Mice and Men: Their Promoter Properties

Vladimir B. Bajic[1,2*], Sin Lam Tan[1,2], Alan Christoffels[3,4], Christian Schönbach[5], Leonard Lipovich[6], Liang Yang[7], Oliver Hofmann[2], Adele Kruger[2], Winston Hide[2], Chikatoshi Kai[8], Jun Kawai[8,9], David A. Hume[10], Piero Carninci[8,9], Yoshihide Hayashizaki[8,9]

1 Knowledge Extraction Laboratory, Institute for Infocomm Research, Singapore, **2** South African National Bioinformatics Institute, University of the Western Cape, Bellville, South Africa, **3** Temasek Life Sciences Laboratory, National University of Singapore, Singapore, **4** School of Biological Sciences, Nanyang Technological University, Singapore, **5** Immunoinformatics Research Team, Advanced Genome Information Technology Research Group, RIKEN Genomic Sciences Center, RIKEN Yokohama Institute, Yokohama, Japan, **6** Genome Institute of Singapore, Singapore, **7** Department of Obstetrics and Gynecology, National University Hospital, National University of Singapore, Singapore, **8** Genome Exploration Research Group (Genome Network Project Core Group), RIKEN Genomic Sciences Center, RIKEN Yokohama Institute, Yokohama, Japan, **9** Genome Science Laboratory, Discovery Research Institute, RIKEN Wako Institute, Wako, Japan, **10** Australian Research Council Special Research Centre for Functional and Applied Genomics, Institute for Molecular Bioscience, University of Queensland, Brisbane, Queensland, Australia

Using the two largest collections of *Mus musculus* and *Homo sapiens* transcription start sites (TSSs) determined based on CAGE tags, ditags, full-length cDNAs, and other transcript data, we describe the compositional landscape surrounding TSSs with the aim of gaining better insight into the properties of mammalian promoters. We classified TSSs into four types based on compositional properties of regions immediately surrounding them. These properties highlighted distinctive features in the extended core promoters that helped us delineate boundaries of the transcription initiation domain space for both species. The TSS types were analyzed for associations with initiating dinucleotides, CpG islands, TATA boxes, and an extensive collection of statistically significant *cis*-elements in mouse and human. We found that different TSS types show preferences for different sets of initiating dinucleotides and *cis*-elements. Through Gene Ontology and eVOC categories and tissue expression libraries we linked TSS characteristics to expression. Moreover, we show a link of TSS characteristics to very specific genomic organization in an example of immune-response-related genes (GO:0006955). Our results shed light on the global properties of the two transcriptomes not revealed before and therefore provide the framework for better understanding of the transcriptional mechanisms in the two species, as well as a framework for development of new and more efficient promoter- and gene-finding tools.

Citation: Bajic VB, Tan SL, Christoffels A, Schönbach C, Lipovich L, et al. (2006) Mice and men: Their promoter properties. PLoS Genet 2(4): e54. DOI: 10.1371/journal.pgen.0020054

The Genome Network Project – FANTOM3 article collection

Introduction

The computational identification and functional analysis of mammalian promoters has, to date, been constrained by the relatively small datasets of experimentally confirmed transcription start sites (TSSs). For example, promoters within dbTSS were recently updated with the mapping of 195,446 FANTOM2 mouse full-length cDNA sequences to 6,875 RefSeq mouse genes [1,2]. Functional analyses of these mammalian promoters have been restricted to shared transcription factor binding sites (TFBSs) between human and mouse datasets [2]. Using the same collection of promoters contained in dbTSS, Aerts et al. embarked on a characterization of promoters by extending their study to *Drosophila melanogaster* and *Fugu rubripes* [3]. Further characterization of mammalian promoters is dependent on the availability of experimentally verified TSSs that would complement and extend existing datasets represented by the FANTOM collection, dbTSS, the H-Invitational database [4], and RefSeq. The latest effort of the FANTOM3 consortium [5] has provided the scientific community with the largest collection of transcriptome data for *Mus musculus* (mouse), and has complemented this with CAGE tags of *Homo sapiens* (human). Based on these data we provide a comprehensive comparative analysis of mouse and human promoters that results in a number of new insights that help us to better understand the transcriptional scenario in these two species.

GC properties are well-known global factors that influence promoter characteristics and gene expression [3,6–9]. In addition, GC characteristics influence important DNA properties such as the "bendability" and curvature of the DNA helix and consequently influence the interplay of DNA and chromatin, which impacts transcription. We set out to

Editors: Judith Blake (The Jackson Laboratory, US), John Hancock (MRC-Harwell, UK), Bill Pavan (NHGRI-NIH, US), and Lisa Stubbs (Lawrence Livermore National Laboratory, US), together with *PLoS Genetics* EIC Wayne Frankel (The Jackson Laboratory, US)

Received August 15, 2005; **Accepted** February 27, 2006; **Published** April 28, 2006

DOI: 10.1371/journal.pgen.0020054

Abbreviations: CGI, CpG island; GO, Gene Ontology; Inr, initiator; ORI, over-representation index; PE, promoter element; TF, transcription factor; TFBS, transcription factor binding site; TSS, transcription start site

* To whom correspondence should be addressed. E-mail: vlad@sanbi.ac.za

Synopsis

Tens of thousands of mammalian genes are expressed in various cells at different times, controlled mainly at the promoter level through the interaction of transcription factors with *cis*-elements. The authors analyzed properties of a large collection of experimental mouse *(Mus musculus)* and human *(Homo sapiens)* transcription start sites (TSSs). They defined four types of TSSs based on the compositional properties of surrounding regions and showed that (a) the regions surrounding TSSs are much richer in properties than previously thought, (b) the four TSSs types are associated with distinct groups of *cis*-elements and initiating dinucleotides, (c) the regions upstream of TSSs are distinctly different from the downstream ones in terms of the associated *cis*-elements, and (d) mouse and human TSS properties relative to CpG islands (CGIs) and TATA box elements suggest species-specific adaptation. The authors linked TSS characteristics to gene expression through categories defined by the Gene Ontology and eVOC classifications and tissue expression libraries. They provided examples of the preference of immune response genes for TSS types and specific genomic organization. Their results shed light on the fine compositional properties of TSSs in mammals and could lead to better design of promoter- and gene-finding tools, better annotation of promoters by *cis*-elements, and better regulatory network reconstructions. These areas represent some of the focal topics of bioinformatics and genomics research that are of interest to a wide range of life scientists.

characterize the regions immediately surrounding TSSs based on such compositional properties. Our determination of tentative TSS locations has been based on the use of CAGE tags [10] and ditags [11] enriched with additional independent pieces of evidence of transcript existence including 5′ expressed sequence tags, long 5′-SAGE, and the 5′ ends of fully sequenced cDNAs from full-length libraries.

In this study, we report several distinctive features in the extended core promoters that helped us delineate the boundaries of the transcription initiation domain space for both mouse and human, as well as delineate species-specific characteristics within that space. We describe the association of TSS types with the initiating dinucleotide, CGIs, TATA boxes, and an extensive collection of statistically significant *cis*-elements in mouse and human, and correlate TSS properties with expression data through comparison with Gene Ontology (GO) [12] and eVOC [13] categories, tissue expression libraries, and specific genome organization.

Table 1. Four TSS Types Defined Based on the GC Content Upstream and Downstream of the TSS

TSS Type	Upstream GC Content	Downstream GC Content
A	GC-rich	GC-rich
B	GC-rich	AT-rich
C	AT-rich	GC-rich
D	AT-rich	AT-rich

GC-rich means G + C > 50% in the considered region. AT-rich (i.e., GC-poor) means G + C ≤ 50% in the considered region. In our case, the upstream region is [−100, −1], and the downstream region is [+1, +100] relative to the TSS.
DOI: 10.1371/journal.pgen.0020054.t001

Results/Discussion

GC Content and TSS Types

We considered TSS properties based on the GC characteristics of the segments immediately upstream and downstream of experimentally estimated TSSs. We split TSSs into four distinct classes based on the GC content upstream and downstream of the TSS, as shown in Table 1 (see Materials and Methods). These four tentative TSS types have been used as a tool to investigate different promoter features in mouse and human. Two TSS types do not differ in GC richness between the upstream and downstream regions. They are GC-rich (GC-GC, type A) or AT-rich (AT-AT, type D) both upstream and downstream. The other two are GC-rich upstream and AT-rich downstream (GC-AT, type B) and, vice versa, AT-rich upstream and GC-rich downstream (AT-GC, type C). The distributions of TSS positions in the case of mouse and human are depicted in Figure 1. A strong polarization of the TSS distribution exists, with TSS types A and D being most prevalent (Figure 1A). The number of TSSs in each of the TSS types remains almost unchanged if the length of the upstream and downstream regions changes

