Genome-wide analysis of mammalian promoter architecture and evolution based upon human and mouse CAGE data, Supplementary Data

Piero Carninci[1, 2*] Albin Sandelin[1, 3*], Boris Lenhard[1, 3*†], Shintaro Katayama[1], Kazuro Shimokawa[1], Jasmina Ponjavic[1◆], Colin A. M. Semple[1, 4], Martin S. Taylor[1, 5], Pär G. Engström[3], Martin C. Frith[1, 6], Alistair R. R. Forrest[6], Wynand B. Alkema[3], Sin Lam Tan[7], Charles Plessy[2], Rimantas Kodzius[1,2], Timothy Ravasi[1, 6, 8], Takeya Kasukawa[1, 9], Shiro Fukuda[1], Mutsumi Kanamori-Katayama[1], Yayoi Kitazume[1], Hideya Kawaji[1, 9], Chikatoshi Kai[1], Mari Nakamura[1], Hideaki Konno[1], Kenji Nakano[1, 9], Salim Mottagui-Tabar[3§], Peter Arner[10], Alessandra Chesi[11], Stefano Gustincich[11], Francesca Persichetti[12], Harukazu Suzuki[1], Sean M. Grimmond[6], Christine A. Wells[19], Valerio Orlando[13], Claes Wahlestedt[3§], Edison T. Liu[14], Matthias Harbers[15], Jun Kawai[1,2], Vladimir B. Bajic[1, 7,16], David A. Hume[1, 6*,**], Yoshihide Hayashizaki[1, 2, 17, 18***]

Affiliations of the authors

[1]Genome Exploration Research Group, RIKEN Genomic Sciences Center (GSC), RIKEN Yokohama Institute, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa, 230-0045, Japan
[2]Genome Science Laboratory, Discovery Research Institute, RIKEN Wako Institute, 2-1 Hirosawa, Wako, Saitama, 351-0198, Japan
[3]Center for Genomics and Bioinformatics, Karolinska Institutet, Berzelius v. 35, S-171 77 Stockholm, Sweden
[4]MRC Human Genetics Unit, Western General Hospital, Crewe Road, Edinburgh, EH4 2XU, UK
[5]University of Oxford, Roosevelt Drive, Oxford, OX3 7BN, UK
[6]ARC Special Research Centre for Functional and Applied Genomics, Institute for Molecular Bioscience, The University of Queensland, Brisbane Qld, 4072, Australia
[7]Knowledge Extraction Laboratory, Institute for Infocomm Research, 21 Heng Mui Keng Terrace, 119613, Singapore,
[8]Department of Bioengineering, University of California San Diego, 9500 Gilman Drive, 0412 La Jolla, CA 92093,
[9]Broadband Communication Service Business Unit, Network Service Solution Business Group, NTT Software Corporation, Teisan Kannai Bldg. 209, Yamashita-cho Naka-ku, Yokohama, Kanagawa, 231-8551, Japan,
[10]Department of Medicine, Karolinska Institute, Huddinge University Hospital, S 141 86 Huddinge, Sweden,
[11]The Giovanni Armenise-Harvard Foundation Laboratory Sector of Neurobiology International School for Advanced Studies I.S.A.S.-S.I.S.S.A. AREA Science Park Padriciano 99, 34012 Trieste Italy,
[12]Sector of Neurobiology International School for Advanced Studies I.S.A.S.-S.I.S.S.A. AREA Science Park Padriciano 99, 34012 Trieste Italy,
[13]Dulbecco Telethon Institute, IGB CNR, Epigenetics and Genome Reprograming lab, Via Pietro Castellino 111, Napoli, 80131, Italy,
[14]Genome Institute of Singapore, 60 Biopolis Street #02-01, Singapore 138672,
[15]Kabushiki Kaisha Dnaform, 1-3-35, Mita, Minato-ku, Tokyo, 108-0073, Japan,

[16]South African National Bioinformatics Institute, University of the Western Cape, Private Bag X17, Bellville, South Africa,
[17]Yokohama City University, 1-7-29 Suehiro-cho Tsurumi-ku Yokohama 230-0045 Japan,
[18]Graduate School of Comprehensive Human Science, University of Tsukuba, 1-1-1 Tennodai, Tsukuba-shi Ibaraki-ken, 305-8577, Japan. School of Biomolecular and Biomedical Science,
[19]The Eskitis Institute for Cell and Molecular Therapies, Griffith University, Nathan Campus, Kessels Rd, QLD 4111, Australia.

Correspondence footnotes

**To whom technical correspondence should be addressed. E-mail: D.Hume@imb.uq.edu.au, tel: +61-7-33462073, Fax: +61-7-33462103

*** To whom general correspondence should be addressed. E-mail: yosihide@gsc.riken.jp, Tel: +81-45-503-9222, Fax: +81-45-503-9215

Additional footnotes

† Current Address: Bergen Center for Computational Science, Unifob AS, University of Bergen, Thormøhlensgate 55, N-5008 Bergen, Norway

§ Current address: Scripps Florida, Jupiter, FL 33458, USA.

◆ Current address: MRC Functional Genetics Unit, Department of Physiology, Anatomy and Genetics, University of Oxford, Oxford OX1 3QX, United Kingdom.

* PC, AS,BL and DAH contributed equally to this work.

This Supplementary On-line Material (SOM) contains supporting information including consideration of the reproducibility and accuracy of the CAGE technology, additional analysis of the CAGE data set and the detailed methodological sections.
The Supplementary Online material is also available as a single consolidated file at http://fantom3.gsc.riken.jp and at www.macrophages.com.

Contents

1. Demonstrations that CAGE technology captures true 5' ends

1-a. Introduction

A key issue underlying the analysis of transcription start sites is the level of confidence one can ascribe to individual CAGE tags. The CAGE technology relies on two independent biochemical events, the extension of reverse transcriptase to the 5-' end of the transcript, and the CAP-dependent second strand synthesis, capture and cloning of a cDNA. If reverse transcriptase fails to generate a complete full length extension product AND that product is captured by the CAP trapping procedure, the result will be a product that is less than full length. As noted in the main text, the intent of this project was to identify the transcription start sites of the largest possible set of protein-coding genes, and also to gain some insight into the dynamic regulation of TSS use (since CAGE tag frequency in a particular library provides an indication of the level of expression). We therefore chose to sequence a very large set of different libraries at around 50,000 to 100,000 tags per library. With this sequencing depth, less abundant transcripts, including the large numbers of non-coding RNAs which represent a major class of rare transcripts, have been sampled randomly. Because of the sampling approach, the coverage of rare transcripts is significantly less than one-fold, and accordingly, there is a strong bias towards singletons in the complete data set (Table 1). We have focused on clusters supported by two or more tags in detailed analyses because the majority is supported by independent libraries, and the probability of two tags occurring in the same location in the number of cases observed by any chance event is infinitesimal（see below). Several lines of evidence indicate that the overall error rate in the CAGE method is very small(see section 1b-g below) and accordingly that even singleton CAGE tags define genuine 5' ends with a high level of confidence. The precise confidence cannot be evaluated without another complete data set, but many individual singleton CAGE tags can be further validated by considering the enrichment of capped molecules, the local sequence context (e.g. Inr sequence), location relative to cDNA and 5'EST sequences, promoter architecture, conservation across species and other supporting evidence.

1-b. Enrichment of capped molecules

We assessed the efficacy of cap selection by examining the enrichment for RNA PolII-derived mRNAs compared to uncapped ribosomal RNAs in random-primed libraries. When using random priming, most of the first-strand cDNA is ribosomal RNA, which by visual inspection was estimated to exceed 90%. Since the cap-trapper procedures enriches for capped RNAs, the decrease of non-capped ribosomal-derived cDNA gives a measure of relative enrichment. We selected 10,000 CAGE tags for both human random primed and mouse random-primed libraries and blasted these sequences against 18,133 ribosomal RNA sequences extracted from the Rfam database (RF00001 5S_rRNA; RF00002 5_8S_rRNA; RF00177 SSU_rRNA_5). The cap-selected random-primed CAGE libraries contain no more than 2.6% (human) and 1.42% (mouse) ribosomal RNA. Assuming that the ribosomal RNA constitutes the 90% of the total RNA, the enrichment of non-ribosomal, presumably capped RNA is between 333 to 625 fold (calculated with human and mouse data, respectively).

The issue of whether all of the CAGE tags are genuine 5' ends can be assessed by looking at the distribution of tags across individual genes. If extension of reverse transcriptase to the 5'-end of the transcript, and the CAP-dependent second strand synthesis, capture and cloning of a cDNA fails for any reason due to incomplete extension, or capture of a transcript that has been subjected to nuclease cleavage, the frequency of tags must decay exponentially from the 3' end to the 5' end of the transcript, since each truncation reduces the proportion of genuine full length transcripts captured. As shown clearly in the example of the albumin gene (alb1) in Fig. S2A, as well as Fig. 6A, this is clearly not the case. In the subset of genes in which we observe apparent start sites over all of the exons, their apparent frequency does not decay from the 3' to the 5' end. Furthermore, there is no correlation between the expression level of the full-length transcript and the amount of tags in inner exons in a given tissue (see above).

1-c. CAGE mapping positions are non-randomly distributed

These observations above alone argue unequivocally that CAGE specifically captures genuine 5' termini. The same point can also be made from purely statistical viewpoint by considering the likelihood of finding the observed number of multitag CTSS (TSS with more than 1 tag exactly aligned at their 5' ends) occurring by chance: the associated p-values are below what normal computers can handle without underflow errors ($<10^{-324}$) (Table S5A). As an example, the number of CTSS composed of two tags occurs

close to 15 times as often than expected: the ratio is exponentially increased when assessing CTSS with increasing tag density (Table S5A). This is an argument for the validity of single tags, since multiple tags in the majority of cases are from more than one library (see below). The high occurrence of multitag CTSS would be impossible if single CAGE tags had significant random errors.

As a further indication of the reproducibility of CAGE, we assessed the proportion of tag clusters (TCs) with more than one tag that come from distinct libraries. The proportion was 77.83% for the set of 67,660 TCs supported by 2 tags, increasing to over 95% for TCs having more than 3 tags (Table S5B). The reproducibility is even higher than this if we assess the replication of exactly aligned 5' ends in distinct libraries by repeating the analysis on CTSS with more than one tag: the proportion was 87.45% for the set of 203,401 CTSS supported by 2 tags within classified TCs, and only marginally lower (77.89%) for the total set of 389,412 CTSS supported by 2 tags (Table S5C). The lower number in the second instance arises because the unclassified set includes rare transcripts, including non-coding RNAs, which have only been detected in the small number of libraries that were sequenced at greater depth. Hence, in these cases there is a greater likelihood that 2 tags have been derived from the same library. If we examine CTSS supported by 5 tags; 99.03% derived from classified TCs have tags from more than one library, and 97.33% derived from unclassified TCs. These data also show that less abundant CAGE tags do, indeed, derive from less abundant mRNAs in that they are unlikely to be sampled again in the same library. By extension, the high incidence of singleton CAGE tags is principally a consequence of the strategy used and the depth of sampling in individual libraries.

