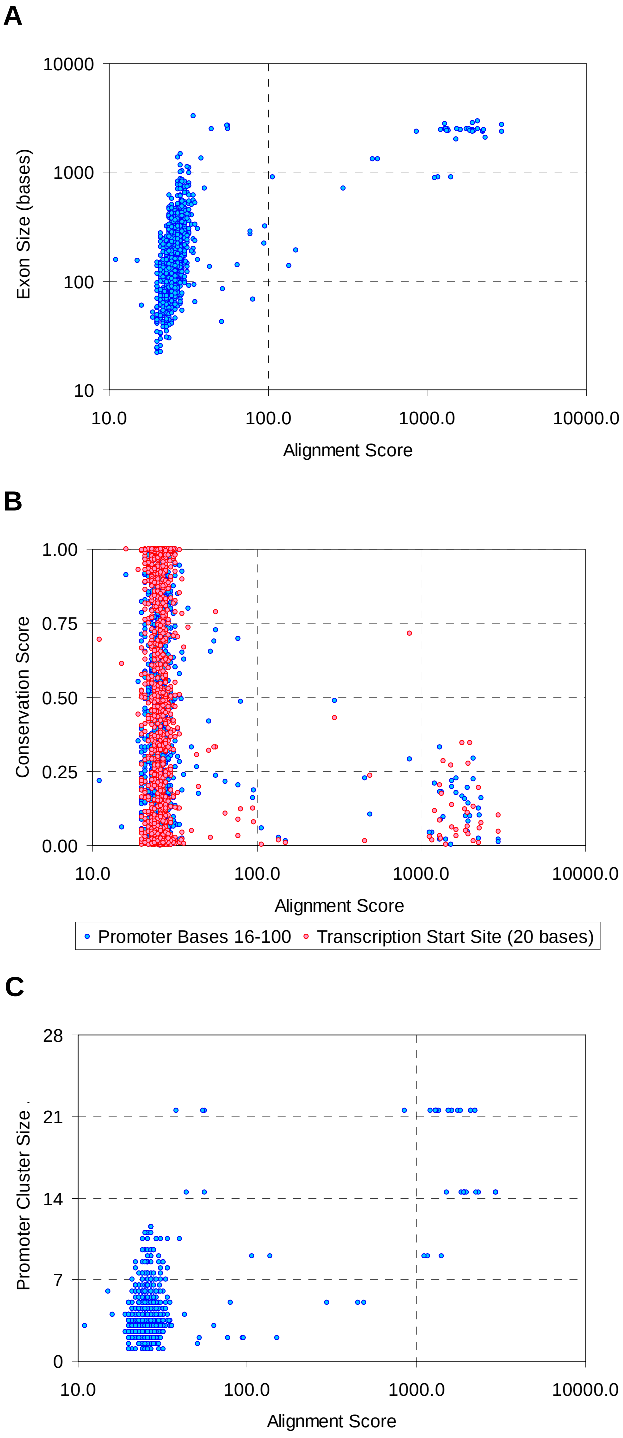
Supplemental Figures



Supplemental Figure S1. Histogram of approximate log-likelihood ratios in CpG-rich (A) and CpG-poor (B) conserved promoters.

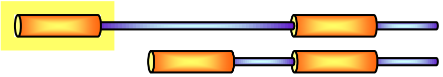
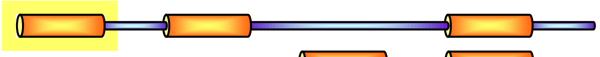
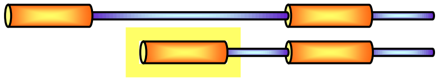
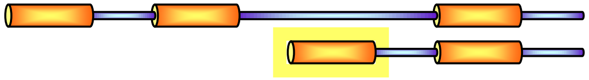


Supplemental Figure S2. Promoter usage rate as a function of approximate log-likelihood ratio in 12,025 conserved promoters.