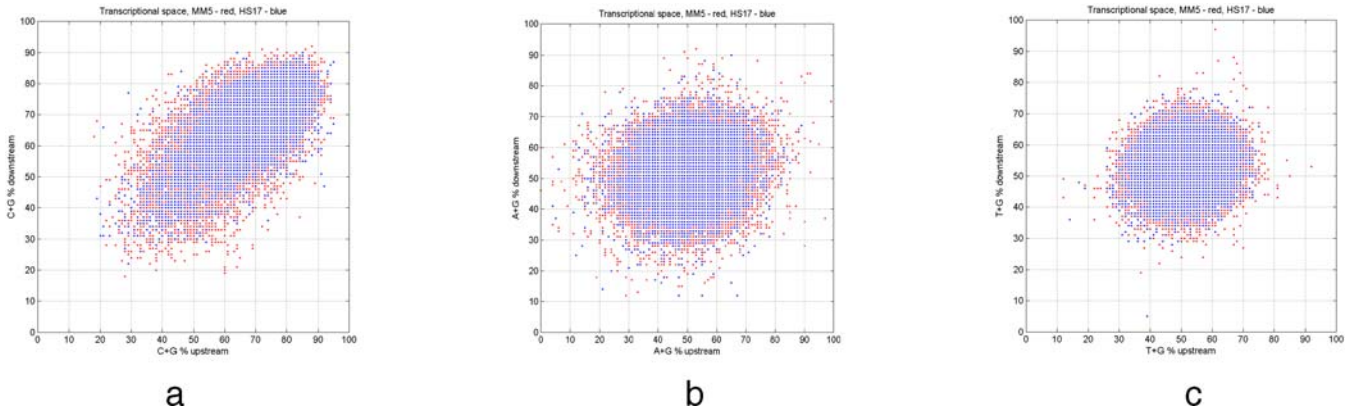


Figure 1. Transcription Initiation Domains for Mouse and Human

Distribution of mouse (red) TSSs overlapped by human (blue) TSSs based on (A) C + G content, (B) A + G content, and (C) T + G content. Nucleotide content is determined for upstream [−100, −1] and downstream [+1, +100] regions relative to the TSS. The distribution of TSS locations is more or less random when viewed in terms of A + G content (B) or T + G content (C). Strong polarization of distributions is evident only in the G + C case (A).
DOI: 10.1371/journal.pgen.0020054.g001

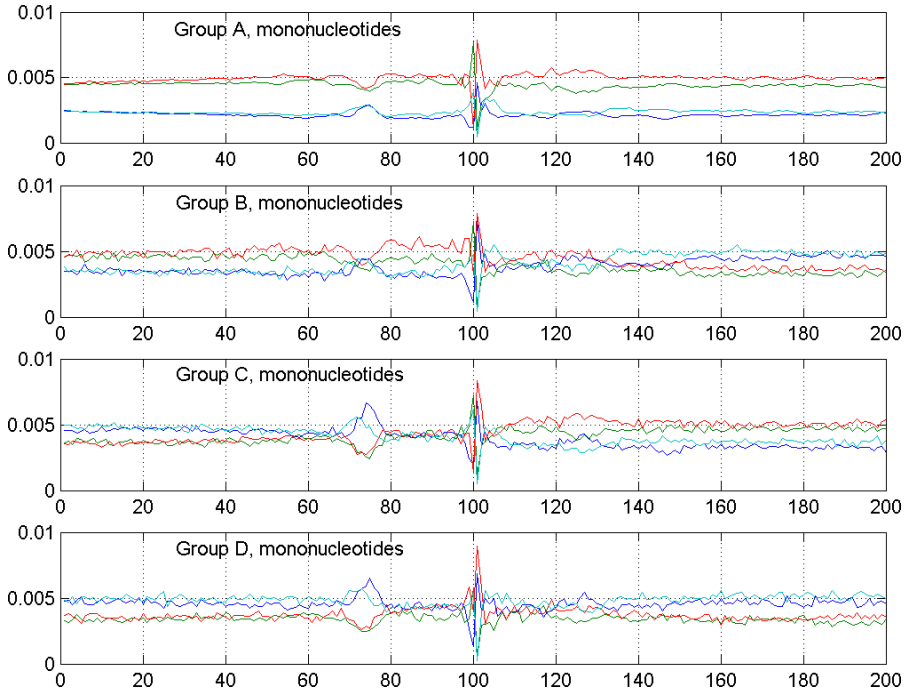


Figure 2. Distribution of Mononucleotides in Mouse Promoters in the Region Surrounding the TSS

The nucleotides adenine, cytosine, guanine, and thymine are represented by blue, green, red, and light blue, respectively. The TSS types that are GC-poor upstream (C and D) show very characteristic enrichment in adenine and thymine nucleotides around [–35, –20], suggesting a potential dominant influence of TATA box and similar AT-rich elements in transcription initiation in these types. In type B and A TSSs, this influence does not seem to be dominant, but the presence of such elements is suggested by a significant reduction of the GC content in the [–35, –20] region. In principle, one could attempt to link the types of AT-rich upstream elements with initiating dinucleotides characteristic of different TSS types.

DOI: 10.1371/journal.pgen.0020054.g002

(Figure S1); it also only gradually changes with a change of threshold for GC richness (Figure S1). These findings suggest robustness of our TSS classification.

Are Two TSS Types (GC-Rich and AT-Rich) Sufficient to Consider?

Promoters are usually classified as either GC-rich or AT-rich, without separating such properties into upstream and downstream characteristics relative to the TSS [3]. In our study we observed that many of the TSSs that are not evidently GC-rich (both upstream and downstream of the TSS) have changing GC content when going from upstream to downstream regions (Figure 2). The types of patterns were AT→GC, AT→AT, and GC→AT, containing 1,911, 1,528, and 1,440 instances, respectively, in our mouse TSS dataset. We found it reasonable to assign the TSSs with a change of GC content around the TSS (AT→GC and GC→AT) to different classes because they represent about 2/3 of all non-GC-GC types. We use this profiling of TSS characteristics as a methodological convenience. However, the biological justification for this relies on the fact that many *cis*-elements have a preference for GC-rich or AT-rich domains, as found in studies of promoter groups [14,15]. Thus, considering separately the GC-rich (AT-rich) upstream and downstream segments around TSSs provides an opportunity to analyze different groups of binding sites that may confer different transcription initiation scenarios.

An essential support for the biological relevance of our introduced TSS classification relies on the fact that some eukaryotic genomes have dominant TSS characteristics of the classes we defined. For example, based on the work of Aerts et al. [3], TSS types B and C appear prevalent in *F. rubripes* and type D in *D. melanogaster*, while type A is characteristic of the human genome. There are other ways to classify promoters using certain functional rather than compositional properties. Kadonaga [16] used the presence of functional core promoter elements (PEs) such as TATA boxes, initiators, and downstream promoter elements (DPEs) to classify promoters into several types. A different approach was used by Kim et al. [17]: the properties of preinitiation complex binding to promoter and the observed transcript expression state were used to define four promoter groups.

Figure 3. Distribution of Densities of Selected PEs in Promoters of the Four TSS Types in Mouse

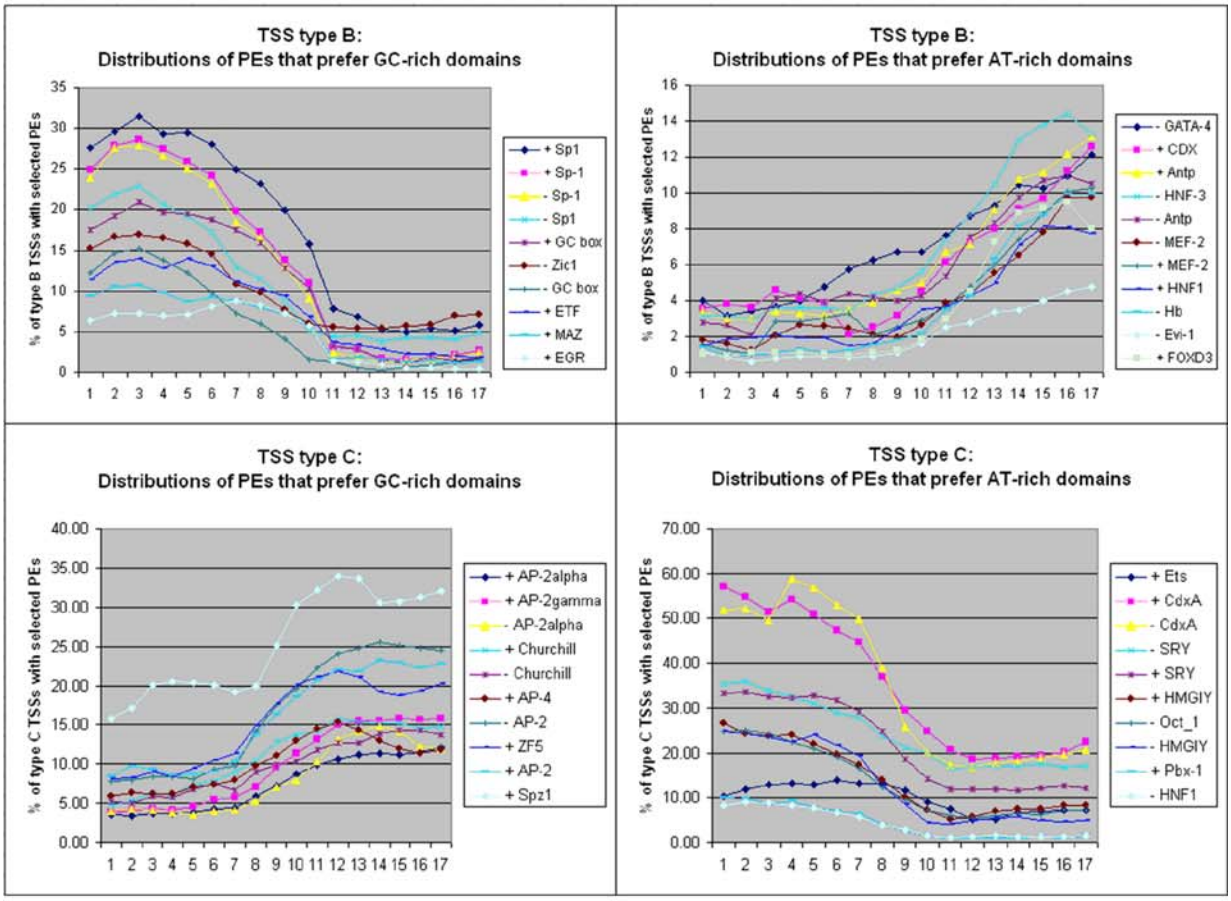
The density of PEs is calculated from the region covering [–100, +100] relative to the TSS. Density is determined for bins of length 50 bp and shifted by 10 bp. In total, there are 17 bins. The vertical axis shows the percentage of TSSs of the considered type that contain the PE.