1-d. Empirical support for promoter identification by CAGE

Amongst the 20-30,000 protein-coding genes in the mouse genome, relatively few transcription start sites have been determined experimentally. Two approaches have generally been used, nuclease protection and primer extension or 5'RACE. Nuclease (RNase or S1) protection assays are relatively insensitive in detecting minor starting sites, and because the protected fragments are usually detected by autoradiography, it does not commonly provide quantitative information about relative promoter usage. Additionally, the number of genes for which the TSS has been determined in 250 different RNA samples is extremely small. Nevertheless, we have undertaken both experimental and published validation of the CAGE data.

For experimental validation, we analyzed the complex OPRM locus, which shows 36 distinct TSS spread over intergenic, exonic and intronic regions using 5' RACE. 5' RACE is distinct from CAGE in that RACE takes place only after RNA ligation adds a primer at the 5' end instead of the cap site. This is obtained after the phosphate group at the 5'-end of non-capped RNA molecules removed by phosphatase treatment. This biochemical reaction is completely distinct from the CAGE technology and therefore serves to validate the CAGE-determined TSSs.
The majority of TSS in the OPRM receptor locus (30 sites) are represented by single CAGE tags; nevertheless, 26 (87%) of these single tags could be validated using RACE even on a single tissue sample.

To compare the CAGE and other methods such as nuclease protection we randomly chose from the literature 19 examples of mouse genes in which the TSS had been determined by nuclease protection, RNase protection and primer extension; in most cases the examples chosen have multiple TSS. Despite the bp-level inaccuracy of the gel-based methods and the use of different cell lines/tissues, in the majority of cases CAGE data is wholly consistent with the previous results (Fig. S7A-S shows a set of published examples). The CAGE tags do not, of course, correlate perfectly with the published start sites in every case. Published data represents start site usage from a single cell type or tissue under a single condition, whereas CAGE is based upon cumulative data from many different libraries. It has much greater depth of coverage than published 5'RACE, and greater sensitivity than nuclease protection for detecting minor start sites. However, the CAGE tags clearly map to the same regions, and where there is a distinction, there is no bias towards mapping downstream (i.e. there is no evidence that they are based upon truncated 5' ends).
In many of the published reports, arrays of weaker start sites are apparent in the gel images are noticable (but not annotated/discussed); in these cases the more comprehensive CAGE tag mapping will indicate

both the major start site indicated in the paper as well as the weaker sites. An example studied in detail by one of the authors (DAH) is the start site of the macrophage-specific CSF-1R promoter. This is a member of the novel purine-rich subclass of broad promoters (see below). The mouse CSF-1R promoter was analyzed previously by both primer extension and RNase protection[1]. Primer extension was based upon primers designed to hybridize to the longest known cDNA. The RNase protection actually detected clusters of start sites downstream of those identified by primer extension. Reevaluation of those results by the CAGE data shows that the RNase protection is precisely congruent with the CAGE information (Fig S7T).

1-e. Validation of CAGE based upon known transcription start site bias

The strong signal-noise ratio in CAGE-based start site identification is evident a priori from the stringent mapping of the start sites in TATA-containing promoters. The tight distribution of TATA-associated TSS is based solely on CAGE tag occurrences and corresponds precisely to the historically defined position around -30.

As noted in the main text, the actual bp detected as 5' ends by CAGE display a profound bp bias towards the initiator consensus (PyPu). Further evidence of the non-randomness of the CAGE tags was obtained by analyzing the distribution of initiation site [-1,+1] dinucleotides for TSS having 1,2 and up to10 or more CAGE tags separately (note that the -1 position is not part of the tag). Regardless of the level of CAGE support, each dinucleotide distribution is significantly different from a distribution of 10,000 randomly sampled non-repetitive, non-overlapping genomic dinucleotides ($p < 2.2* 10^{-16}$ in all cases, chi-square test). Simply stated, the identified start sites correspond to the initiator sequence. While the distributions with different level of CAGE support overall are similar (Figure S4A-B), there is a higher preference to PyPu nucleotides, especially CA, in high-expressed transcripts, while other dinucleotides (in particular GG) are gradually favored as the CAGE tag count is lowered.

1-f. Support for CAGE data from cross-species validation

Start sites identified by CAGE not only have a non-random sequence preference, they are correlated across species; the human data set can be considered an independent experimental validation of the mouse data. We report a strong correlation between the TSS usage in orthologous mouse/human promoters, both in terms of distribution shape (Fig S3J-K) and even on base pair level (Fig. S4C-H). The discrepancies in the TSS distribution can be correlated to single nucleotide substitutions (Fig. 4I). From a different perspective, we can ask what proportion of the CTSS is conserved across species (Table S5D). For this purpose, we selected a set of mouse and human genes where the promoters can be unambiguously aligned, and determined the numbers of CTSS that are replicated in liver-derived libraries across species as a function of tag density in the human CTSS. Liver was chosen for this comparison because it has been polled to approximately equal depth in the two species. If CAGE accurately detects 5' ends, detection of any particular 5' end will obey a Poisson distribution, with a frequency determined by relative abundance and depth of sampling.

This series reached a saturation at around 81.9% for CTSS with 20 tags. If one assumes that the detection of TSS is purely a function of their relative use to initiate transcription and their sampling density, then this series should also obey a Poisson distribution, and for singletons sampled to equivalent depth in two samples, the expected replication rate should be 1-1/e, or 67%. In this case, the support for singletons across species was 59%, or 72% of the observed asymptote. This is remarkably close to the prediction given the possible caveats. Firstly, the alignment cannot be perfect. To take account of arbitrary positioning of bp insertions or deletions, we allowed a 1bp slack in the actual 5' end allowed as identical. Secondly, the sampling depth on any individual promoter across species cannot be precisely equivalent because it depends on relative abundance and depth of sequencing. The human data comes with a small number of libraries sequenced to great depth, whereas mouse comes from many libraries with liver in different states. Finally, on individual broad promoters mouse and human differ in precise start site dominance; so the frequency of use of weaker initiators is influenced by the "strength" of other sites in the region. The degree of replication can also be seen by examining particular genes in detail. The highly-conserved tissue-specific brain proteolipid protein 1 gene (Plp1) has a broad CTSS region, which in humans is dominated to a greater extent than in mouse by one major start site, but the overall profile is strikingly similar (Fig S7U). The human promoter has been polled to a 3-fold greater extent than in the mouse. Out of the few start sites (including singletons) that are not replicated in both species, the majority have mutations in the initiator sequence.

1-g. Correlation between CTSS and active promoters identified by ChIP of transcription complexes in humans

If CAGE is identifying active promoters, we would expect the sites to be occupied by transcription initiation complexes. Chromatin immunoprecipitation linked to genome tiling arrays is clearly not feasible on the same range of tissues that has been used to generate the mouse and human CAGE data. However, it is possible to assess the correlation between CAGE and genomic regions occupied by TAF1 (TBP-associated factor 1 of the TFIID, associated with the preinitiation complex) defined using chromatin immunoprecipitation (ChIP) and hybridized to Nimblegen tiling arrays in a recent study[2]. We investigated how many of these sites within the human ENCODE regions had CAGE tags within 100 bp of inferred TAF1 binding sites. Specifically, the "known sites" and "novel sites" ENCODE track from the UCSC ENCODE browser in human genome assembly hg17 was examined; the human CAGE data is mapped to the same assembly. The 100 bp limit was chosen because i) CAGE corresponds to the actual starts of transcripts, while TAF1 is part of the pre-initiation complex ii) the resolution of ChIP is at best in the range of 10s of bp and iii) more importantly, the 50-mer Nimblegen probes are placed at 100bp genomic intervals. In the validation of the ChIP study itself[2] using GenBank cDNA, the corresponding limit was 2.5kb. On this basis, 58.3% of the TAF1 sites were supported by CAGE. Given that this study only interrogated a single cell line (which has no direct correspondence to the tissue libraries used for CAGE), the agreement is remarkable and similar to the analogous overlap study using cDNA ends (described in main text).

2. Extended Biological findings

2-a. Detailed analysis of exonic promoter activity

For a given tissue, we counted the CAGE tag density (tags/bp) in the +-100bp of the 5' edge of the TUs representative cDNA transcript. Only transcripts with at least 100 tags from the relevant tissue in this region were considered. We compared this value to the overall CAGE tag density for the inner exons (tags/bp) of the corresponding cDNAs. Lung, liver and macrophage transcripts from mouse were analyzed this way, since these were the tissues in which multiple CAGE Tag libraries were generated and therefore those in which there were sufficient tags for analysis of a substantial number of TUs. Fig 1b shows that there is no relation between the two variables: hence that there is no correlation between overall expression level and extent of exonic promoter activity. This finding also supports the view that transcripts arising from within exons are not attributable to incomplete synthesis or capture of full length cDNA, since such truncations would have to be proportional in some measure to the abundance of the corresponding full length transcript.

To determine whether there was any correlation between primary (i.e. 5'end) promoter architecture and internal exonic promoter activity the 5% most and least extreme cases of exonic promoter activity in inner exons were collected and annotated for each tissue. The SP class of promoter was strongly over-represented amongst the 5' promoters of transcripts with high exonic promoter while corresponding promoters for transcripts with low exonic promoter activity are enriched for the BR class (Fig. S2B-D). Consistently, transcripts with high exonic promoter activity have significantly fewer CpG islands in their major promoter compared to transcripts with no or little exonic promoter activity (Fisher 2-tail test p values: 1.48E-17 (liver), 2.31E-3 lung, 3.03E-5 (macrophages))(Fig. S2E-G). Given that the SP class is also over-presented amongst tissue-specific promoters, high exonic promoter activity might also be associated tissue-specific transcription. To test this proposition for liver, lung and macrophages we plotted the average tag density (mean tags/bp) in inner exons versus the tissue specificity of the full-length transcript. The tissue specificity was assessed empirically by calculating the proportion of the total number of tags in the major 5' promoter that had been derived from the tissue of interest (Figure 1D). In all three tissues studied, high exonic promoter activity is correlated with high tissue-specificity.