Supplemental Figure S3. Scatter plots of alignment scores of the first exons in AP data set against flanking 200-kb regions, versus exon size (A), sequence conservation (B), and promoter cluster size (C). Promoters likely arising from exon duplication (those with high alignment scores) are atypical of other APs in having very large first exons, many mutually exclusive promoters per gene, and relatively diverged promoter regions.

Supplemental Tables

Group		Structure of cDNA/EST Isoform Pairs[*]	Num of Skipped Exons in Isoform 1[†]	Num of Skipped Exons in Isoform 2[†]	Evidence for Full-Length Status[‡]		Count				
					Isoform 1[†]	Isoform 2[†]	CpG-Rich	CpG-Poor	Combined	Duplicated First Exon Promoters[§]	Analysis Set (Genes)
AP	1		1	1	Required	Required	249	74	323	-	-
	2		a (a+b)≥3, a≥1, b≥1	b	Required	Required	125	33	158		
	3		1	1	Required	Not Required	374	249	623		
	4		a (a+b)≥3, a≥1, b≥1	b	Required	Not Required	113	155	268		
	Total (After Removing Redundancy)						707	373	1,080	35	1,045 (610)
SP	Total						2,509	600	3,109	29	3,080 (3,080)
All	Total						9,748	2,385	12,133	108	12,025 (11,035)

Supplemental Table S1. Criteria for identifying alternative promoters, and counts of promoter types. We only identified alternative promoters of 'mutually exclusive' type, i.e. the first exon of each representative transcript isoform is entirely absent from an isoform expressed from another promoter. Isoforms 1 and 2 were required to have identical first common exons (3'-most exons in the diagram), and to have 5' ends ≥500 bases apart. All exons shown in the diagram were required to be orthologous (reciprocal best match) between human and mouse. Note that more downstream promoters are detected than upstream promoters, a consequence of our evidence requirements for full-length status. [*]Identified first exons are highlighted in yellow. [†]A transcript that includes the highlighted exon is denoted as Isoform 1 and the other as Isoform 2. [‡]Reference cDNA or 5'-complete cDNA. [§]Promoters originating from recent duplication of first exons have atypical characteristics (Supplemental Fig. S3) and were excluded from subsequent analyses.