(A) Distribution of selected PEs that prefer GC-rich (left) and AT-rich (right) domains in type B (above) and type C (below) TSS groups. Bin number 9 is centered around the TSS. It can be seen that groups of PEs change significantly in their concentrations in transition from upstream to downstream regions and characterize two distinct TSS types (B and C).

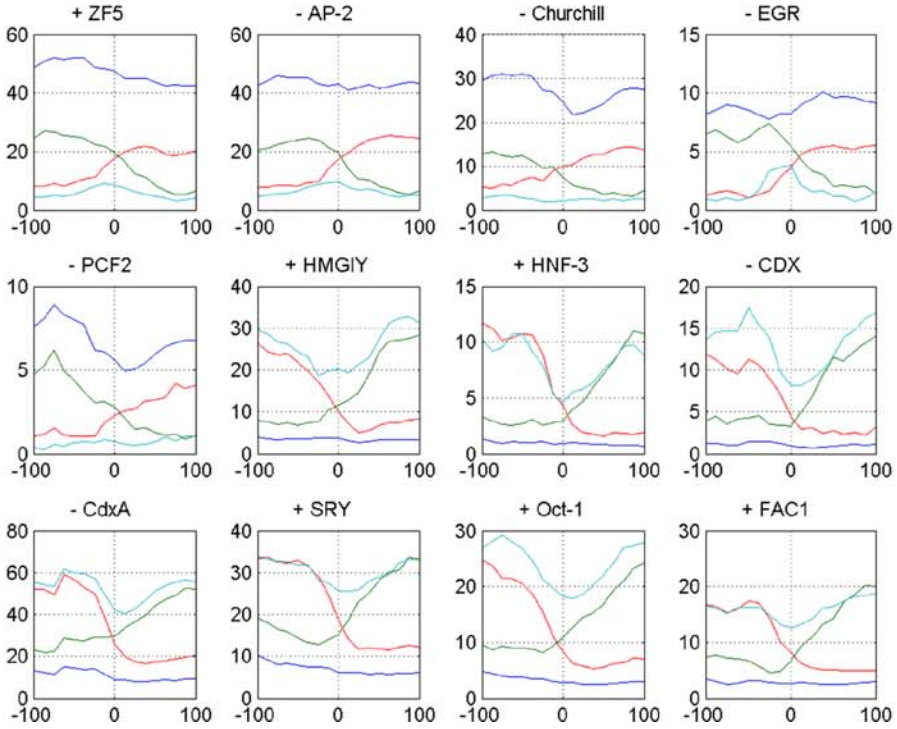
(B) Distribution of selected PEs across all four TSS types. Blue, green, red, and light blue correspond to distributions characterized by type A, B, C, and D TSSs. The first five PEs are those that prefer GC-rich regions, and the last seven PEs prefer AT-rich regions (the plus or minus sign in front of the TFBS symbol denotes the strand where the TFBS is found).

DOI: 10.1371/journal.pgen.0020054.g003

A


TSS type B:
Distributions of PEs that prefer GC-rich domains
% of type B TSSs with selected PEs
+ Sp1
+ Sp-1
- Sp-1
- Sp1
+ GC box
- Zic1
- GC box
+ ETF
+ MAZ
+ EGR
TSS type B:
Distributions of PEs that prefer AT-rich domains
% of type B TSSs with selected PEs
- GATA-4
+ CDX
+ Antp
- HNF-3
- Antp
- MEF-2
+ MEF-2
+ HNF1
- Hb
- Evi-1
+ FOXD3
TSS type C:
Distributions of PEs that prefer GC-rich domains
% of type C TSSs with selected PEs
+ AP-2alpha
+ AP-2gamma
- AP-2alpha
+ Churchill
- Churchill
+ AP-4
- AP-2
+ ZF5
+ AP-2
+ Spz1
TSS type C:
Distributions of PEs that prefer AT-rich domains
% of type C TSSs with selected PEs
+ Ets
+ CdxA
- CdxA
- SRY
+ SRY
+ HMGIY
- Oct_1
- HMGIY
+ Pbx-1
- HNF1

B


+ ZF5
- AP-2
- Churchill
- EGR
- PCF2
+ HMGIY
+ HNF-3
- CDX
- CdxA
+ SRY
+ Oct-1
+ FAC1

We found through several sources of evidence that expanding a crude classification of GC-rich and AT-rich TSSs by two additional subclasses makes biological sense and presents certain fine details more explicitly than is possible if all TSSs are lumped into only two (GC-rich and AT-rich) classes. Very obvious examples of such details, in addition to largely different compositions of the putative *cis*-elements that reside in the upstream and downstream regions, are (a) specialized, but different, initiating dinucleotides overrepresented in a statistically significant manner in TSSs of different types, (b) clear differences in the surrounding environment of the initiating dinucleotides between the four TSS types, and (c) different preferences of some functional gene groups for particular TSS types. These features cannot be observed if the groups are lumped.

GC Content of TSS Surroundings Reflects Types of Putative *cis*-Elements

By considering the GC content upstream and downstream separately, we allowed for one more degree of freedom in observing global TSS properties. Here we denote a PE as a TFBS and the strand where it is found. Many PEs have preferences for either GC-rich or GC-poor regions [14,15]. For example, the well-known TATA box element, being AT-rich, will be found more frequently in AT-rich regions, while the Sp1-binding sites, being GC-rich, will be found more frequently in GC-rich regions. Thus, the four TSS types that we consider could be correlated in a global manner with the potential PEs that may control the respective genes. Support for the influence of potential PEs on specific TSS types is obtained from the distributions of PE densities (Figure 3). Density distributions of selected PEs that prefer GC-rich (AT-rich) domains in type B and type C TSSs are depicted in Figure 3A. We observe that PE groups change their concentrations significantly in transition from upstream to downstream regions. Moreover, in Figure 3B we present distributions for selected PEs across all four TSS types. The first five PEs in Figure 3B (+ZF5, −AP-2, −Churchill, −EGR, and −PCF2) are those that prefer GC-rich regions (the plus and minus signs in front of the TFBS symbols denotes the strand where the TFBS is found). It is interesting to observe that these PEs occur in high concentrations in the type A group (GC-GC), occur in considerably lower concentrations in type D (AT-AT), and follow the change of GC content in types B and C. We observe the converse for the remaining seven PEs, which prefer AT-rich regions. These properties suggest that the four TSS types selectively associate with different groups of PEs.

Upstream and Downstream Regions Are Different: Enrichment by Specific PEs

We analyzed the preference of upstream and downstream regions in the four TSS types for significantly enriched (at least 3-fold) PEs in one region as opposed to the other region. The results are presented in Figure 4. To our surprise, we found that for all TSS types the number of enriched PEs in the upstream region is much higher than in the downstream region. In three types (A, C, and D) the number of PEs in the downstream region is minimal compared to the upstream region. The only exception is type B, for which there are a significant number of enriched PEs in the downstream region. The data suggest for type A TSSs a high influence of PEs that reside upstream and prefer GC-rich domains, while for type C TSSs such influence is likely through PEs that are located upstream of the TSS but prefer AT-rich domains. Contrary to these patterns, promoters with type B TSSs seem to utilize a mix of both GC-rich-preferring and AT-rich-preferring PEs. A conclusion cannot be made for type D TSSs because of the very small number of highly enriched elements overall. Moreover, applying the Chi-square test for the equality of distributions in the upstream and downstream regions we get $p = 1.34 \times 10^{-07}$, which strongly rejects the null hypothesis that these distributions are the same. All these findings suggest that upstream and downstream regions should be considered separately (as we do). The results emphasize enrichment of different PE groups associated with upstream and downstream regions in the promoters of the four TSS types.

Four TSS Types Associate with Different Sets of PEs

Different compositional properties of the four TSS types suggest that the TSSs may be controlled by specialized collections of transcription factors (TFs). Thus, we attempted to find the potential TFs that could play dominant roles in the four TSS types by identifying (a) the specificity of the top-ranked PEs (relative to overrepresentation index [ORI]; see Materials and Methods) in different TSS types, (b) unique and common motifs in the GC-rich/AT-rich upstream/downstream regions for different TSS types, and (c) the most significant PEs/TFs upstream/downstream of TSSs of types A, B, C, and D.

To carry out these analyses we initially compared the incidence of predicted DNA-binding sites of known TFs in the different promoter segments in mouse in the four TSS types against those in random mouse DNA. For the top 150 predicted motifs (representing approximately 10% of all elements found in these comparisons) determined based on ORI [15], we calculated Bonferroni corrected p-values for enrichment in the considered promoter segments. In the selection of these top 10% of motifs we required that they be present in at least 10% of the promoters in the target groups and that they have an ORI value not less than 1.5. In these comparisons we found that the corrected p-value was below the threshold of 0.05 for the great majority of cases. These comparisons indicate that most of the motifs for the considered TSS types are highly specific relative to random DNA (Table S1).

Next we aimed to see if promoter segments with the same GC richness share the same set of PEs. We compared the upstream regions of groups A versus B and C versus D, and the downstream regions of groups A versus C and groups B versus D. It is interesting to note that the upstream (GC-rich) regions of type A and B TSSs do share, as expected, a subset of predicted motifs, but each type is characterized also by a specialized collection of putative binding sites that do not appear in the top 150 ranked sites of the other type (for example, ETS appears in promoters of type B TSSs, but not in promoters of type A TSSs) (Table S2).