2-b.Promoter evolution in mammalian genomes

Comparisons of human with mouse and rat suggest that human promoter sequences are also more slowly evolving than randomly sampled sequence (Table S2B). In contrast, the comparison with dog shows that human promoters, particularly CpG promoters (substitution rate = 0.3259+/-0.0025), have evolved significantly more rapidly than randomly sampled sequences (substitution rate = 0.3070+/-0.0017), while the core 200 bp of these promoters evolved close to random sequences (Table S2B). Since this effect is

not seen for mouse versus dog TSSs (Table S2A), it could reflect events in the human lineage since divergence from dog. This is supported by comparisons between human and chimpanzee, which suggest a striking lack of constraint has been a general feature of the recent evolution of human promoters (Table S2B). This is consistent with recent findings that human promoters exhibit higher substitution rates (relative to the mutation rate) than mouse promoter regions.

2-c. Different tissues utilize different promoter types

To support the hypothesis that particular tissues preferentially utilize particular promoter architectures, we analyzed the CAGE data in two other ways. We extracted a subset of library-specific TSS sets, consisting of clusters in which more than 80% of tags (normalized for library size) come from only one tissue (or embryo). Then we determined the observed/expected ratios and the associated P-values for tissue versus shape category associations (Table S3A). Single start site promoters are strongly overrepresented in the tissue-specific transcripts, with the notable exceptions of the macrophage- and the CNS-specific genes.
Secondly, we analyzed the association of transcribed genes with their gene ontology (GO) terms. We show (Table S3B) that many specific GO categories are correlated significantly with single peak promoters, while only a handful of general categories exist where broad peaks are significantly overrepresented.

2-d. Alternative promoters are common in protein coding genes

Our data shows that previous estimates on alternative promoter usage of TUs were conservative. A majority of protein-coding genes have at least two alternative promoters; in most cases, these promoters show differential expression and are therefore likely to be subject to different regulatory mechanisms. This concept has important implications for array-based gene expression measurements – using only one type of probe per gene will increase the risk of not observing biologically significant expression patterns, as in the case of the gelsolin gene (below). The TCs described herein provide the basis for development of specific promoter arrays.

2-e. Detailed analysis of alternative promoters in the gelsolin gene

The gelsolin gene (Gsn) contributes to actin filament remodeling (Fig. S6A,B). The Gsn gene has two alternative promoters (T02F02195C0E and T02F021984A4) potentially producing the same protein product, and a third alternative promoter, T02F0219C1BB, which directs a distinct 5'UTR encoding not only a distinct methionine but an N-terminal signal peptide permitting protein secretion. Although the T02F02195C0E and T02F021984A4 have the same cytoplasmic protein product, they belong to different TC supergroups. The promoter T02F02195C0E is in the same cluster as the core promoter of the vimentin gene (Vim), an intermediary filament, and is the dominant form of Gsn expressed in macrophages. Conversely, the promoter T02F021984A4 is in the same supergroup as the laminin beta 3 gene (Lamb3), which is a part of the basement laminins, and is the main Gsn promoter in liver cells (Hepa 1-6). We infer that T02F02195C0E is used in cellular contexts where the Gsn and Vim need to be coexpressed while T02F021984A4 is used when Gsn and Lamb3 are needed. The secreted protein product of the third alternative promoter, mainly found in cerebellum and heart libraries, encodes a plasma form of gelsolin, which has a potential role to solubilize actin molecules derived from damaged cells to prevent thrombosis [3].

2-f. An in-depth look: promoting the macrophage-specific transcriptome

The [-1000, +200] region of all 159,075 TCs used in the cluster analysis was extracted, and centered on the dominant start site within each TC. A comparative analysis of the incidence of TRANSFAC[4] predicted motifs between the 450 macrophage-specific TCs, the 295 LPS-inducible TCs and a random set of TCs was carried out (40,000). The comparative data was used to calculate an over-representation index (Experimental procedures) for each promoter cluster (Table S6 and Table S7). For the macrophage-specific set (Table S6A, Table S7A), many over-represented sites are motifs recognized by members of the Ets transcription factor family, consistent with the known unique architecture of myeloid promoters[5]. Such promoters lack TATA box, GC or CCAAT box elements and belong to the broad class of promoters. The minimal requirement for a macrophage-specific promoter is multiple Ets sites, one of which must be recognized by the lineage-specific transcription factor PU.1, and the other by another Ets family member[6]. We further compared the set of BR type promoters in which the normalized fraction of CAGE tags comes from either macrophage or a CNS library that has similar properties to all BR-type promoters, even if they are tissue specific. We specifically looked for core Ets[7] and CAGA motifs. The

latter is a motif identified in the CSF-1R promoter and other myeloid promoters[5], recognized by the Ewing sarcoma proteins, a component of the basal transcription machinery (Hume, D. A., in preparation) (Fig. S5B-D). Unlike the 42 CNS-specific ones, the 63 macrophage specific promoters have a high incidence of core Ets sites within 100 bp upstream of TSS, while most CAGA sites cluster downstream of it. These preferences cannot be explained solely by different nucleotide compositions along the TC region. This example of a simple characterization of a highly specialized type of core promoters is likely to provide enough data for the construction of a predictive model for macrophage-specific promoters. We assume that our data will enable the same for other subcategories of core promoters.

The LPS-inducible set (Tables S6 and S7) is clearly distinguishable from the constitutive set by an over-representation of Rel/NFkappaB motifs, as one might presume given the well-documented roles of these factors[8]. More dramatic is the over-representation of interferon-responsive elements (IRFs), which highlights the likely involvement of interferon in a large proportion of the downstream transcriptional regulation by lipopolysaccharide[9].

3. Methods and Resources

3-a. Links to databases and resources

Novel, publically available databases and resources integrating CAGE, ESTs, full-length cDNAs and other genomic elements are described in Table S4. Sequences of tags are made available to the community also through DDBJ at the site http://www.ddbj.nig.ac.jp/whatsnew/050124-e.html.

3-b. Dataset and basic tag analysis

Except for a substantial (4.17 million tags) part of human CAGE set here described for the first time, the dataset production is described elsewhere[11], including the detailed mapping conditions for all the CAGE tags. The process to build the TCs is described in Fig. S8A, and the assignment to transcripts is in Fig. S8B.

3-c. CAGE libraries preparation

CAGE libraries[12] were prepared with a protocol developed based on the described procedures[13]. Briefly, the CAGE technology is based on priming the first strand cDNA with an oligo-dT or a random primer, starting from total RNA and synthesize the first-strand cDNA at high temperature (55-60^{o}C) in presence of trehalose and sorbitol to increase the full-length cDNA rate even in presence of strong secondary RNA structure. Then, cap-trapping is performed. Cap-trapping is a method to enrich cDNA/RNA hybrids through the cap-structure, only when the hybridized cDNA is a full-length one. After chemical biotinylation, RNase I (which cleaves only single strand mRNA at any base) is used to remove any ssRNA linking the biotinylated cap and the double-strand RNA/truncated cDNA. RNA molecules hybridized with full-length cDNA molecules are left undigested, and are next captured with streptavidin beads. After several stringent washings of the beads, full-length cDNAs are removed with mild alkali treatment. After specific addition of a linker, which contains the class-IIs restriction enzyme *Mme*I site next to the ligation junction with the 5' end of cDNAs, the second strand cDNA is synthesized. Subsequently, the cDNA is cleaved with *Mme*I, only the initial 20-21 nt of the cDNA are left attached to the 5'-end linker, while cDNA is removed. After addition of appropriate linkers and cycles of PCR and purification, restriction-digested double strand sequencing tags are obtained. After formation of concatamers, these are cloned and sequenced. The whole procedure is described in details elsewhere[13].The sequenced CAGE tags are extracted and aligned to the genome by using BlastN. Only CAGE tags without base-calling problems (no "N" nucleotides in the sequence) were used for mapping, and tags mapping on multiple genomic regions (such as tags consisting of repeats) were not used for the current analysis. CAGE alignment to the genome is complicated by the addition of one template-free C (sometimes two Cs) to the end of the first strand cDNA, which results in one (or sometimes two) G's on the DNA strand. For this reason, best alignments of at least 18 nt long or better was chosen. The G addition bias is corrected using the algorithm described below, based on unambiguous mapping cases. More technical details of mapping are described elsewhere[11].

3-d. Rules for the assignment of CAGE tags

The hierarchical structure of CAGE data is explained briefly in this paragraph and the associated methodology described in details below. Two or more individual CAGE tags that have identical sequences (and therefore identical genomic mappings) are grouped into a Representative CAGE tag. Representative CAGE tags that have exactly the same genomic starting point and strand define a CAGE tag-defined transcriptional start site (CTSS). CTSSs are grouped into tag clusters (TCs), where the member tags map to the same strand and overlap by at least one bp.

The Tag Cluster (TC) definition is exemplified in Fig S8A. A TC is a cluster of overlapping tags, spanning from the 5'-end of its 5'-most tag, to the 3' end of its 3'-most tag (1). RIKEN 5'-ESTs and 5'-end of FANTOM3 clones were also used. At first, all of the cDNAs, CAGE tags and ditags were mapped to the genome using BlastN[14]. The threshold of the Blast was at least 18 nt match for CAGE tags and at least 16 nt of each side of the ditags (32 nt alignment within a 2.5Mbp of the genome). Only best matches were considered (usually, 19-20bp) and ambiguous cases were not used in the analysis. After the genome assignment, we further grouped the tags to form TCs. We grouped all tags overlapping with one or more bp (on the same strand) into a single TC. Human TCs are defined by CAGE, 5'-end of Long-5' SAGE and dbTSS. The representative position (rep. pos.) of a TC defines the location where the largest number of mapped tags occurs based upon a prioritized source map; CAGE tags have the highest priority, then GIS, GSC, RIKEN 5'-EST, FANTOM3 clone, Long-5' SAGE and dbTSS tags have decreasing priority in the

order listed. (1). When the number of tags is equal in more than one position, the rep. pos. is taken as the 5'-end (1, 3). A tag in a TC needs to be overlapping with at least one base of another tag (3), hence, two adjacent but non-overlapping tags contribute to separate TCs unless both share sequence with a bridging tag (4). Tags on opposite strands are not considered overlapping and contribute to different TCs (5).
(Fig S8B). Rules to assign CTSSs (CAGE tag starting sites) to mRNA. This protocol is used to associate CTSS with the transcript that is the presumptive product (i.e. to give the promoter a name associated with the transcript annotation). In order of priority, the promoter is associated with a transcript annotation based upon mapping within the first exon and CDS, 3'UTR, any undefined UTR (CDS was not defined), inside an exon after the first exon and 5'UTR, any CDS, any 3'UTR, any UTR, inside intron, upstream 10bp or less, upstream 100bp or less, far upstream (more than 100bp), and downstream have decreasing priority.
The detailed rules to assign CAGE TCs to mRNA are as follows. Rule 1: a TC is assigned to an mRNA in which first exon includes the rep. pos. of the TC. Rule 2: if such mRNA is not found, the TC is assigned to an mRNA having any exon including the rep. pos. of the TC. Rule 3; when more than one mRNA complies to rule 1 or 2, the TC is assigned to the one in which 5'UTR (1st priority), CDS (2nd), or 3'UTR (3rd) includes the rep. pos. of the TC. Rule 4: if no such mRNA is found after applying rules 1 and 2, the TC is assigned to an mRNA having any intron including the rep. pos. of the TC. Rule 5: when more than one mRNA complies with rule 1-4, the TC is assigned to the one in which coding sequence is the longest. Rule 6: if no such mRNA is found following rules 1, 2 and 4, the TC is assigned to the one which 5'-end is nearest to the rep. pos. of TC. See examples in Supplementary Fig. S8 Example (1): the TC is assigned according to rule 1. Example (2): the TC is assigned according to rule 1, 3 & 5. Example (3): the TC is assigned according to rule 6.