Promoter Type	Motif[*]	Enrichment[†]	P Value[†]	Num of Occurrences[†]	TRANSFAC Binding Factors
CpG-Rich SP	CCCGCCCM[‡]	2.1	2.5E-191	4514	SP1, KROX, GC BOX
	VCGGAAGA[‡]	2.5	1.6E-100	1721	NRF-2, GABP, STAT1, PEA3
	CACTTCCC[‡]	2.6	9.0E-81	1290	TEL-2, ETS, ELK-1, C-ETS-1, MAF
	ACCAATCA[‡]	2.9	1.3E-72	997	CCAAT, NF-Y
	CATTGGC[‡]	2.9	3.1E-68	925	NF-Y, ALPHA-CP1
	CCCCGC[‡]	1.6	4.1E-66	3476	SP1, GC BOX
	AGGCGGA[‡]	1.9	1.3E-57	1852	
	CTCCGC[‡]	1.7	2.7E-51	2101	
	ACGTCAC[‡]	3.0	6.9E-45	594	V-JUN, ATF, ATF3, HTF, ATF-1, E4F1, CREB, CRE-BP1:C-JUN, CRE-BP1, XBP-1, ATF6, CREBATF
	CCTGCGCA[‡]	1.6	1.9E-41	2011	HEN1
	GCGGGAV[‡]	1.7	1.7E-40	1759	BSAP, E2F-1:DP-1, E2F-4:DP-2
	CGGGGC[‡]	1.4	3.4E-30	2831	GC BOX, SP1, HIC1
	VAGGAAG[‡]	1.7	1.6E-29	1331	C-ETS-1, NERF1A, PU.1, TEL-2, ETS
	CGCCTCC[‡]	1.5	1.6E-28	1892	
	CTGGGA[‡]	1.6	7.6E-27	1488	IK-3, STAT, LUN-1
	ACGTGAC[‡]	2.5	3.2E-26	481	SREBP-1, ARNT, USF, STRA13
	RGGAGGA[‡]	1.5	1.7E-24	1594	MAF
	CGGAGCC[‡]	1.5	3.1E-23	1736	
	CCGGGA[‡]	1.5	3.7E-23	1451	STAT1, STATX, STAT3
	CCACGCCC[‡]	1.6	3.7E-23	1212	SREBP-1, AHR, PAX-9, PAX-4
	GCTTCC[‡]	1.5	9.1E-21	1420	NERF1A, C-ETS-1 68
	VGGAAAC[‡]	1.7	5.8E-20	864	PTF1-BETA, NF-AT, DEAF1
	VCGCGCAC[‡]	1.5	6.6E-19	1249	ZF5, NRF-1, GZF1
	ATTTCC[‡]	1.9	7.0E-19	590	NF-KAPPAB (P65), STAT3, ELK-1, C-ETS-2, STAT5B (HOMODIMER), ELF-1, HMG IY
	SCCCACCC[‡]	1.4	3.8E-17	1711	KROX, CAC-BINDING PROTEIN, UF1H3BETA
	CCCTCCCA	1.4	1.5E-15	1436	
	CAGCCAA[‡]	1.6	3.0E-15	912	ALPHA-CP1, NF-Y
	CACCGC[‡]	1.5	7.4E-14	1075	NRSF
	TTAAAA[‡]	1.8	2.2E-13	535	AMEF-2, MEF-2, NKX6-1
	CCTCCAGA[‡]	1.4	3.6E-13	1443	
	AATCAG[‡]	1.8	4.7E-13	482	AP-4, CCAAT, NF-Y, PBX1, GFI1
	GCATGCGC[‡]	1.7	5.2E-13	620	NRF-1
	ACCCGGA[‡]	1.4	3.4E-12	1400	STATX, STAT3, STAT1
	GTGGGA[‡]	1.5	1.7E-11	886	IK-1, IK-2, NKX25
	CCAGGAC[‡]	1.4	2.0E-11	1206	NERF1A
	ACGTGGC[‡]	1.7	8.8E-11	543	C-MYC:MAX, ATF6, N-MYC, USF
	TTCCCA	1.7	3.6E-10	532	STAF
	GCGTCC[‡]	1.4	2.1E-09	992	HEN1, WHN
	CTCCAC[‡]	1.4	2.6E-09	969	MUSCLE INITIATOR
	CCTTCC[‡]	1.3	4.0E-09	1488	NRF-2, GABP, BLIMP1
	AAGTGACA[‡]	1.6	7.1E-09	525	AP-1, LXR DIRECT REPEAT 4, LXR
	CCTGGAA[‡]	1.3	4.4E-08	1147	PAX6
	ACAGAAA[‡]	1.6	6.9E-08	526	
	CCGCAGC[‡]	1.3	1.2E-07	1556	HEN1
	ACAGGA[‡]	1.5	1.7E-07	634	ELK-1, C-ETS-1(P54), C-ETS-1, C-ETS-2
	CGGAACC[‡]	1.5	1.9E-07	660	BSAP, PAX-1
	GGTGAC[‡]	1.5	2.5E-07	622	AP-1, BACH2, ATF3, PPARG, SREBP
	CTTTAA[‡]	1.6	7.9E-07	446	MEIS1B:HOXA9, PLZF