Even those TFs that are found to be common in the upstream parts of both type A and type B TSSs appear in significantly different proportions of promoters of these types, as summarized in Figure S2. For example, Ets (Table S1) appears in AT-rich downstream segments (types B and D). However, in type B TSSs it appears in 17.08% of promoters,

but only on the minus strand, while in type D it appears in 13.48% of promoters, but only on the plus strand.

Moreover, if we consider unique motifs that appear in different groups, they are commonly present in large proportions of promoters of those target groups. For example, in transcripts initiated from type D TSSs, we find only two unique PEs in the downstream region. One is DBP, a transcriptional activator in hepatic cells [18] and member of the C/EBP family, which appears in 26.77% of promoters with type D TSSs and only on the minus strand. The other element, Ncx, is enteric neuron homeobox and acts as an activator [19] that is required for proper positional specification, differentiative cell fate, and maintenance of proper function of enteric neurons [20,21]. It is present in 41.75% of promoters with type D TSSs and only on the plus strand.

Since any two of the four TSS types could differ in their GC content in the upstream, downstream, or both regions, and consequently harbor different sets of significant motifs, we conclude that, overall, TSS types contain sets of significant signature motifs (denoted by a plus sign next to the ORI value in Table S1 and a plus sign in Table S2) that potentially may contribute to orientation, and are likely to interact with distinct set of TFs. This concurs with the results of the preceding two subsections and suggests overall different transcriptional programs present in the transcripts of these TSS types. Lists of the most significant PEs that appear in the TSS groups are provided in Table S3.

The Initiating Dinucleotide and Its Environment

We analyzed in mouse and human datasets the initiating dinucleotide, that is, the one that occupies positions [−1, +1] relative to the TSS. We found that a number of different initiating dinucleotides are statistically significant across various TSS types and that they show certain regularities related to the GC content of upstream and downstream regions surrounding the TSS. Table 2 shows for mouse and human data all statistically significant cases based on the *p*-value obtained by the right-sided Fisher's exact test and corrected for multiplicity testing by the Bonferroni method. The association of initiating dinucleotide to TSS properties is very specific. It is interesting to note that the initiating dinucleotide TA is significantly enriched in TSS types that are AT-rich upstream, downstream, or both (B, C, and D), while dinucleotides that start with guanine (GA or GG) are significantly enriched in TSS types that are AT-rich specifically downstream (B and D). Type A TSSs are significantly enriched for dinucleotides that start with cytosine (CC, CG, and CT). However, the canonical initiating dinucleotide CA appears statistically significant only for TSS types that change GC richness (B and C). Finally, the TSS type C group contains AG and TG dinucleotides at a statistically significant level, while these do not appear significant in any other TSS type.

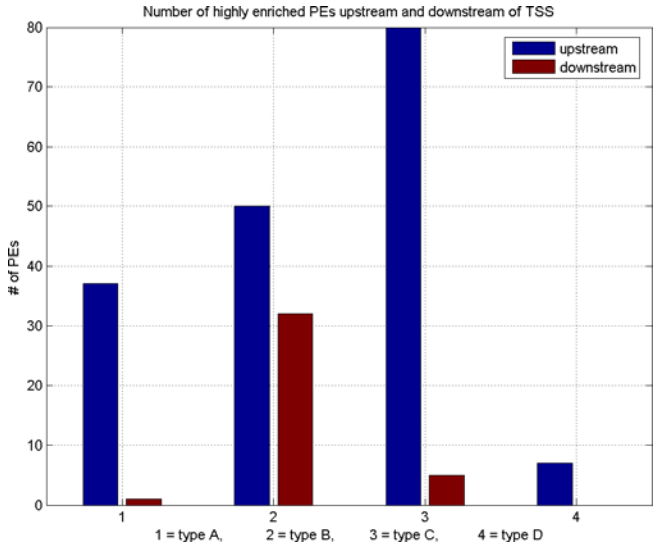


Figure 4. Distribution of Selected Groups of PEs That Are Highly Enriched (at Least 3-Fold) Upstream or Downstream of the TSS

The upstream region considered covers [−100, −1], while the downstream region covers [+1, +100] relative to the TSS. In all TSS types, the upstream region contains significantly more enriched PEs than the downstream region.

DOI: 10.1371/journal.pgen.0020054.g004

This compositional property of the initiating dinucleotide being linked in a statistically significant manner to the GC properties of the upstream and downstream regions would not be detectable if the TSS groups were lumped. We see that these properties characterize significant numbers of TSSs in our mouse dataset, namely, 10,547 (30.80%), 889 (61.74%), 1,372 (70.61%), and 534 (34.95%) of TSSs of type A, B, C, and D, respectively, and thus they do not appear to be artifacts of the proposed TSS classification that we have introduced. The conclusion is that the initiating dinucleotides show specific preferences at statistically significant levels to different TSS environments and that a significant portion of TSSs in our datasets are characterized by these initiating dinucleotides. Moreover, almost all of them are different from the canonical CA dinucleotide.

This last observation leads us to hypothesize that different TSS types may be controlled by different initiator (Inr) elements. Figure 2 depicts the quite different composition of the regions immediately surrounding tentative TSSs. The Inr elements—if they appear biologically relevant for these groups—should overlap TSSs and may be qualitatively different for different TSS types. Different initiating dinucleotides of highly statistically significant enrichment support such a hypothesis, and, at the same time, the variability of the observed initiating dinucleotides could explain the non-specific consensus of the octamer Inr element [22]. We have generated sequence logos of the TSS surroundings [−5, +5] in both mouse and human, and present them in Figure 5A. We observe that the nucleotide distributions for type A (GC-GC) TSSs are about the same in mouse and human. However, for TSS types B, C, and D, there is evident difference in these distributions in the region surrounding the TSS, which does not contradict our hypothesis of potentially different Inr elements for different TSS types. Figure 5B shows logos of regions [−35, +20] for the four TSS types in mouse and human. Again, we observe significant similarity between the species in the composition of the region for type A TSSs, while the other TSS types show significantly more variability. This may suggest species-specific organization of the core promoters for these minority TSS types (B, C, and D).

Relation of TSS Types to TATA Box Elements and CpG Islands

We analyzed the four TSS types in mouse and in human (Tables 3 and 4) for the presence of TATA box elements and

Table 2. Starting Dinucleotide [–1, +1] for Various TSS Types in Mouse and Human Datasets

Organism	Starting Dinucleotide	TSS Type	Number of Cases	Number of TSSs with Starting Dinucleotide	Total Number of TSSs in the Same TSS Group	Total Number of TSSs	Multiplicity Correction Factor	*p*-Value	Bonferroni Corrected *p*-Value
Mouse	AG	C	172	1,943	2,524	39,156	16	1.41×10^{-5}	2.25×10^{-4}
	CA	B	458	1,440	10,000	39,156	16	3.25×10^{-8}	5.20×10^{-7}
	CA	C	558	1,943	10,000	39,156	16	6.09×10^{-4}	9.75×10^{-3}
	CC	A	1,299	34,245	1,410	39,156	16	7.17×10^{-9}	1.15×10^{-7}
	CG	A	8,669	34,245	9,076	39,156	16	1.06×10^{-185}	1.69×10^{-184}
	CT	A	579	34,245	635	39,156	16	1.80×10^{-3}	2.88×10^{-2}
	GA	B	16	1,440	171	39,156	16	6.09×10^{-4}	9.75×10^{-3}
	GA	D	15	1,528	171	39,156	16	2.99×10^{-3}	4.79×10^{-2}
	GG	B	264	1,440	2,952	39,156	16	1.32×10^{-42}	2.12×10^{-41}
	GG	D	350	1,528	2,952	39,156	16	8.28×10^{-83}	1.33×10^{-81}
	TA	B	151	1,440	2,703	39,156	16	1.86×10^{-7}	2.97×10^{-6}
	TA	C	187	1,943	2,703	39,156	16	2.30×10^{-6}	3.68×10^{-5}
	TA	D	169	1,528	2,703	39,156	16	7.82×10^{-10}	1.25×10^{-8}
	TG	C	455	1,943	7,381	39,156	16	1.55×10^{-7}	2.48×10^{-6}
Human	AA	D	12	385	88	10,255	16	1.03×10^{-4}	1.65×10^{-3}
	CG	A	2,777	9,269	2,878	10,255	16	2.37×10^{-46}	3.79×10^{-45}
	GG	D	85	385	578	10,255	16	4.28×10^{-29}	6.85×10^{-28}
	TA	B	25	244	575	10,255	16	2.55×10^{-3}	4.07×10^{-2}
	TA	C	35	357	575	10,255	16	8.68×10^{-4}	1.39×10^{-2}

We show only statistically significant cases.
DOI: 10.1371/journal.pgen.0020054.t002

association with CGIs. Globally, there are similarities in these properties of TSS types between these two species, but there are also significant differences. This mouse–human comparison must be treated with some caution, since the mouse and human datasets are based upon analysis of distinct tissues, and the human set is probably less comprehensive. In some measure, the distinctions may also relate to depth of coverage in the two species. However, since we considered a statistically large number of well-defined TSS locations in mouse (39,156) and in human (10,255), this makes comparison between the two species feasible.

Based on Bonferroni corrected *p*-values we find that the mouse and human datasets differ significantly in a number of promoter features (Tables 3 and 4). Mouse promoters are significantly enriched in (a) the number of promoters not associated with CGIs in TSS types A and B, and overall; (b) the number of TATA-less promoters in group A; (c) the overall number of promoters that have TATA boxes but are not associated with CGIs; and (d) the number of TATA-less promoters not associated with CGIs in TSS groups A and B, and overall. Conversely, human promoters are significantly enriched in (a) the number of promoters associated with CGIs in TSS types A and B, and overall; (b) the number of TATA-box-containing promoters in TSS type A; (c) the number of TATA-box-containing promoters associated with CGIs in TSS types A, B, and C, and overall; and (d) the number of TATA-less promoters associated with CGIs in TSS types A and B, and overall. These data suggest that there are species-specific solutions for transcriptional initiation in mouse and human for the analyzed TSS types.