3-e. Transformation function correcting for systematic G-addition bias

The experimental protocol for preparing CAGE uses the MmeI restriction enzyme for separating the 5'-end of the full-length cDNA and a linker sequence (the protocol steps are described in detail in elsewhere[12,13]). Because of the template-free activity of the reverse transcriptase used to prepare the cDNA, an additional G nucleotide is often attached to the 5' side of the tag. In cases where the added G does not map to corresponding genome sequence, the extra G can be confidently classed as added; we call those cases unambiguous. However, when the first nucleotide in a tag is a G that maps to the genome, we cannot know with certainty if the tag starts with a G or the following nucleotide. In situations where we have two or more Gs after each other in the genome sequence, and all have one or more tags, the ambiguity situation is worse, since for a given position *i*, tags with an added G will be mapped to the position *i-1*.
We estimated how often a G is added in tags whose mappings were unambiguous. We found that a single G is added in 87.4% of all tags, while the addition of more than one G is rare (1.95% of unambiguous cases). Since the chance of adding more than one G is small, we constructed a transformation algorithm based on that either a single G or no G is added to the tag. Each TC used for classifications below was first subject to the transformation function.
Algorithm: A CTSS is defined by the genome position of the first nucleotide and the strand. CTSSs that start with something other than a G and do not have a G directly upstream in the genome are not altered. Where a CTSS *I* starts with a mapped G and does not have another CTSS starting with G starting just before it, all CTSS following *I* (given strand) are identified until a CTSS is reached that does not start with a mapped G. This gives a set of *n* CTSSs, each harbouring a number of tags. As an example, in the sequence HHHGGGHH (where H is a nucleotide which is not a G), positions 1, 2, 3 and 8 are non-ambiguous: the number of tags starting in each position is not altered. Positions 4,5,6,7 are ambiguous and can be partitioned into three different cases: Position 4 marks the start of a series of Gs (start case), while position 7 marks the end (end case). Positions 5 and 6 are inside the series (general case).
In such a series, we are interested in two values for each position: the true number of tags N starting in the position and the number of tags falsely assigned to this CTSS F (= belonging to next CTSS). We are assuming that the number of additions of more than one G is negligible. We traverse each CTSS in the series from 5' to 3'.
Let *X* denote the number of tags observed in the CTSS, *N* denote the corrected count of tags in the CTSS, *F* denote the number of tags observed in this CTSS that belong to the next CTSS and *P* denote the chance of adding a G, assessed to be 0.8935878.
Start case: The start case is special since the number of tags with falsely added Gs can be counted (as

they are not mapping to the genome). Thus, X contains tags that belong to this CTSS but have an extra G added (denoted A), tags that belong in this CTSS but have no extra G added (denoted U) tags that belong in the next CTSS (F).
Since we can count the number corresponding to A, we can estimate the true number of tags N starting in this position:

$$N = A/P \quad (1)$$

Since we know that N cannot be larger than X, we let X be the upper bound of N.
The expected number of tags U starting in this position with no false G added is then

$$U = N*(P-1) \quad (2)$$

As discussed above, the number of tags in X that have no falsely added G must belong to this CTSS or the next. We expect that the number of tags in this position without an added G is U. Thus, the number of tags F belonging to the next CTSS is

$$F = (X-A)-U \quad (3)$$

This can be expressed as

$$F = (X-A)-N*(P-1) \quad (4)$$

We impose two constraints F must be positive and cannot be larger than X-A.
General case: The general case differs from the previous case in that only two types of tags are observed in this CTSS: tags that belong in this CTSS but have no extra G added (=U) , and tags that belong in the next CTSS (=F). Therefore,

$$X = F+U \quad (5)$$

The number of tags A belonging to this CTSS with a falsely added G is observed in the previous CTSS (due to the definition of CTSS). However, the F value calculated in the previous CTSS can be used as an estimate for A.

$$A = F_{[\text{previous CTSS}]} \quad (6)$$

Given $F_{[\text{previous CTSS}]}$, we can estimate N,U and F analogous to equation (1,2,3):

$$N = A/P \quad (7)$$

$$U = N*(P-1) \quad (8)$$

$$F = (X)-U \quad (9)$$

which can be simplified to

$$F=(x)-N*(1-P) \quad (10)$$

A is not subtracted from X, since the count A is from the previous CTSS. Again, we impose two constraints; F must be positive and cannot be larger than X.
End case: The end case differs from the previous case in that only one type of tag is present: tags that belong in this CTSS but have no extra G added (=U), since the number of tags A belonging to this CTSS that has an added G are observed in the previous CTSS, and the number of tags F belonging to the next position is zero (since this position does not start with G).

$$X=U; \quad (11)$$

Thus, we only need to calculate N, which simply is the sum

$$N = F_{[\text{previous CTSS}]}+X. \quad (12)$$

3-f. Classification of tag clusters
Tag clusters (TCs) containing 100 tags or more were classified into four mutually exclusive shape categories, using four different criteria, which were applied in a specific order. 1) A TC is assigned a single peak (SP) shape if the distance between the 75 and 25 tag density percentile within a TC is less than 4 bp; 2) If the ratio between the first and second tag peak >2 and the TC is not classed as SP, it is classed as broad with dominant peak (PB); 3) If the TC is not classed as SP or PB and there is one or more consecutive tag density 5^{th} percentile pairs with a distance exceeding 10bp a TC is classed as bi- or multimodal (MU); 4) If none of the above apply, the TC is classed as broad (BR).

3-g. Analysis of initiator sites of TSSs located in inner exons
We focused on TUs with more than 10,000 tags mapped, and analyzed those CTSSs situated in their inner exons and containing more than one tag. Since the G addition correction algorithm was designed for TCs containing a large number of tags, we choose to only analyze unambiguous CTSSs (defined above). For any such CTSS, we investigated the [-1,+1] position relative to the CTSS, counting nucleotide base combinations (one count per tag).

3-h. Calculation of over and under-represented features associated with TC shape classes.
To calculate over- or underrepresented features (tissue or transcription starting site sequence) with distinct TC shape classes two tests were applied: 1) a test to assess whether a TC has shape class (X) of interest; 2) a test to assess whether a TC is associated with the feature (Y) of interest. The results of these two tests are summarized in a contingency table, which contains the number of TCs that failed in both tests, the number of TCs that failed in test 1, the number of TCs that failed in test 2 and the number of TCs that passed both tests. The p-value for the null hypothesis that both tests are independent can be calculated with the Fisher Exact test. A low p-value indicates that the tests are not independent and that feature Y is over or underrepresented in shape class X

3-i. TATA box spacing
The occurrence of TATA-boxes in the -50 to -15 region relative to the most dominant tag peak in SP class promoters was assessed using the Bucher's weight matrix [15] and the TFBS Perl module[16] for binding site prediction. Only predictions on the forward strand having a score >75% of maximum were accepted.

3-j. Sp1 over-representation
The (-200,+200) region around every CTSSs in all TCs of each class was extracted. The occurrence of TATA and Sp1 motifs [17] was assessed using a 80% relative cutoff in the region, non-normalized (one per CTTS) and intensity-normalized. For each TC, the weight of each individual CTSS was proportional to the number of tags in it, and sum of all weights in a TC was equal to 100; this accounts for relative strengths of different CTSSs positions within a TC, while avoiding the domination of pattern by a small number of TCs with a large number of tags.

3-k. Tissue specificity comparisons in PB-class promoters
We measured the transcriptional specificity of the dominant TSS, which contains the most tags within the TC, and then the specificity of the remaining TSSs in the TC.
A discrepancy would indicate that the dominant TSS and the rest of the promoters are subject to different mechanisms, creating a "hybrid" promoter: an overlaid SP and BR class promoter. We measured transcriptional specificity of a set of tags by measuring by the relative entropy (the Kullback-Leibler distance)[18]of the RNA library distribution of a sample tag cluster with respect to the RNA library distribution of all CAGE tags:

$$d = \sum_k p_k \log_2 \left(\frac{p_k}{q_k} \right)$$

where k is the number of different libraries (103), q is the discrete probability distribution of RNA libraries for all tags and p is the discrete probability distribution of RNA libraries in the sample tag cluster. Only tags from cDNA libraries with more than 10,000 tags were used for the calculation.
We first compared the tissue specificity of the dominant peaks to the tissue specificity of the surrounding tags in the promoter, using Wilcoxon tests on the two vectors of relative entropy values. The dominant peaks have significantly higher tissue specificity ($p< < 2.2*10^{-16}$).Secondly, we divided the dominant peaks into two subsets: those that had a predicted TATA-box in the –40 to –19 region, and those lacking the TATA box. We then compared the distribution of relative entropy values for the two subsets, as above. The TATA containing set has a significantly higher tissue specificity ($p< 3.823*10^{-07}$). This indicates that the TATA box has a clear effect on the tissue specificity of the dominant TSS.

3-l. Analysis of cross-species conservation in mouse and human promoters
TSSs were derived from mouse and human CAGE TCs composed of 10 or more tags. We analyzed pairwise alignments from multiple species in the region 1Kb upstream and 200 bp downstream of mouse and human TSS. Promoters were categorized according to their TSS categories (MU, BR, PB and SP) and according to whether they possessed a CpG island or a TATA box. In addition we examined a large number of 1.2Kb alignments ('random' below) randomly sampled from the mouse and human genomes (these were intended to give us a 'base line', near-neutral substitution rate to compare to the promoter rates). We have estimated evolutionary divergence in the form of nucleotide substitution rates calculated using the REV model in PAML 3.14[19] for each TSS alignment as in[20] . The mean rate and 95% confidence intervals were calculated for each category of promoters.