CpG-Rich SP	CTGTCC‡	1.4	1.1E-06	859	NRSF
	ACCGGA‡	1.5	1.5E-06	586	NRF-2, E2
	AGTTCC‡	1.5	1.6E-06	621	STAT1, PAX, PAX6
	CGCCGC‡	1.2	1.7E-06	2126	
	AGGGCG‡	1.3	1.8E-06	1581	
	RAGAAAC‡	1.5	2.9E-06	588	PTF1-BETA, ISRE
	GCCGGA‡	1.3	3.1E-06	1197	C-ETS-1 P54
	GGGACAC‡	1.4	3.2E-06	781	
	AGAAAA‡	1.4	3.3E-06	768	PTF1-BETA
	CTTTCC‡	1.3	3.6E-06	1059	NF-KAPPAB, BLIMP1
	GGTCACA‡	1.5	3.7E-06	591	V-ERBA, PPARG
	AGTCCCA‡	1.3	1.0E-05	934	OLF-1, NF-KAPPAB
	AGGGAA‡	1.3	1.2E-05	914	BLIMP1, EBF
	TCCCCA	1.4	1.4E-05	602	OLF-1
	ACACAC‡	1.6	1.5E-05	406	POLY A
	AGGTGGA‡	1.4	1.7E-05	813	MUSCLE INITIATOR
	CGGGGA‡	1.3	2.0E-05	1341	MZF1, NF-KAPPAB, SP1
	CACAGCC‡	1.3	2.5E-05	938	AP-4
	CCCAGAA‡	1.3	3.6E-05	1193	OLF-1, STAF, STAT
	GCAGGA‡	1.3	4.3E-05	1083	C-ETS-1, C-ETS-2
	CAGGCGC‡	1.2	6.9E-05	1440	AP-2, PAX-4, USF2, MYOGENIN / NF-1
	AGAAAG‡	1.3	7.9E-05	848	BLIMP1
	AGGATG‡	1.5	1.8E-04	475	C-ETS-1(P54), PEA3
	ATCCTG‡	1.5	2.8E-04	423	C-ETS-1(P54)
	AGTTTC‡	1.4	2.9E-04	552	ISRE, ICSBP
	CGCAGA‡	1.3	4.3E-04	795	
	CGCGCG	1.3	4.4E-04	911	E2F, ZF5, E2F-1
	TCCGGA	1.5	5.1E-04	403	DEAF1
	CGCTCC‡	1.2	5.5E-04	1464	
	CTCCTC‡	1.2	7.0E-04	1556	
	AGCCAC‡	1.3	9.4E-04	835	
	AGGACA‡	1.4	1.2E-03	550	PR, GR, GR, T3R
	CGCGGA‡	1.3	1.5E-03	1008	E2F-1, NRSF
	GGAGAC‡	1.3	4.6E-03	871	
	GAAGTC‡	1.3	7.3E-03	591	ELK-1, NF-KAPPAB
	AAACAC‡	1.4	7.8E-03	408	FAC1, HFH-4
	AGAACT‡	1.4	9.7E-03	481	PPARALPHA:RXR-ALPHA
	CCCACA‡	1.3	9.8E-03	801	
CpG-Poor SP	SAGGAAGA‡	2.0	8.1E-19	563	C-ETS-1, NERF1A, PU.1, TEL-2, ETS
	TATAAAW	4.4	5.8E-15	150	TATA, XFD-2, MUSCLE TATA BOX, TBP
	CCTCCC‡	1.6	1.9E-11	619	MAZ, UF1H3BETA
	AGGAAA‡	1.8	2.2E-10	404	HELIOS A, NF-KAPPAB, STAT1, STAT6
	CCCAGCC‡	1.6	4.3E-09	514	LUN-1, CAC-BINDING PROTEIN
	CCACAGAG‡	1.8	3.0E-08	336	PEBP, SEF-1
	ATTTCC‡	1.9	1.3E-06	238	NF-KAPPAB (P65), STAT3, ELK-1, C-ETS-2, STAT5B (HOMODIMER), ELF-1, HMG IY
	CCCCGCCC‡	2.3	1.3E-06	172	SP1, KROX, GC BOX
	TAAATAA‡	2.0	3.3E-06	222	MEF-2, FREAC-3, AMEF-2, FOXP1