There are a number of core PEs other than TATA boxes and Inr elements, such as the downstream promoter element (DPE) [23–26], the TFIIB response element (BRE) [27], the motif ten element (MTE) [28], and the downstream core element (DCE) [29,30]. It would be of interest to investigate their presence around mammalian TSSs. Unfortunately, such an analysis represents a study on its own and requires reliable matrix models of these elements in mammals that are not yet available.

Linking TSS Properties and Gene Expression

We were interested to find out if the TSS types show any correlation with broad expression categories. We used association of transcripts with different GO [12] and eVOC [13] categories, as well as FANTOM3 tissue expression libraries, and analyzed their TSS distribution across the four types in mouse. While it is not possible to make definite conclusions because of incomplete GO, eVOC, tissue library, and transcript data, we were able to find a number of classes that associate with specific TSS types in a statistically significant manner (Tables 5, 6, S4, and S5). Moreover, we searched for ortholog transcript groups in mouse and human whose promoters preserve enrichment in specific TSS types in both species (Table S4). Under the conditions of our study we found that 100% of GO categories whose mapped transcripts emanate from type B TSSs preserve their enrichment; this is true for 64% of GO categories associated with type C TSSs and for 80% of GO categories associated with type D TSSs. These results suggest that between mouse and human the TSS character within the GO categories is largely conserved. Distributions of all mouse TSSs across the four TSS types for GO categories and FANTOM3 tissue libraries are provided in Table S5.

We further analyzed several specific cases. For many GO categories we found that transcripts associated with them prefer specific GC-rich/GC-poor transcription initiation frameworks (Table 5). For example, the immune response group (GO:0006955) (Figure 6) appears with 1.58-, 4.85-, and

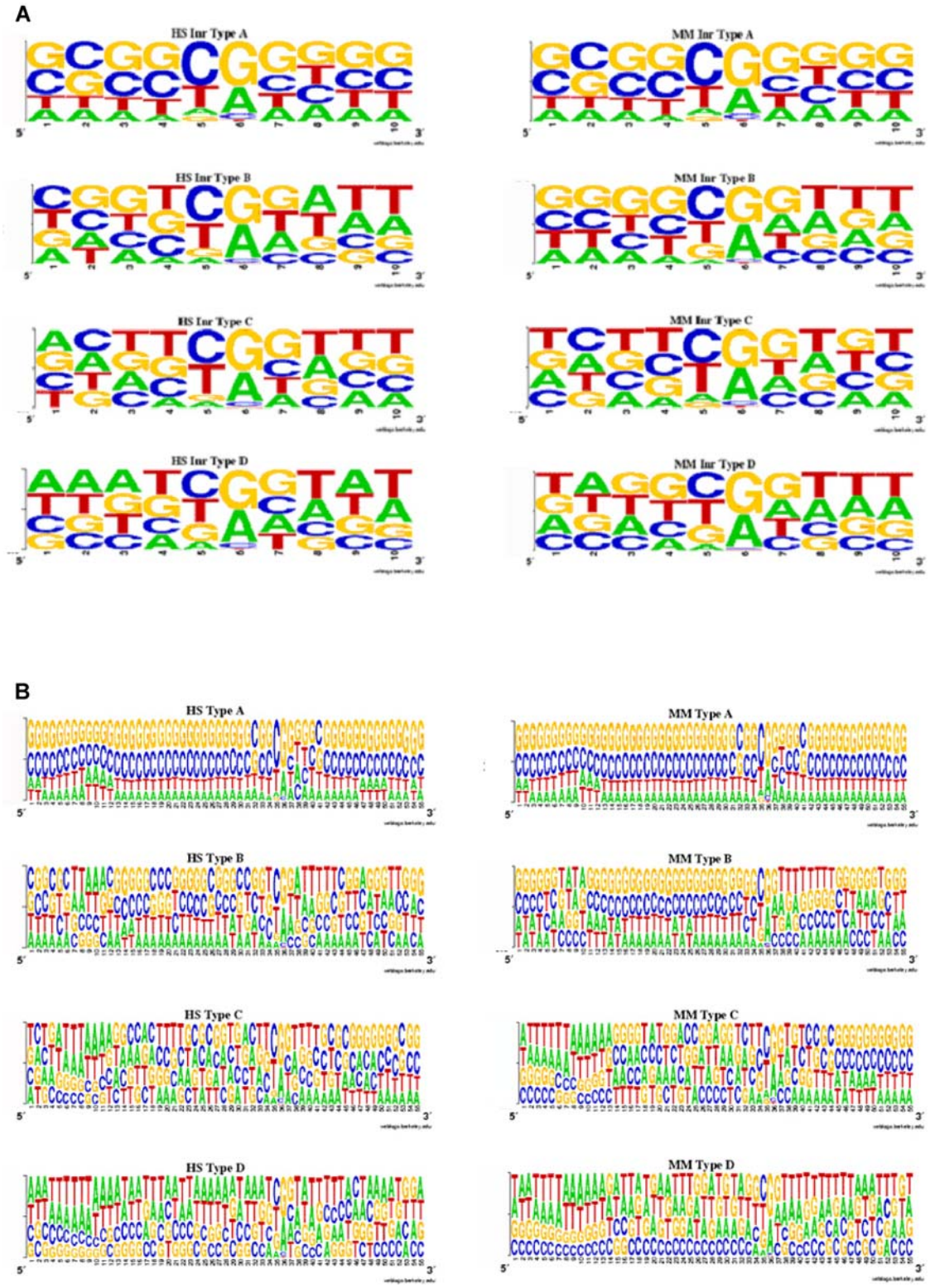


Figure 5. Sequence Logos

(A) Sequence logos for Inr in human (left) and mouse (right) obtained using [–5, +5] segments relative to TSS locations. There is an evident bias in the nucleotide composition surrounding the TSS that effectively determines different Inr elements.

(B) Sequence logos for segments [–35, +20] relative to TSS locations. Strong similarity exists between human (left) and mouse (right) in TSS type A, while that similarity is considerably reduced for the other TSS types.

DOI: 10.1371/journal.pgen.0020054.g005

Table 3. Basic Statistics on Relation of TATA Box Motifs, CGIs, and Four TSS Types for MM5 Transcripts

Category	TSS Type				
	Type A	Type B	Type C	Type D	Overall
Number of promoters	34,245	1,440	1,943	1,528	39,156
CGI	27,026 (78.92) [1]	253 (17.57) [1]	363 (18.68) [1]	9 (0.59) [1]	27,651 (70.62) [1]
No CGI	7,219 (21.08) [2.74×10^{-41}]	1,187 (82.43) [4.87×10^{-5}]	1,580 (81.32) [9.58×10^{-2}]	1,519 (99.41) [8.82×10^{-2}]	11,505 (29.38) [6.26×10^{-59}]
TATA	2,539 (7.41) [1]	188 (13.06) [1]	567 (29.18) [1]	434 (28.40) [1]	3,728 (9.52) [1]
TATA-less	31,706 (92.59) [1.63×10^{-3}]	1,252 (86.94) [1.43×10^{-1}]	1,376 (70.82) [1]	1,094 (71.60) [1]	35,428 (90.48) [2.02×10^{-1}]
CGI + TATA	1,613 (4.71) [1]	33 (2.29) [1]	58 (2.99) [1]	1 (0.07) [1]	1,705 (4.35) [1]
CGI + TATA-less	25,413 (74.21) [1]	220 (15.28) [1]	305 (15.70) [1]	8 (0.52) [1]	25,946 (66.26) [1]
No CGI + TATA	926 (2.70) [2.19×10^{-1}]	155 (10.76) [1]	509 (26.20) [1]	433 (28.34) [1]	2,023 (5.17) [2.09×10^{-4}]
No CGI + TATA-less	6,293 (18.38) [3.72×10^{-41}]	1,032 (71.67) [1.12×10^{-4}]	1,071 (55.12) [1]	1,086 (71.07) [1]	9,482 (24.22) [2.11×10^{-52}]

We present for each category (CGI, no CGI, etc.) the number of cases for each TSS type, the percent (in parentheses) of the total population in that TSS type, and the Bonferroni corrected *p*-value (in brackets) calculated from a right-sided Fisher's exact test based on the hypergeometric distribution.

DOI: 10.1371/journal.pgen.0020054.t003

3.35-fold more transcripts having TSSs of type B, C, and D, respectively, than one would expect based on the proportion of transcripts in these groups in our reference mouse data. The enrichment in type C and D TSSs is statistically significant (Bonferroni-corrected right-sided Fisher's exact test, $p = 1.33 \times 10^{-18}$ and $p = 2.60 \times 10^{-4}$, respectively). Based on this, we conclude that the transcript group GO:0006955 is characterized by increased participation of transcripts from TSS types that are AT-rich upstream or downstream. We analyzed in more detail the genomic organization of loci corresponding to genes from the most overrepresented TSS type (type C) for this GO. We found that TSSs of type C map to 36 nonredundant genes, of which two are in bidirectional promoters (2/36), which means these are underrepresented for type C TSSs relative to the genome average. There are 23 genes (64%) that are appearing in gene family clusters, that is, these genes are highly overrepresented for type C TSSs relative to the genome average. Finally, genes with type C TSSs have small genomic span: 34 out of 36 are less than 25 kb long, which is again more than one would expect based on the genome average. Most genes in the category GO:0006955 are short (the majority are actually less than 10 kb), are clustered with other members of the same families, and are not bidirectionally transcribed. This analysis illustrates a specific genomic organization of genes with TSSs of type C in this GO group. Thus, TSS properties may be associated with genomic organization.