3-m. Promoter evolution in mammalian genomes

TSSs were derived from mouse and human CAGE tag clusters composed of 10 or more tags. Alignment data for TSSs refers to 1.2Kb regions around CAGE TSSs (1Kb upstream and 200bp downstream of the TSS) from mouse (30,898 TSSs derived from ≥10 tag clusters examined) and human (26,290 TSSs derived from ≥10 tag clusters examined). Each region was compared to the aligned, putatively orthologous regions from various species taken from the UCSC MULTIZ whole genome multiple alignments for the mouse mm5 and human hg17 assemblies[21]. For the mouse TSSs we estimated rates from comparisons to aligned rat (Rn), human (Hs) and dog (Cf) sequences, for the human TSSs we made estimates for comparisons to chimpanzee (Pt), dog (Cf), mouse (Mm) and rat (Rn) sequences. Promoters were categorized according to their TSS categories (MU, BR, PB, and SP) and according to whether they possessed a CpG island (taken from UCSC annotation) or a TATA box (predicted using EMBOSS/TRANSFAC profile within 50bp of the TSS). Two other categories were examined: those promoters with TSSs supported by ≥100 tags (Tables 1A and B) and those supported by <100 tags ('Low'), as a simple way to examine rate differences between promoters associated with relatively high and low rates of transcription. In addition we examined a large number (23,993 in mouse, 20,605 in human) of 1.2Kb alignments ('random' below) randomly sampled from the mouse and human genomes (these were intended to give us a 'base line', near-neutral substitution rate to compare to the promoter rates). We also extracted a large number (17,025 in mouse, 33,407 in human) of 1.2Kb alignments centered upon start codons to provide a relatively well-conserved set of alignments.

We have estimated evolutionary divergence in the form of nucleotide substitution rates calculated using the REV model in PAML 3.14[19] for each TSS alignment as in[20]. The mean rate and 95% confidence intervals were calculated for each category of promoters. All alignments were masked for CpG islands and simple repeats (based upon UCSC annotation for the human and mouse genomes) before rate estimates were made as such regions are known to evolve by mechanisms other than point mutation. For each alignment, divergence was measured for the entire upstream and downstream regions as well as in the 200bp core promoter region immediately upstream of the TSS. To ensure the accuracy of these estimates alignments with fewer than 100 nucleotides were ignored. Rate estimates are presented with 95% confidence intervals.

3-n Analysis of substitutions effects of pyrimidine-purine dinucleotide start sites

Human-mouse genomic alignments of orthologous TC pairs corresponding to the SP, BR, PB and MU-classed TCs were analyzed. Using the mouse position as a template, the genomic position, the corresponding human genomic region was retrieved by using the NET alignment data from the UCSC genome browser database (assembly MM5 and HG17, respectively). Only unambiguous coordinates were considered. Having defined a corresponding location in the human genome, we identified TCs using the following criteria: (i) TC is within the corresponding region or (ii) overlaps this region totally or partially with more than 75% tag coverage. In order to ensure orthologous one-to-one mappings, only those cases where only one human TC, containing more than 100 tags, fulfilled this criteria were considered. This resulted in 2890 orthologous promoter pair alignments.

We used a sliding 2bp window [-1, +1] along the mouse and human alignments in the direction of the TC, (which then corresponds to the [-1, +1] positions of a start site). Each such window was analyzed in terms of the tag counts at position +1 using the following metric:

(#tags in mouse/#total tags in mouse) /(#tags in human/#total tags in human)

We required that the sum of tag counts in both species exceeded 30 tags, and added a pseudocount of 1 to all counts to avoid division by zero. Mutations were classified with respect to changes on purine/pyrimidine level.

3-o. Promoter-based clustering

The 127 libraries obtained from mouse CAGE data were analyzed by cluster analysis.

CAGE expression values for 159,620 TCs were measured as $\log_2$ (tags per million). We applied the K-means clustering procedure using Euclidean distance as implemented by the Eisen Cluster 3.0 program[22]. The identification of 70 clusters was decided based investigation of mean square error and application of cross-validation methodology. Within each of the 70 supergroups we applied hierarchical clustering.

CpG rich ubiquitous transcripts are clearly depleted of the AT-rich binding sites (e.g. Forkhead, HMG, and Homeobox), while tightly regulated transcripts have the opposite properties. On a more detailed level, REL and Ets binding sites are over-represented in macrophage-specific promoters.

3-p. Supergroup shape and TATA/CpG island association
Within a supergroup, we classified a TC as being peaked if the highest tag peak and its two neighboring nucleotides contained more than 80% of all tags within the TC, or broad otherwise. If the tag count of the TC was less than 20 it was not classified. CpG island and TATA association for classed TCs were measured as above.

3-q. TFBS density of TC super groups
For each TC in a supergroup, we analyzed the 300 bp region upstream of the highest CAGE tag peak in the TC. Familial binding profiles corresponding to major transcription factor classes described in [7] were used to scan this region using a 70% score cutoff [23]. The relative density of sites of a certain class on each strand was evaluated by the following metric:
(sum of sites within super group/number of TCs with super group)/
(total sum of sites within all super groups/total sum of TCs within all super groups)

3-r. Definition of alternative promoters with differential expression patterns
For each TU with multiple TCs, we evaluated the distributions of RNA libraries in the TCs using pair-wise chi-square tests for independence ($P<0.05$). The TU was defined as having at least one differential alternative promoter if at least one pair-wise comparison was significantly different.

3-s. Analysis of 3' UTR promoters
To seek evidence for specific control of such transcripts, we analyzed 1,327 transcripts that have at least 30 tags mapping to their terminal exons, and at least 300 CAGE tags mapping to their corresponding full-length transcript. The majority of these transcripts, 770 (58%), have at least 20% of tags mapping in the most distal 20% of the last exon.
The conservation of the 1kb regions centered around these prominent 3' UTR TSS was determined based upon Phastcons scores (based on cross-species comparison between human, chimp, mouse, rat, dog, chicken, fugu, and zebrafish)[24] from the UCSC genome browser database[25].

3-t. Analysis of differential expression of 3' UTRs compared to associated 5' region
Tag counts from the 5' region of representative transcripts (-100 to +100 relative to the cDNA start) were compared with the tag counts from the 3' region (the last 20% of the terminal exon). Tag counts were subdivided depending on tissue origin. In those cases where the total tag count from a certain tissue exceeded 30, we tested if the tags counts from that tissue in both regions are likely to be drawn from the same population using Fisher's exact test, corrected for the number of tests (number of tissues) by the Bonferroni method.

3-u. RACE validation of 3' UTR transcripts
RACE was performed by using oligo-capping based on the method described[26]. Calf Intestinal Phosphatase (CIP) was used to dephosphorylate truncated RNA from total RNA. Tobacco Acid Pyrophosphatase (TAP) was used to decap full-length RNA. A RNA oligo was ligated to decapped RNA. Reverse transcription reaction was performed using random primers (N6). Two nested PCR reactions were performed using gene specific primers and primers on sequence from ligated RNA oligonucleotide. The PCR products were cloned into pCR4-TOPO vector and transformed to competent *Escherichia coli*. After sequencing, originating DNA sequence was used for BLAST. The generated gff file was imported to Genome Element Viewer to display the RACE products.

3-v. Experimental validation of the promoter activity of the 3'-UTR transcripts
We selected four putative 3'-end promoters of the genes Aldoc (chr11, 77939144..77939393), Ilk (chr7, 93271482..93271731), Lass2 (chr3, 95318188.. 95318437) and Klhl5 (chr5, 63803784.. 63804033). For each of them, two distinct regions were cloned into the luciferase pGL3-Basic vector (Promega, accession number U47295) (Fig. S6E). We selected two regions, a shorter one (from -100 to +10 respect to the GGG identified starting site), and a larger region (-100 to +120 from the GGG-identified starting site. The sequences of these constructs were verified by sequencing. HepG2 cells were each cultured in minimum essential medium (Eagle) with non-essential aminoacid, and 1% non-essential amino acid, 1mM sodium pyruvate with Earle's BSS, 10% heat-inactivated fetal bovine serum (FBS), 200 U/ml penicillin and 200 mg/ml streptomycin. The reporter gene vectors (200 ng each) were transfected into either $4x10^4$ HepG2

cells in 96-well assay plates using the Lipofectamine 2000 reagent (Invitrogen). After 24 h of incubation, the luciferase activity of the reporter gene was measured with the Steady-Glo luciferase assay system (Promega). Experiments were repeated at least three times.

3-w. Determination of over-represented and under-represented TFBSs in macrophage-expressed promoters

For this purpose we analyzed 450 genes in which the promoter-usage is specifically enriched in macrophage libraries, and 295 in which promoter usage is specifically enriched in macrophages treated for 7 hours with lipopolysaccharide, the time of maximal gene induction in this system. The aim was to find the TFBSs that appear either over-represented or under-represented in these promoter groups when compared to mouse promoters in general. As a control group, we randomly selected over 40,000 TSSs out of approximately 160K. Consequently, these 40K TSSs correspond to an 'average' mouse promoter. Segments [-500, +200] relative to the peak of TSSs were selected and all promoter that had 5% or more ambiguous nucleotides within their sequences were excluded. The test set was therefore 287 and 430 promoters corresponding to induced and constitutively expressed genes, respectively, and 39,090 'average' mouse promoters. Then, we mapped all available matrix models of TFBSs contained in Transfac (Professional Ver. 7.4) database[4]. We used *minSUM* profiles for matrix models parameters, as these represent optimized thresholds for the core and matrix scores used in Transfac models[27]. The thresholds in *minSUM* are based on optimization that provides the minimum sum of false positive and false negative TFBS predictions. We used models of TFBSs from all species with the rationale that TFBSs known in other species than mouse, if found in the mouse promoters, could still represent real binding sites. In determination of over-representation we used the method presented in elsewhere[28]. We ranked all TFBS mapped to promoters based on their over-representation index (ORI) as defined elsewhere[28]. If ORI = 1 or is close to this value, then our estimate is that there is no over-representation of the motif in the target promoter groups. The results for the two promoter sets are summarized in Table S6 and S7. These tables also contain the actual probabilities of motifs as found in the target sets and in the background set obtained as follows:

P = (# of motifs found in the dataset) / (total length of the dataset).