CpG-Poor SP	AAACAC‡	2.0	6.2E-06	216	FAC1, HFH-4
	GCTTCCC‡	1.7	3.2E-05	312	NERF1A, C-ETS-1 68
	CCCCAC‡	1.5	6.0E-05	407	SREBP-1, EGR, UF1H3BETA
	CAGAAAC‡	1.8	6.1E-05	249	PTF1-BETA, ISRE
	GGAGGA‡	1.5	8.8E-05	429	MAF
	AGGCAG‡	1.4	3.8E-04	476	
	CAGCCC‡	1.4	5.5E-04	464	CAC-BINDING PROTEIN
	CACAGC‡	1.5	3.4E-03	299	AP-4
	CCACCC‡	1.5	5.2E-03	376	MUSCLE INITIATOR, ZIC1, ZIC2, GLI
	AGATAA‡	1.8	7.0E-03	157	EVI-1, GATA-1, GATA-2, GATA-3, HFH-8
CpG-Rich AP	CCCTCCCC‡	2.0	5.8E-32	881	MAZ, UF1H3BETA
	AGGAGGAA‡	2.2	4.6E-22	490	MAF
	CCCGCCCC‡	1.6	5.0E-15	830	SP1, KROX, GC BOX
	CATTGGC‡	3.0	4.4E-12	174	NF-Y, ALPHA-CP1
	CTCCTC‡	1.7	1.6E-09	473	
	CACACAC‡	2.9	3.4E-08	134	POLY A
	CGTCACA‡	2.9	3.3E-07	120	PAX-3, E4F1, CRE-BP1
	CCCCTC‡	1.4	3.9E-07	672	UF1H3BETA
	TCCCCACC‡	1.6	4.1E-07	412	SREBP-1, EGR, UF1H3BETA
	CCACCCC‡	1.6	4.7E-07	397	MUSCLE INITIATOR, ZIC1, ZIC2, GLI
	CTGGGA‡	1.7	2.7E-06	314	IK-3, STAT, LUN-1
	AGGAAG‡	1.7	1.2E-05	282	C-ETS-1, NERF1A, PU.1, TEL-2, ETS
	CGCCGC‡	1.4	1.3E-05	596	
	CCAATC‡	2.1	2.0E-05	157	CCAAT, NF-Y
	AAAATA‡	2.8	4.2E-05	97	MEF-2, AMEF-2, RSRFC4, FOXJ2
	GGCGGA‡	1.5	2.0E-03	338	
	CCCCGC‡	1.3	3.0E-03	747	SP1, GC BOX
	AGAAAA‡	1.7	5.6E-03	179	PTF1-BETA
CpG-Poor AP	CCCTCCC‡	1.9	1.9E-06	255	MAZ, UF1H3BETA
	CCCACCC‡	2.2	1.1E-05	160	MUSCLE INITIATOR, ZIC1, ZIC2, GLI
	AAAATA‡	1.9	1.8E-04	202	MEF-2, AMEF-2, RSRFC4, FOXJ2
	AGGAAG‡	1.7	4.2E-03	210	C-ETS-1, NERF1A, PU.1, TEL-2, ETS
	GGAGGA‡	1.8	6.9E-03	174	MAF

Supplemental Table S2. Overrepresented hexamers (relative to simulated sequences having the same dinucleotide compositions). Only hexamers occurring in at least 10% of promoters of the given type are included. [*]Overlapping hexamers are merged (most statistically significant hexamer is underlined). [†]Values refer to the most significant (underlined) hexamer. [‡]No strand bias was detected, so occurrences of the hexamer and its complement were combined.