In Table 5, one can see that GC-rich TSSs relate to genes responsible for various binding and protein transport activities. These functions usually occur in different regions of the cell and are reflected in the diverse compartments that are enriched for type A TSSs. AT-rich TSSs (types C and D), on the other hand, are enriched in processes relating to defense responses to the environment. TSSs of the membrane attack complex (GO:0005579), defense response (GO:0006952), and immune response (GO:0006955) are enriched in type D TSSs, while the last two of these (defense and immune response) and cytocine activity (GO:0005125) are enriched in type C TSSs. Globin group (GO:0001524) and hemoglobin complex (GO:0005833) are enriched in type B TSSs. These findings suggest a preference of different functional transcript groups for specific TSS types.

Similarly, for transcript groups based on eVOC terms, we find that they prefer GC-rich or GC-poor transcription initiation frameworks, depending on the eVOC category. For example, thymus-expressed transcripts (EVM:2270063 and EVM:2280063) (Table 6) seem to prefer either type A or D TSSs. The same is the case for transcripts classified according

Table 4. Basic Statistics on Relation of TATA Box Motifs, CGIs, and Four TSS Types for HS17 Transcripts

Category	TSS Type				
	Type A	Type B	Type C	Type D	Overall
Number of promoters	9,269	244	357	385	10,255
CGI	7,887 (85.09) [2.74×10^{-41}]	74 (30.33) [4.87×10^{-5}]	86 (24.09) [9.58×10^{-2}]	8 (2.08) [8.82×10^{-2}]	8,055 (78.55) [6.26×10^{-59}]
No CGI	1,382 (14.91) [1]	170 (69.67) [1]	271 (75.91) [1]	377 (97.92) [1]	2,200 (21.45) [1]
TATA	791 (8.53) [1.63×10^{-3}]	45 (18.44) [1.43×10^{-1}]	106 (29.69) [1]	101 (26.23) [1]	1,043 (10.17) [2.02×10^{-1}]
TATA-less	8,478 (91.47) [1]	199 (81.56) [1]	251 (70.31) [1]	284 (73.77) [1]	9,212 (89.83) [1]
CGI + TATA	574 (6.19) [7.00×10^{-8}]	16 (6.56) [7.01×10^{-3}]	22 (6.16) [2.99×10^{-2}]	0 (0.00) [1]	612 (5.97) [1.05×10^{-10}]
CGI + TATA-less	7,313 (78.90) [2.62×10^{-20}]	58 (23.77) [7.80×10^{-3}]	64 (17.93) [1]	8 (2.08) [5.64×10^{-2}]	7,443 (72.58) [4.31×10^{-34}]
No CGI + TATA	217 (2.34) [1]	29 (11.89) [1]	84 (23.53) [1]	101 (26.23) [1]	431 (4.20) [1]
No CGI + TATA-less	1,165 (12.57) [1]	141 (57.79) [1]	187 (52.38) [1]	276 (71.69) [1]	1,769 (17.25) [1]

We present for each category (CGI, no CGI, etc.) the number of cases for each TSS type, the percent (in parentheses) of the total population in that TSS type, and the Bonferroni corrected *p*-value (in brackets) calculated from a right-sided Fisher's exact test based on the hypergeometric distribution.

DOI: 10.1371/journal.pgen.0020054.t004

Table 5. Enrichment of TSS Types in Selected GO Categories in Mouse

GO Category	GO ID	Term	Bonferroni Corrected *p*-Values for the TSS Types			
			A	B	C	D
Cellular component	GO:0005833	Hemoglobin complex	1	1.74×10^{-6}	1	1
	GO:0005579	Membrane attack complex	1	1	1	1.24×10^{-5}
	GO:0005576	Extracellular region	1	1	4.79×10^{-2}	2.09×10^{-6}
	GO:0005794	Golgi apparatus	2.84×10^{-12}	1	1	1
	GO:0005634	Nucleus	6.15×10^{-12}	1	1	1
	GO:0005737	Cytoplasm	3.25×10^{-4}	1	1	1
	GO:0005739	Mitochondrion	1.23×10^{-9}	1	1	1
	GO:0005829	Cytosol	2.28×10^{-2}	1	1	1
Molecular function	GO:0001524	Globin	1	1.74×10^{-6}	1	1
	GO:0005125	Cytokine activity	1	1	1.98×10^{-7}	1
	GO:0003677	DNA binding	1.63×10^{-2}	1	1	1
	GO:0003723	RNA binding	3.38×10^{-2}	1	1	1
	GO:0003925	Small monomeric GTPase activity	1.39×10^{-4}	1	1	1
	GO:0005524	ATP binding	4.48×10^{-7}	1	1	1
	GO:0005525	GTP binding	1.62×10^{-4}	1	1	1
	GO:0008565	Protein transporter activity	2.11×10^{-7}	1	1	1
	GO:0016301	Kinase activity	6.82×10^{-5}	1	1	1
	GO:0016740	Transferase activity	3.19×10^{-4}	1	1	1
Biological process	GO:0006935	Chemotaxis	1	1	1.32×10^{-3}	1.36×10^{-2}
	GO:0006952	Defense response	1	1	3.12×10^{-8}	5.11×10^{-2}
	GO:0006955	Immune response	1	1	1.33×10^{-18}	2.60×10^{-4}
	GO:0006886	Intracellular protein transport	1.77×10^{-12}	1	1	1
	GO:0007049	Cell cycle	3.66×10^{-3}	1	1	1
	GO:0007264	Small GTPase-mediated signal transduction	2.76×10^{-4}	1	1	1
	GO:0015031	Protein transport	3.36×10^{-8}	1	1	1

The table shows some statistically significant examples of biased distribution of transcripts from different GO categories in specific TSS groups from all mouse data.
DOI: 10.1371/journal.pgen.0020054.t005

to cardiovascular function (EVM:2280037 and EVM:2250045) (Table 6).

Conclusions

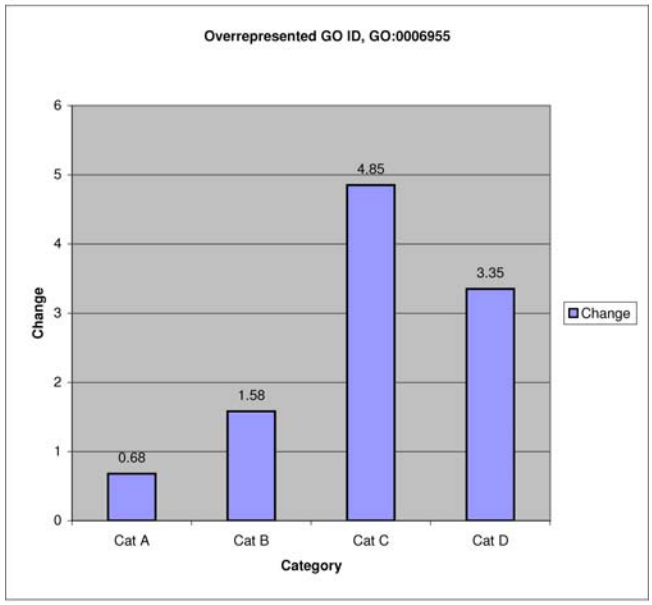
We have introduced a different way to characterize TSSs, which connects TSS properties to the GC content of the immediately upstream and downstream regions. This implicitly links the TSS type with PEs that are residing in the TSS neighborhood. We were able to delineate transcription initiation active domains in the mouse and human genomes and observed fundamental similarities in the transcription initiation active domains in the two species. Looking separately at the GC content upstream and downstream of TSSs provides a useful paradigm to view certain phenomena in a clearer and more meaningful manner. We found that two of the TSS types, types C and D, possess positionally very well defined AT-rich regions [–35, –20] relative to the TSS, suggesting the significant role of AT-rich sequences such as TATA boxes in the control of TSSs of these types. Our analysis documents that various initiating dinucleotides show

Table 6. Enrichment of TSS Types in Selected eVOC Categories and Tissue Libraries in Mouse

EVOC ID or Tissue Library	Terms	Bonferroni Corrected *p*-Value for the TSS Types			
		A	B	C	D
EVM:2280168	Lung, male, adult	2.22×10^{-2}	1	1	1
EVM:2120010	Whole body, mixture, embryo	3.05×10^{-2}	1	1	1
EVM:2270063	Thymus, mixture, embryo	1	1	1	7.51×10^{-3}
EVM:2280037	Aorta and vein, male, adult	1	1	1	4.98×10^{-11}
EVM:2280087	Cortex, mixture, embryo	1	4.56×10^{-2}	1	1
EVM:2280063	Thymus, mixture	9.28×10^{-4}	1	1	1
EVM:2250045	Heart, mixture, embryo	3.43×10^{-3}	1	1	1
I1	Blastocyst	1.50×10^{-4}	1	1	1
I4	Osteoclast-like cell	4.99×10^{-2}	1	1	1
I8	LPS-treated bone marrow, macrophage	4.85×10^{-3}	1	1	1
24	ES cell	8.20×10^{-4}	1	1	1
C7	Liver, tumor, adult	1	1	3.13×10^{-2}	2.51×10^{-16}

Some examples of statistically significant enrichment of different TSSs types in eVOC categories and tissue libraries from all mouse data.
DOI: 10.1371/journal.pgen.0020054.t006

Type	Total	Number in category	Percent	Change
Cat A	253	152	60.08	0.68
Cat B	253	14	5.53	1.58
Cat C	253	57	22.53	4.85
Cat D	253	30	11.86	3.35



GO ID Description	GO:0006955
biological_process	immune response

Figure 6. Distribution of TSSs for Transcripts Related to Immune Response through GO:0006955

There are 1.58-, 4.85-, and 3.35-fold more transcripts having TSS types B, C, and D than one would expect based on the proportion of transcripts in these groups in our reference mouse data. Enrichment is statistically significant for types C and D based on Bonferroni corrected *p*-values obtained by the right-sided Fisher's exact test (Table 5).