3-x. Analysis of Ets and CAGA-site over-representation in macrophage and CNS promoters

Macrophage-specific (>80% tags in the TC originating from macrophage RNA libraries, normalized in respect to library sizes) BR-classed promoters were scanned with ETS (SPI-1 matrix from the JASPAR database with 90% cutoff, essentially capturing the core GGAA motif) and exact CAGA motif. The spacing of such sites in relation to individual CTSSs in the TC was investigated by plotting the site distribution in a 400bp window centered on the CTSS. Sites were analyzed in 10bp bins. The same procedure was used to investigate CNS-specific BR promoters and all BR promoters.

3-y. Comparison of human liver-containing CTSS conservation

We used the set of aligned promoters as described above (the analysis of substitutions effects of pyrimidine-purine dinucleotide start sites). We located all CTSS in human containing at least one tag from liver libraries. Using the same alignments, we investigated if the position and strand of each such CTSS was replicated by a liver CTSS in the mouse genome, and analyzed this rate as a function of the number of liver tags in the human CTSS. The criterion for "replication" was that the CTSS must be on same strands and the start position of the CTSSs must match within 1 bp: this is much more conservative than the commonly applied criteria of any subsequence overlap.

3-z. Calculation of the probability of the observed number of multitag CTSS

We calculated the probability of observing the actual number of CTSS with given tag density by applying the method described in figure 5 in reference[29]. The same statistical method also gives the expected occurrence. The "cluster width" was set to the minimum possible: 2bp. Since the definition of a CTSS is a single bp position, the p-values are conservative calculations.

4. Supplementary References

1. Yue, X., Favot, P., Dunn, T.L., Cassady, A.I. & Hume, D.A. Expression of mRNA encoding the macrophage colony-stimulating factor receptor (c-fms) is controlled by a constitutive promoter and tissue-specific transcription elongation. *Mol Cell Biol* **13**, 3191-201 (1993).
2. Kim, T.H. et al. A high-resolution map of active promoters in the human genome. *Nature* **436**, 876-80 (2005).
3. Kwiatkowski, D.J., Mehl, R., Izumo, S., Nadal-Ginard, B. & Yin, H.L. Muscle is the major source of plasma gelsolin. *J Biol Chem* **263**, 8239-43 (1988).
4. Matys, V. et al. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res* **31**, 374-8 (2003).
5. Rehli, M., Lichanska, A., Cassady, A.I., Ostrowski, M.C. & Hume, D.A. TFEC is a macrophage-restricted member of the microphthalmia-TFE subfamily of basic helix-loop-helix leucine zipper transcription factors. *J Immunol* **162**, 1559-65 (1999).
6. Ross, I.L., Yue, X., Ostrowski, M.C. & Hume, D.A. Interaction between PU.1 and another Ets family transcription factor promotes macrophage-specific Basal transcription initiation. *J Biol Chem* **273**, 6662-9 (1998).
7. Sandelin, A. & Wasserman, W.W. Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. *J Mol Biol* **338**, 207-15 (2004).
8. Kornblihtt, A.R., de la Mata, M., Fededa, J.P., Munoz, M.J. & Nogues, G. Multiple links between transcription and splicing. *Rna* **10**, 1489-98 (2004).
9. Yamamoto, M., Takeda, K. & Akira, S. TIR domain-containing adaptors define the specificity of TLR signaling. *Mol Immunol* **40**, 861-8 (2004).
10. Kodzius, R. et al. Absolute expression values for mouse transcripts: re-annotation of the READ expression database by the use of CAGE and EST sequence tags. *FEBS Lett* **559**, 22-6 (2004).
11. Carninci, P. et al. The transcriptional landscape of the mammalian genome. *Science* **309**, 1559-63 (2005).
12. Shiraki, T. et al. 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**, 15776-81 (2003).
13. Kodzius, R. et al. CAGE: cap analysis of gene expression. *Nat Methods* **3**, 211-22 (2006).
14. Altschul, S.F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-402 (1997).
15. Bucher, P. Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J Mol Biol* **212**, 563-78 (1990).
16. Lenhard, B. & Wasserman, W.W. TFBS: Computational framework for transcription factor binding site analysis. *Bioinformatics* **18**, 1135-6 (2002).
17. Sandelin, A. et al. Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. *BMC Genomics* **5**, 99 (2004).
18. Kullback, S. & Leibler, R. On information and sufficiency. *Ann. Math Stat* **22**, 79-86 (1951).
19. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* **13**, 555-6 (1997).
20. Gibbs, R.A. et al. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**, 493-521 (2004).
21. Blanchette, M. et al. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res* **14**, 708-15 (2004).
22. de Hoon, M.J., Imoto, S., Nolan, J. & Miyano, S. Open source clustering software. *Bioinformatics* **20**, 1453-4 (2004).
23. Wasserman, W.W. & Sandelin, A. Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet* **5**, 276-87 (2004).
24. Siepel, A. & Haussler, D. Combining phylogenetic and hidden Markov models in biosequence analysis. *J Comput Biol* **11**, 413-28 (2004).
25. Karolchik, D. et al. The UCSC Genome Browser Database. *Nucleic Acids Res* **31**, 51-4 (2003).
26. Suzuki, Y. & Sugano, S. Construction of a full-length enriched and a 5'-end enriched cDNA library using the oligo-capping method. *Methods Mol Biol* **221**, 73-91 (2003).
27. Kel, A.E. et al. MATCH: A tool for searching transcription factor binding sites in DNA

sequences. *Nucleic Acids Res* **31**, 3576-9 (2003).
28. Bajic, V.B., Choudhary, V. & Hock, C.K. Content analysis of the core promoter region of human genes. *In Silico Biol* **4**, 109-25 (2004).
29. Karlin, S. & Macken, C. Assessment of inhomogeneities in an E. coli physical map. *Nucleic Acids Res* **19**, 4241-6 (1991).
30. Galus, A., Lagos, A., Romanik, E.A. & O'Connor, C.M. Structural analysis of transcripts for the protein L-isoaspartyl methyltransferase reveals multiple transcription initiation sites and a distinct pattern of expression in mouse testis: identification of a 5'-flanking sequence with promoter activity. *Arch Biochem Biophys* **312**, 524-33 (1994).
31. Moffat, J.G., Edens, A. & Talamantes, F. Structure and expression of the mouse growth hormone receptor/growth hormone binding protein gene. *J Mol Endocrinol* **23**, 33-44 (1999).
32. Samuelson, L.C. et al. Isolation of the murine ribonuclease gene Rib-1: structure and tissue specific expression in pancreas and parotid gland. *Nucleic Acids Res* **19**, 6935-41 (1991).
33. Contente, S., Csiszar, K., Kenyon, K. & Friedman, R.M. Structure of the mouse lysyl oxidase gene. *Genomics* **16**, 395-400 (1993).
34. Ziober, B.L. & Kramer, R.H. Identification and characterization of the cell type-specific and developmentally regulated alpha7 integrin gene promoter. *J Biol Chem* **271**, 22915-22 (1996).
35. Gotoh, K., Yokota, H., Kikuya, E., Watanabe, T. & Oishi, M. Genomic structure of MUNC18-1 protein, which is involved in docking and fusion of synaptic vesicles in brain. *J Biol Chem* **273**, 21642-7 (1998).
36. Okladnova, O. et al. The genomic organization of the murine Pax 8 gene and characterization of its basal promoter. *Genomics* **42**, 452-61 (1997).
37. Zhang, L., Xiao, H., Schultz, R.A. & Shen, R.F. Genomic organization, chromosomal localization, and expression of the murine thromboxane synthase gene. *Genomics* **45**, 519-28 (1997).
38. Downes, G.B., Copeland, N.G., Jenkins, N.A. & Gautam, N. Structure and mapping of the G protein gamma3 subunit gene and a divergently transcribed novel gene, gng3lg. *Genomics* **53**, 220-30 (1998).
39. Moore, X.L., Hoong, I. & Cole, T.J. Expression of the 11beta-hydroxysteroid dehydrogenase 2 gene in the mouse. *Kidney Int* **57**, 1307-12 (2000).
40. Young, D.A. et al. Identification of an initiator-like element essential for the expression of the tissue inhibitor of metalloproteinases-4 (Timp-4) gene. *Biochem J* **364**, 89-99 (2002).
41. Youn, B.S. et al. Structure of the mouse pore-forming protein (perforin) gene: analysis of transcription initiation site, 5' flanking sequence, and alternative splicing of 5' untranslated regions. *J Exp Med* **173**, 813-22 (1991).
42. Ogawa, K., Burbelo, P.D., Sasaki, M. & Yamada, Y. The laminin B2 chain promoter contains unique repeat sequences and is active in transient transfection. *J Biol Chem* **263**, 8384-9 (1988).
43. McGrogan, M., Simonsen, C.C., Smouse, D.T., Farnham, P.J. & Schimke, R.T. Heterogeneity at the 5' termini of mouse dihydrofolate reductase mRNAs. Evidence for multiple promoter regions. *J Biol Chem* **260**, 2307-14 (1985).
44. Yoder, J.A., Yen, R.W., Vertino, P.M., Bestor, T.H. & Baylin, S.B. New 5' regions of the murine and human genes for DNA (cytosine-5)-methyltransferase. *J Biol Chem* **271**, 31092-7 (1996).
45. Killen, P.D., Burbelo, P.D., Martin, G.R. & Yamada, Y. Characterization of the promoter for the alpha 1 (IV) collagen gene. DNA sequences within the first intron enhance transcription. *J Biol Chem* **263**, 12310-4 (1988).
46. Nusse, R. et al. The Wnt-1 (int-1) oncogene promoter and its mechanism of activation by insertion of proviral DNA of the mouse mammary tumor virus. *Mol Cell Biol* **10**, 4170-9 (1990).