Promoter Type	P Value	Num of ESTs	Enrichment	Tissue Type
CpG-Rich SP	1.6E-65	1509	2.0	blood
	2.7E-87	3523	1.7	early embryonic cells
	2.4E-07	576	1.4	egg
	0.0E+00	85883	1.4	cancer-related
	1.7E-08	1093	1.3	thyroid
	7.6E-104	14421	1.3	stem cells
	9.0E-05	835	1.3	skin
	2.8E-46	11504	1.2	testis
	1.0E-02	1170	1.2	lymph
CpG-Poor SP	0.0E+00	1098	7.3	muscle
	0.0E+00	2004	6.7	pancreas
	7.1E-31	81	6.2	diaphragm
	3.1E-186	569	5.1	inner ear
	1.9E-12	63	3.4	salivary gland
	1.2E-279	1768	2.8	spleen
	1.9E-187	1942	2.1	eye
	9.8E-24	336	1.9	adrenal gland
	5.8E-12	179	1.9	tongue
	2.0E-27	490	1.8	colon
	3.1E-25	459	1.8	joint
	4.2E-41	1099	1.6	kidney
	5.2E-70	1962	1.6	placenta
	3.6E-09	438	1.4	intestine
	3.2E-17	919	1.4	lymphocyte
	1.1E-03	245	1.4	ascites
	6.4E-20	1146	1.4	nasopharynx
	8.1E-15	877	1.4	bone marrow
	1.7E-03	377	1.3	prostate
CpG-Rich AP	1.2E-50	682	2.8	cervix
	9.2E-35	950	1.9	ovary
	1.1E-190	8267	1.6	uterus
	0.0E+00	105002	1.3	brain
	7.4E-11	1449	1.3	umbilical cord
	9.7E-42	6577	1.3	lung
	1.5E-03	1230	1.2	bladder
	1.5E-08	9707	1.1	thymus
CpG-Poor AP	0.0E+00	4849	7.8	liver
	0.0E+00	1976	3.4	heart
	8.6E-279	1740	2.8	breast
	5.7E-11	174	1.9	amnion
	1.1E-14	302	1.7	blood vessel
	4.2E-30	682	1.7	stomach

Supplemental Table S3. Associations of tissue types with promoter types. For this analysis we used all conserved promoters that were strongly predicted by our discriminator to be AP or SP (having aLLRs in the top and bottom quartiles of the aLLR distribution, respectively).

CpG Island Class		Test Data Set Size	Predicted APs	Predicted SPs
CpG-Rich		9692	4375 (45%)	5317 (55%)
CpG-Poor		2333	1134 (49%)	1199 (51%)
Overall		12025	5509 (46%)	6516 (54%)
Overall (Genes)		11035	5084 (46%)*	6336 (57%)*
After correction†	Overall	12025	5206 (43%)	6819 (57%)
	Overall (Genes)	11035	4888 (44%)*	6532 (59%)*

Supplemental Table S4. Estimated genome-wide prevalence of AP and SP. *Failure to sum to 100% reflects a small fraction of genes (~3.5%) having multiple promoters with some (incorrectly) predicted to be SPs and some predicted to be APs. †The number of promoters was corrected to reflect the false positive/negative rate observed in our experimental tests.

Validation	CpG Island Class	Best Performing Model	Sample Size		Sensitivity	Specificity	(Sn+Sp)/2
			AP	SP			
Leave-one-out cross-validation (training set -- 80% of known promoters)	CpG-Rich	15 Parameters + Trimers	544	1998	0.73	0.78	0.76
	CpG-Poor	15 Parameters + Tetramers	292	465	0.75	0.70	0.73
Validation for test set (20% of known promoters)	CpG-Rich	15 Parameters + Trimers	136	500	0.72	0.70	0.71
	CpG-Poor	15 Parameters + Tetramers	73	117	0.74	0.62	0.68
Leave-one-out cross-validation (95% of known promoters)	CpG-Rich	15 Parameters + Pentamers	646	2373	0.69	0.80	0.75
	CpG-Poor	15 Parameters + Tetramers	346	552	0.78	0.68	0.73

Supplemental Table S5. Accuracy summary for leave-one-out cross-validation and test data.

(%)		Human		
		SP	AP	Unclassified
Mouse	SP	0.0	4.3	10.2
	AP	15.3	7.0	10.4
	Unclassified	28.3	6.5	17.9

Supplemental Table S8. Subclassification of promoters which failed to be classified as conserved APs or conserved SPs by our criteria. For some promoters, aligned ESTs/cDNAs do not meet the criteria either for APs or SPs; for example, cases in which mutually exclusive first exons of two 5'-complete transcripts have 5' ends <500 bases apart, or which do not have identical first common exons (3'-most exons in the diagram of Supplemental Table S1). APs for which two 5'-complete transcripts have overlapping first exons also fall into this category. We designate such cases as "unclassified". The center cell above corresponds to cases where both human and mouse promoters are classified as APs, but a mutually exclusive exon, or the first common exon, is not conserved between human and mouse.