DOI: 10.1371/journal.pgen.0020054.g006

very specific preferences for the TSS types we considered, are present in statistically significant proportions of the TSSs in our datasets, and are almost all different from the consensus dinucleotide. Very specific sets of initiating dinucleotides are associated with different TSS types, and surrounding GC content is well correlated with the types of these dinucleotides. This suggests the potential presence of different Inr elements that may be characteristic for each of the TSS types and associated with different nucleotide characteristics of the surrounding domain.

We have shown that different TSS types associate with different PEs, that regions upstream and downstream of different TSS types are characterized by different collections of PEs, and that the putative PE content (for the top 10% of PEs) of the TSS surroundings generally differs for the TSS types. All these findings suggest likely control of the respective transcripts by different collections of significant PEs residing upstream or downstream of the TSS. Our results on TSS properties relative to CGIs, TATA boxes, and Inr elements in mouse and human suggest species-specific adaptation. Finally, we have shown a number of examples of transcript groups obtained on the basis of different ontologies or tissue libraries that have statistically significant enrichment in at least one of the TSS types. This has provided a link between TSS characteristics and expression data.

We believe that the results of this analysis will help in better understanding the general transcription regulation properties of mammalian promoters, and prove useful for further development and enhancement of promoter and gene prediction tools.

Materials and Methods

TSSs. We constructed two highly accurate sets (one for mouse and one for human) of TSSs and of the promoter sequences covering the span [−100, +100] relative to these TSSs. These datasets are available at http://www.sanbi.ac.za and were obtained as follows. If the first 5′ nucleotide of the CAGE tag or 5′ ditag (http://fantom31p.gsc.riken.jp/cage__analysis/export) coincided with the first 5′ nucleotide of the full-length cDNA (http://fantom.gsc.riken.go.jp/download.html), the TSS determined by this tag was selected. Also, in cases when this condition did not hold, we selected TSSs based on the following requirements: the TSS is a representative TSS location from a tag cluster that has at least ten tags, the representative TSS is supported by at least six tags, and there is at least one other piece of transcriptional evidence associated with this tag cluster (expressed sequence tag, full-length cDNA, or long SAGE; http://fantom.gsc.riken.go.jp/download.html). In this way, we compiled a mouse reference promoter set of 39,156 promoters and a human reference promoter set of 10,255 promoters. These two sets are used for all our analyses.

Randomly selected DNA sequences from mouse were used as the background set for analysis of TF binding sites in mouse promoters. These DNA sequences were 200 bp long and selected randomly from all mouse chromosomes, with the number of sequences from each chromosome proportional to the length of the chromosome. In total we selected 41,000 such random DNA sequences (Dataset S1).

TSS types. We determined the GC content of the [−100, −1] region and the [+1, +100] region relative to TSS location for each individual TSS. The TSS is considered to be between positions −1 and +1. The upstream or downstream segment was defined as GC-rich if $G + C > 50\%$ in the region. Otherwise, the region was defined as AT-rich. Four types of TSSs were defined based on the GC richness in the upstream and downstream segments as follows (Table 1): type A, GC-rich upstream and downstream (GC-GC); type B, GC-rich upstream and AT-rich downstream (GC-AT); type C, AT-rich upstream and GC-rich downstream (AT-GC); and type D, AT-rich upstream and downstream (AT-AT). Each TSS can be represented as a point in the *x*–*y* plane, where *x* corresponds to the GC content upstream and *y* corresponds to the GC content downstream of the considered TSS. For mouse and human these distributions are depicted in Figure 1A.

TF binding sites in promoters. We used all available matrix models of TFBSs contained in the TRANSFAC Professional (version 8.4) database [31] and mapped them to the extracted sequences. We used minSUM profiles for the threshold of the matrix models since these contain the optimized threshold values for the core and matrix scores [32]. The thresholds in minSUM are based on optimization that provides the minimum sum of false positive and false negative TFBS predictions. To determine the overrepresentation of TFBSs found in the target set, we used the method of Bajic et al. [15]. All TFBSs mapped to target promoters were ranked based on their ORI as defined by Bajic et al. [15]. For ORI = 1 or close to this value, there is no overrepresentation of the motif in the target promoter group. We also estimated the likelihood of observing these TFBSs in the target set using the background random promoter set as a reference. The null hypothesis was that the proportion of sequences in the target set in which a particular PE was found was the same as that in the background set. The *p*-values were calculated using right-sided Fisher's exact tests based on hypergeometric distribution. The original *p*-values were subjected to Bonferroni correction for multiplicity testing. If the corrected *p*-value of the pattern was not greater than 0.05, we placed a plus sign after the ORI value in the provided tabular reports.

Most significant PEs. For each of the TSS types in mouse, we analyzed the 150 top-ranked PEs (based on the values of ORI). This represents about 10% of all (1,428) PEs analyzed. We also required that the PEs have an ORI of at least 1.5 and that the PE be found in at least 10% of the target sequences. Details are explained in Tables S1–S3.

TATA boxes. The TATA box model used was based on that of Bucher [22]. The threshold used was 0.75, while score was normalized between zero and one (analogous to Bajic et al. [33]). A TATA box was considered detected if the maximum value of the score in the [−50, −1] region was higher than the threshold. Only one TATA box was assumed in the [−50, −1] region.

eVOC, GO, and tissue expression libraries. In order to assess the biological significance of our TSS classification system, we assigned TSSs according to different GO and eVOC categories, as well as tissue libraries in FANTOM3 collection (http://fantom.gsc.riken.go.jp/download.html). GO–FANTOM mapping data was downloaded from the RIKEN Web site (ftp://fantom.gsc.riken.jp/FANTOM3/annotation/fantomdb-3.0/anndata.txt.gz). The eVOC system consists of a set of orthogonal controlled vocabularies that unify gene expression data by mapping between the genome sequence and expression phenotype information. The eVOC human anatomy ontology [13] and the newly developed mouse adult and developmental ontologies (http://www.evocontology.org) have been mapped to the FANTOM3 library descriptions, providing a hierarchical representation of tissues, cell types, and developmental stage information. This allows for a standardized analysis of gene expression and promoter profiles independent of the original annotation vocabulary used in the original dataset.

For the generation of the results presented in Table S4, we used ortholog gene groups between mouse and human as defined at ftp://ftp.ncbi.nih.gov/pub/HomoloGene. Table S5 for mouse data contains statistics of all GO and tissue expression libraries from FANTOM3, complemented by the Bonferroni corrected p-value (right-sided Fisher's exact test based on hypergeometric distribution) for the null hypothesis that the proportion of TSSs of a specific type in the considered GO/tissue library is the same as what one can expect based on the distribution of these TSSs in mouse.

Supporting Information

Dataset S1. Supplementary Nonpromoter Data

Found at DOI: 10.1371/journal.pgen.0020054.sd001 (2.5 MB ZIP).

Figure S1. Number of TSSs of the Four Types in Human and Mouse Genomes under the Change of Parameters

Blue, green, red, and light blue correspond to TSSs of type A, B, C, and D, respectively. From graphs in the first row we observe that when the length of the region considered changes, the numbers of TSSs of the different types remain almost unchanged. We changed the length of upstream and downstream regions from $[-x, -1]$ and $[+1, +x]$, respectively, with values of x from 50 to 150. From graphs in the second row we observe that the numbers of TSSs within the four types gradually change with the change of threshold for GC content. We changed this threshold from 40% to 60%.

Found at DOI: 10.1371/journal.pgen.0020054.sg001 (24 KB PDF).

Figure S2. Distributions of TFs Found to Be Common among the Top 150 PEs in Comparisons of Different TSS Types

(A) Comparison of types A and B upstream regions.
(B) Comparison of types B and D downstream regions.
(C) Comparison of types A and C downstream regions.
(D) Comparison of types C and D upstream regions.

Found at DOI: 10.1371/journal.pgen.0020054.sg002 (40 KB PDF).

Table S1. List of Top 150 PEs That Appear with a Frequency of 10% or Greater in Upstream and Downstream Regions of Different TSS Categories

Comparison is carried out against a background of random mouse sequences. Ranking is based on ORI value. The higher the ORI, the higher the rank. We present results for the four TSS types (A, B, C, and D). For each PE we give the strand where it is found (+1 or −1), name of TFBS, ORI value, percentage of promoters in the target set that contain the PE, percentage of sequences in the background set that contain the PE, probability of finding the PE in the target set (given as one prediction per nucleotide), probability of finding the PE in the background set (given as one prediction per nucleotide), and Bonferroni corrected p-value. A plus sign added after the ORI value indicates that the PE is enriched in a statistically significant manner at the level 0.05. Almost all top-ranked elements appear to be statistically significantly enriched in the target sets.

Found at DOI: 10.1371/journal.pgen.0020054.st001 (329 KB PDF).

Table S2. Common and Specific TFBSs in the Four TSS Types

PEs are compared relative to the same GC richness and same location (upstream or downstream) in different TSS types. The signs plus or minus indicate whether the PE was found to be significantly enriched in the considered region for the considered TSS type. For example, the first column (yellow), which shows comparison between the AT-rich downstream domains in TSS types B and D, contains a common element denoted as "+ − Ets." This means that Ets was found significantly enriched for the B type, but its enrichment was not significant for D type. When an element is unique for one or another group, then it is associated only with one plus or minus sign.