Figure S1:Mapping of CAGE starting sites to the genome

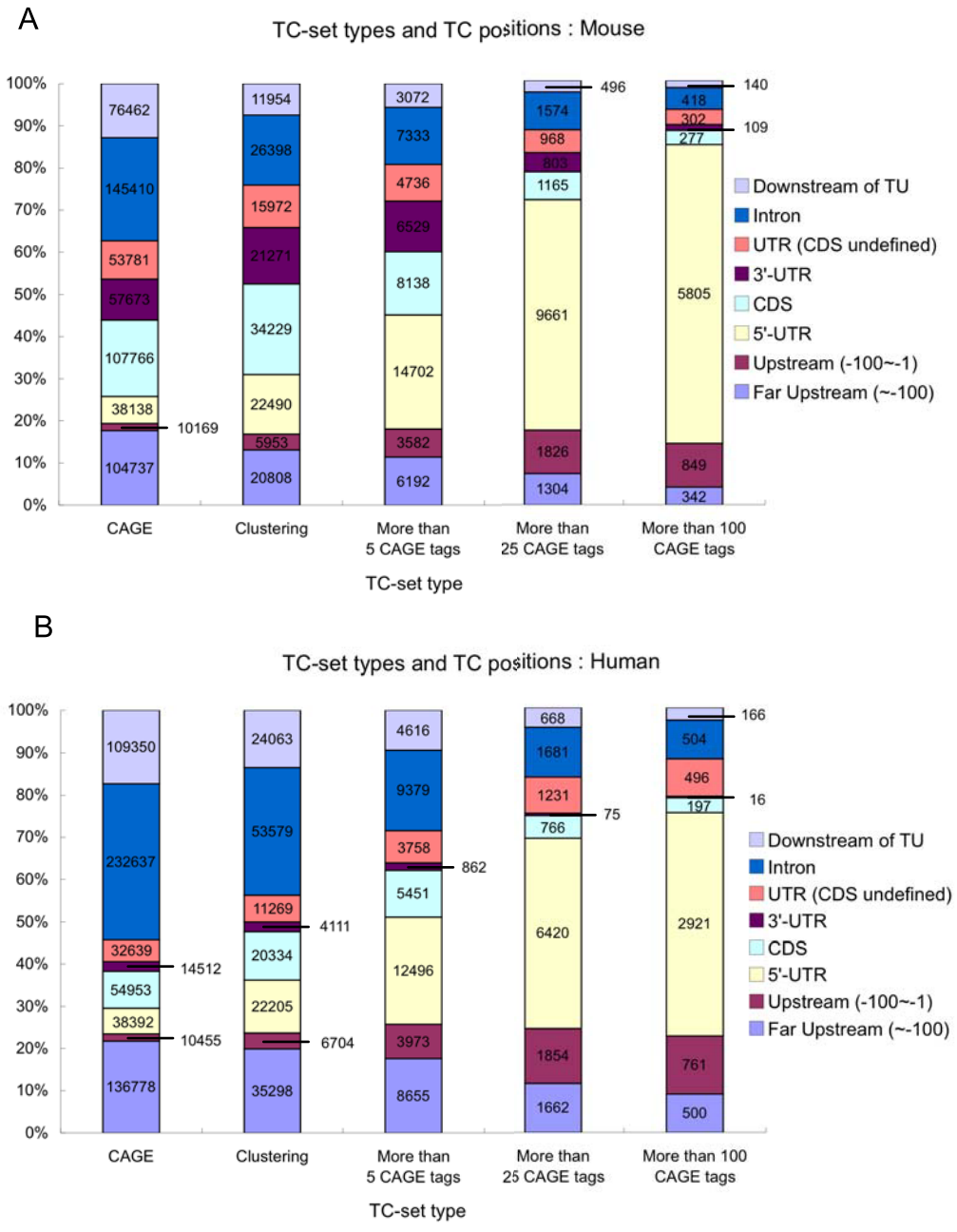


Fig. S1. Mapping of CAGE starting sites to the genome.
Partition of the CAGE tags map with respect to mRNAs for mouse (A) and human (B). From left to right, we have grouped TC with progressively larger numbers of CAGE tags. Categories are indicated on the right. -100~1, tags mapping up to 100 nucleotides upstream than 5'-ends. ~101, tags mapping more than 101 nucleotide upstream of known mRNA sequences. "CAGE" indicates all of the CTSS (one tag sufficient); "clustering" indicates cluster having at least 2 tags per TC deriving from libraries having >1500 tags.

Figure S2(A-G) Assessing exonic promoter activity:

Fig. S2A.

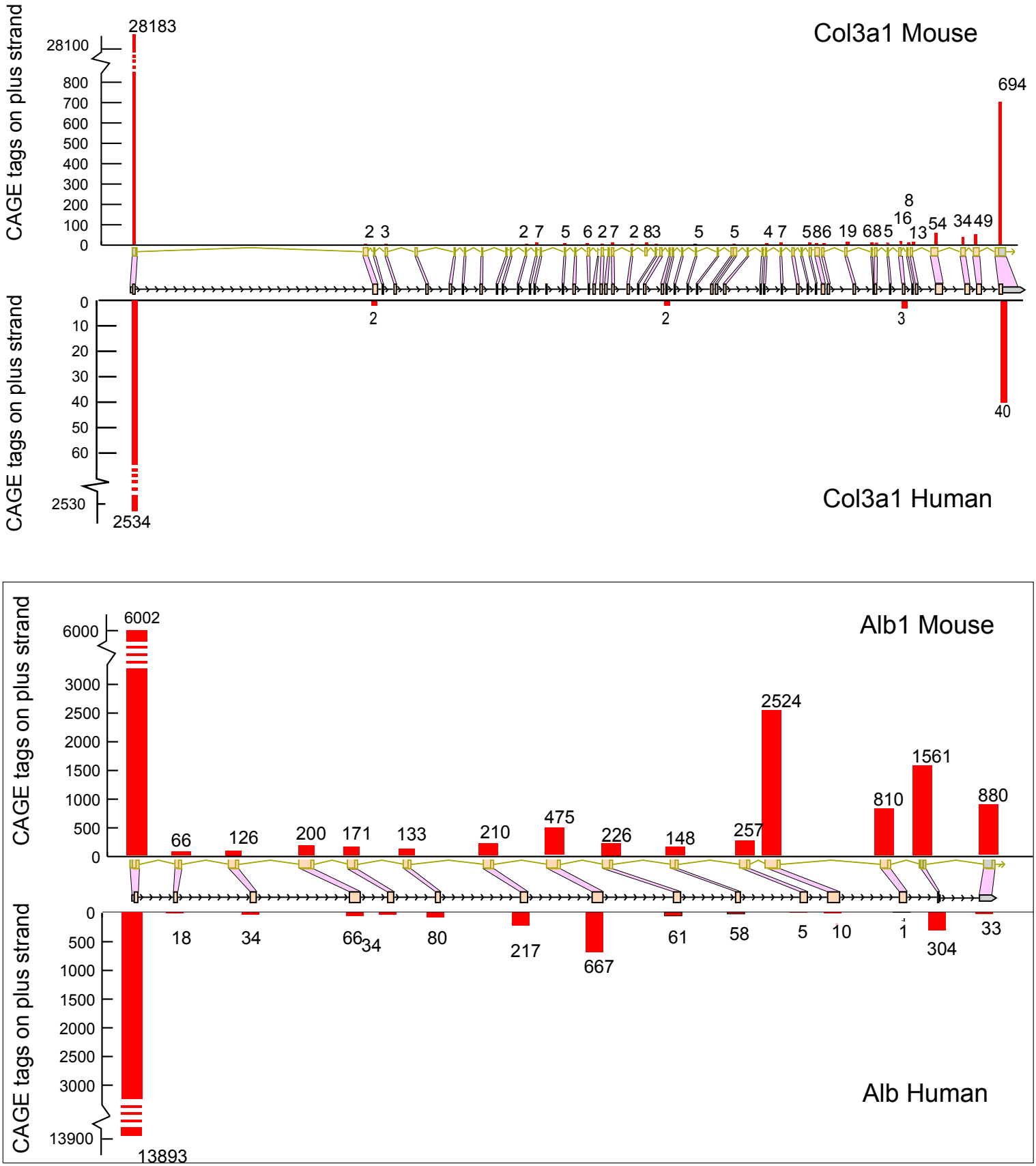


Fig. S2A. Exonic promoter activity is conserved between species.
The number or CAGE tags mapping to the exons of the Col3a1 and Alb1 genes are shown. Col3a1 is an ubiquitously expressed gene with almost no exonic promoter activity in its inner exons in mouse or human. Conversely, the Alb gene, expressed almost exclusively in liver, show high exonic promoter activity over all exons in both mouse and human. The large number of CAGE tags in exon12 in mouse Alb1 are due to a single peak promoter with over 2000 tags which is missing in the corresponding human exon. The sequence conservation between human and mouse in the proximal promoter to this peak is significantly smaller than the rest of the exon (data not shown), indicating that this promoter might be rodent-specific.