Found at DOI: 10.1371/journal.pgen.0020054.st002 (53 KB PDF).

Table S3. List of Significant PEs Unique and Common for Different TSS Types in the Upstream and Downstream Segments

The yellow highlighted TFs are unique for the considered groups when compared with the same upstream or downstream segment of another TSS type with the same GC richness.

Found at DOI: 10.1371/journal.pgen.0020054.st003 (47 KB PDF).

Table S4. GO Categories That Preserve Enrichment in Specific TSS Types between Human and Mouse

We considered only those TSSs whose generated transcripts belong to the same homology group as defined on the NCBI Web site (ftp://ftp.ncbi.nih.gov/pub/HomoloGene). We only considered GO categories that were supported by at least 60 TSSs and where the target TSS type was supported by at least three TSSs.

Found at DOI: 10.1371/journal.pgen.0020054.st004 (96 KB PDF).

Table S5. All GO and Tissue-Specific Libraries with Distribution of TSSs across the Four TSS Types in Mouse

The table presents the total number of TSSs associated with the category (GO or expression library), the number of TSSs of individual TSS type, the percentage of TSSs in that TSS type, enrichment of TSSs in the TSS type relative to what can be expected based on the distribution of all TSSs in mouse across all four TSS types, and Bonferroni corrected p-values calculated based on right-sided Fisher's exact tests for the null hypothesis that the proportion of TSS type found in the target group is the same as that of the general mouse distribution. For example, there are 253 transcripts associated with GO:0006955. Of these, 52 transcripts have a TSS of type C. For the number of transcripts in this GO category, one would expect only 11 transcripts with TSSs of type C. Thus, in this GO category, we have 4.85-fold enrichment of transcripts of this type (compared to what we would expect based on the distribution of all transcripts across the four TSS types). If in any of the GO/eVOC categories or tissue libraries, at least one of the TSS groups of transcripts has enrichment that is 1.5-fold or greater than the expected value, we consider such TSS type overrepresented.

Found at DOI: 10.1371/journal.pgen.0020054.st005 (2.9 MB PDF).

Acknowledgments

Author contributions. VBB, JK, PC, and YH conceived and designed the experiments. VBB, SLT, and CK performed the experiments. VBB, SLT, AC, CS, LL, and OH analyzed the data. VBB, SLT, LY, OH, AK, WH, CK, JK, PC, and YH contributed reagents/materials/analysis tools. VBB, AC, CS, OH, and DAH wrote the paper.

Funding. This study was supported by a research grant for the RIKEN Genome Exploration Research Project from the Ministry of Education, Culture, Sports, Science and Technology of Japan to YH, a grant of the Genome Network Project from the Ministry of Education, Culture, Sports, Science and Technology of Japan to YH., and a grant for the Strategic Programs for R&D of RIKEN to YH.

Competing interests. The authors have declared that no competing interests exist. ■

References

1. Suzuki Y, Yamashita R, Sugano S, Nakai K (2004) DBTSS, Database of transcriptional start sites: Progress report 2004. Nucleic Acids Res 32: D78–D81.
2. Suzuki Y, Yamashita R, Shirota M, Sakakibara Y, Chiba J, et al. (2004) Large-scale collection and characterization of promoters of human and mouse genes. In Silico Biol 4: 0036.
3. Aerts S, Thijs G, Dabrowski M, Moreau Y, De Moor B (2004) Comprehensive analysis of the base composition around the transcription start site in metazoan. BMC Genomics 5: 34.
4. Imanishi T, Itoh T, Suzuki Y, O'Donovan C, Fukuchi S, et al. (2004)

Integrative annotation of 21,037 human genes validated by full-length cDNA clones. PLoS Biol 2: e162. DOI: 10.1371/journal.pbio.0020162
5. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, et al. (2005) The transcriptional landscape of the mammalian genome. Science 309: 1559–1563.
6. Vinogradov AE (2005) Noncoding DNA, isochores and gene expression: Nucleosome formation potential. Nucleic Acids Res 33: 559–563.
7. Vinogradov AE (2003) Isochores and tissue-specificity. Nucleic Acids Res 31: 5212–5220.
8. Vinogradov AE (2003) DNA helix: The importance of being GC-rich. Nucleic Acids Res 31: 1838–1844.
9. Levitsky VG, Podkolodnaya OA, Kolchanov NA, Podkolodny NL (2001) Nucleosome formation potential of eukaryotic DNA: Calculation and promoters analysis. Bioinformatics 17: 998–1010.
10. Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, et al. (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc Natl Acad Sci U S A 100: 15776–15781.
11. Ng P, Wei CL, Sung WK, Chiu KP, Lipovich L, et al. (2005) Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. Nat Methods 2: 105–111.
12. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene Ontology: Tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25–29.
13. Kelso J, Visagie J, Theiler G, Christoffels A, Bardien S, et al. (2003) eVOC: A controlled vocabulary for unifying gene expression data. Genome Res 13: 1222–1230.
14. Kel-Margoulis OV, Tchekmenev D, Kel AE, Goessling E, Hornischer K, et al. (2003) Composition-sensitive analysis of the human genome for regulatory signals. In Silico Biol 3: 0013.
15. Bajic VB, Choudhary V, Hock CK (2004) Content analysis of the core promoter region of human genes. In Silico Biol 4: 109–125.
16. Kadonaga JT (2004) Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors. Cell 116: 247–257.
17. Kim TH, Barrera LO, Zheng M, Qu C, Singer MA, et al. (2005) A high-resolution map of active promoters in the human genome. Nature 436: 876–880.
18. Mueller CR, Maire P, Schibler U (1990) DBP, a liver-enriched transcriptional activator, is expressed late in ontogeny and its tissue specificity is determined posttranscriptionally. Cell 61: 279–291.
19. Shimizu H, Kang M, Iitsuka Y, Ichinose M, Tokuhisa T, et al. (2000) Identification of an optimal Ncx binding sequence required for transcriptional activation. FEBS Lett 475: 170–174.
20. Shirasawa S, Yunker AM, Roth KA, Brown GA, Horning S, et al. (1997) Enx (Hox11L1)-deficient mice develop myenteric neuronal hyperplasia and megacolon. Nat Med 3: 646–650.
21. Hatano M, Aoki T, Dezawa M, Yusa S, Iitsuka Y, et al. (1997) A novel pathogenesis of megacolon in Ncx/Hox11L.1 deficient mice. J Clin Invest 100: 795–801.
22. Bucher P (1990) Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. J Mol Biol 212: 563–578.
23. Burke TW, Kadonaga JT (1996) *Drosophila* TFIID binds to a conserved downstream basal promoter element that is present in many TATA-box-deficient promoters. Genes Dev 10: 711–724.
24. Burke TW, Kadonaga JT (1997) The downstream core promoter element, DPE, is conserved from *Drosophila* to humans and is recognized by TAFII 60 of *Drosophila*. Genes Dev 11: 3020–3031.
25. Kutach AK, Kadonaga JT (2000) The downstream promoter element DPE appears to be as widely used as the TATA box in *Drosophila* core promoters. Mol Cell Biol 20: 4754–4764.
26. Willy PJ, Kobayashi R, Kadonaga JT (2000) A basal transcription factor that activates or represses transcription. Science 290: 982–984.
27. Lagrange T, Kapanidis AN, Tang H, Reinberg D, Ebright RH (1998) New core promoter element in RNA polymerase II-dependent transcription: Sequence-specific DNA binding by transcription factor IIB. Genes Dev 12: 34–44.
28. Lim CY, Santoso B, Boulay T, Dong E, Ohler U, et al. (2004) The MTE, a new core promoter element for transcription by RNA polymerase II. Genes Dev 18: 1606–1617.
29. Lewis BA, Kim TK, Orkin SH (2000) A downstream element in the human beta-globin promoter: Evidence of extended sequence-specific transcription factor IID contacts. Proc Natl Acad Sci U S A 97: 7172–7177.
30. Lee DH, Gershenzon N, Gupta M, Ioshikhes IP, Reinberg D, et al. (2005) Functional characterization of core promoter elements: The downstream core element is recognized by TAF1. Mol Cell Biol 25: 9674–9686.
31. Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, et al. (2003) TRANSFAC: Transcriptional regulation, from patterns to profiles. Nucleic Acids Res 31: 374–378.
32. Kel AE, Gossling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, et al. (2003) MATCH: A tool for searching transcription factor binding sites in DNA sequences. Nucleic Acids Res 31: 3576–3579.
33. Bajic VB, Seah SH, Chong A, Krishnan SP, Koh JL, et al. (2003) Computer model for recognition of functional transcription start sites in RNA polymerase II promoters of vertebrates. J Mol Graph Model 21: 323–332.

Author/s:
Bajic, VB;Tan, SL;Christoffels, A;Schonbach, C;Lipovich, L;Yang, L;Hofmann, O;Kruger, A;Hide, W;Kai, C;Kawai, J;Hume, DA;Carninci, P;Hayashizaki, Y

Title:
Mice and men:: Their promoter properties

Date:
2006-04

Citation:
Bajic, V. B., Tan, S. L., Christoffels, A., Schonbach, C., Lipovich, L., Yang, L., Hofmann, O., Kruger, A., Hide, W., Kai, C., Kawai, J., Hume, D. A., Carninci, P. & Hayashizaki, Y. (2006). Mice and men:: Their promoter properties. PLOS GENETICS, 2 (4), pp.614-626. https://doi.org/10.1371/journal.pgen.0020054.

Persistent Link:
http://hdl.handle.net/11343/259659