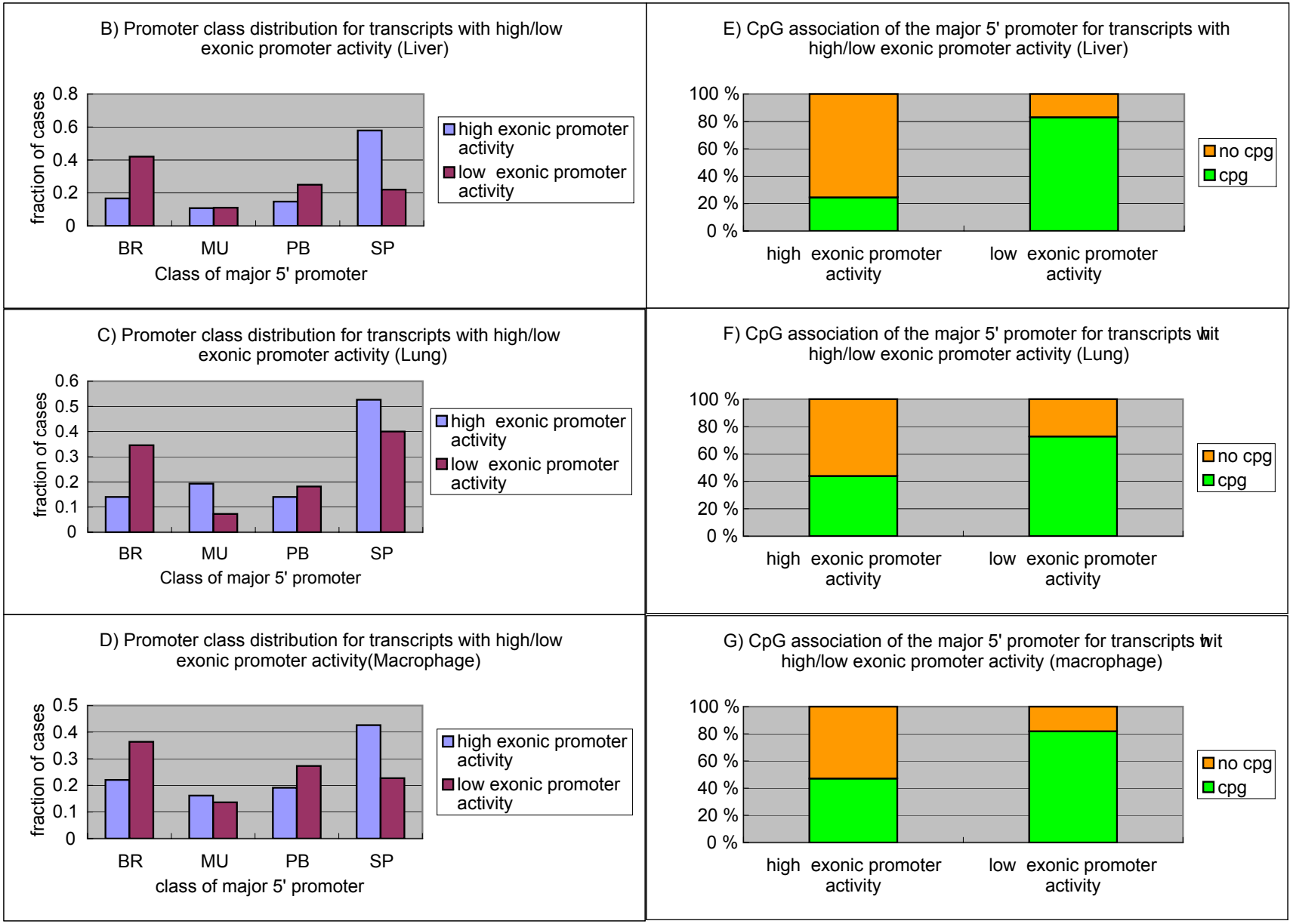


Fig. S2B Properties of genes with high/low exonic promoter activity.
The 5% most and least extreme cases of exonic promoter activity were analyzed for liver, lung and macrophage tissues. B-D) Distribution of shape classes of the major promoter of the genes in the two groups E-G) CpG and TATA-box association of the major promoters of genes in the two groups. Genes with low exonic promoter activity generally have BR-class promoters associated with CpG islands, while high exonic promoter activity is associated with SP-class promoters and lack of CpG islands.

Figure S2(A-G) Assessing exonic promoter activity: Fig. S2B-G

Figure S3 A-K : Conservation of promoters and TSS shapes over evolution

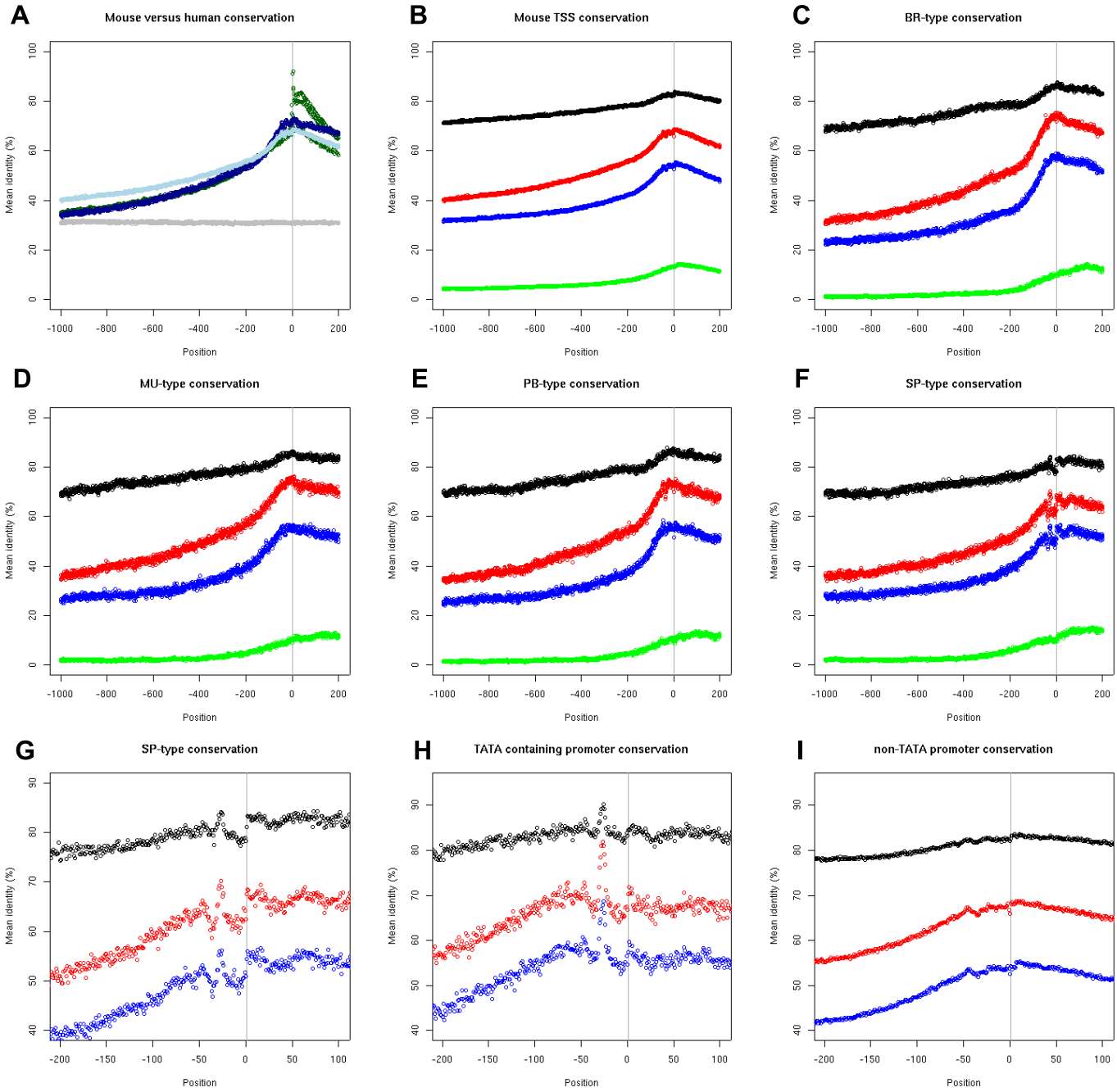
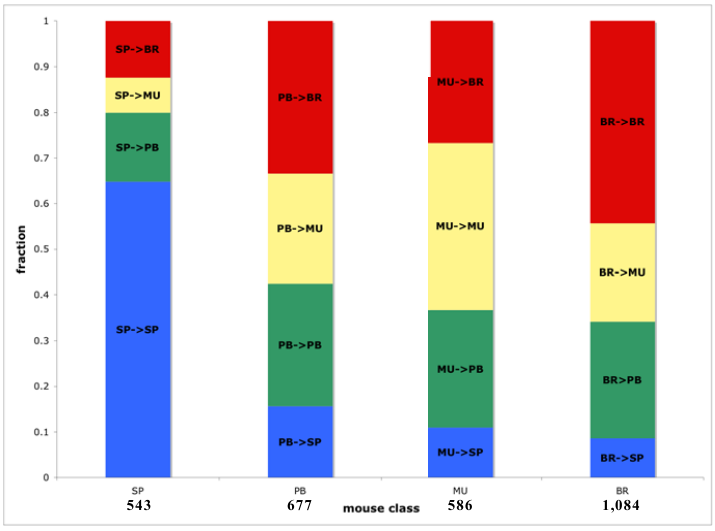


Fig. S3A-I. Average pairwise nucleotide identities for mouse

TSSs aligned to orthologous sequences. The percentage identity is calculated for each nucleotide of the mouse sequence, summing all identities across pairwise alignments and dividing by the number of alignments. Nucleotide coordinates are relative to the mouse sequence, +1 is the TSS reference position (most common tag position in TC). (A) Shows mouse to human pairwise conservation for TCs containing greater than 100 tags (dark blue), TCs containing 10 to 100 tags (light blue) and for comparison, conservation based on UCSC (http://www.genome.ucsc.org/) annotated mouse ATG-translation start sites (dark green) and randomly selected sites (gray). Trifurcation of the ATG-translation start site graph highlights the differing levels of first, second and third codon position conservation. (B-I) Conservation profiles for mouse TSSs aligned with orthologous rat (black), human (red), dog (blue) and chicken (green) genomic sequences. (B) Shows overall conservation for all TSSs with 10 or more tags in the TC. (C-F) Categories of TSS based on tag distribution within the TC. (G) Highlights the peaks and troughs of conservation most noticeable around SP-type TSSs. (H) Conservation profile for mouse TSSs with matches to the TRANSFAC TATA-box matrix (see Experimental procedures) within 29 to 35 nucleotides upstream of the TSS. (I) Conservation profile of TSSs with 10 or more tags in the TC but no TATA-box 29 to 35 nucleotides upstream of the TSS.

Figure S3 A-K : Conservation of promoters and TSS shapes over evolution

J



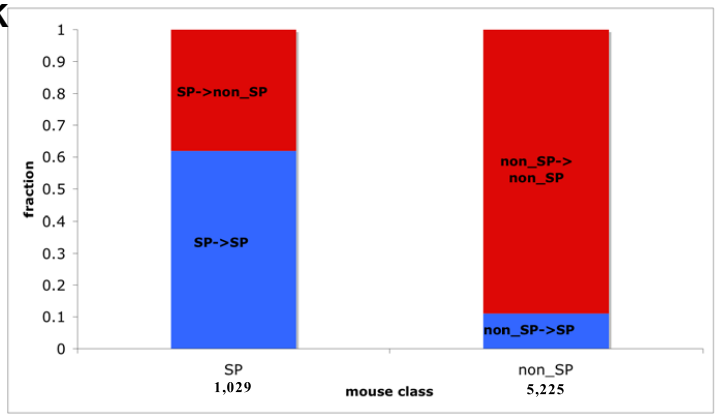
Orthologous mouse and human tag clusters classed as SP, PB, MU and BR

Species	Number of classed TCs with > 100 tags
Mouse	8,185
Human	5,532

Orthology mapping success rate				
Class	SP	PB	MU	BR
Total TC number in mouse	1,875	1,880	1,699	2,702
Number of successful mappings*	543	677	586	1,084

* not necessarily to same class

K



Orthologous mouse and human tag clusters classed as SP and non_SP

Species	Number of classed TCs with > 30 tags
Mouse	15,848
Human	12,207

Orthology mapping success rate		
Class	SP	non_SP
Total TC number in mouse	3,984	11,864
Number of successful mapping*	1,029	5,225

* not necessarily to same class

Fig. S3J-K. Core promoter shapes are retained over evolution.
We measured the fraction of retained and changed promoter shapes in orthologous mouse-human promoters using either four promoter classes (SP, PB, MU, and BR) (J), or two classes (SP and non-SP)(K). All changes are relative to mouse promoters. Summary tables for each analysis are shown in the right panel.

Figure S4 A-H : Initiation site properties and evolutionary changes

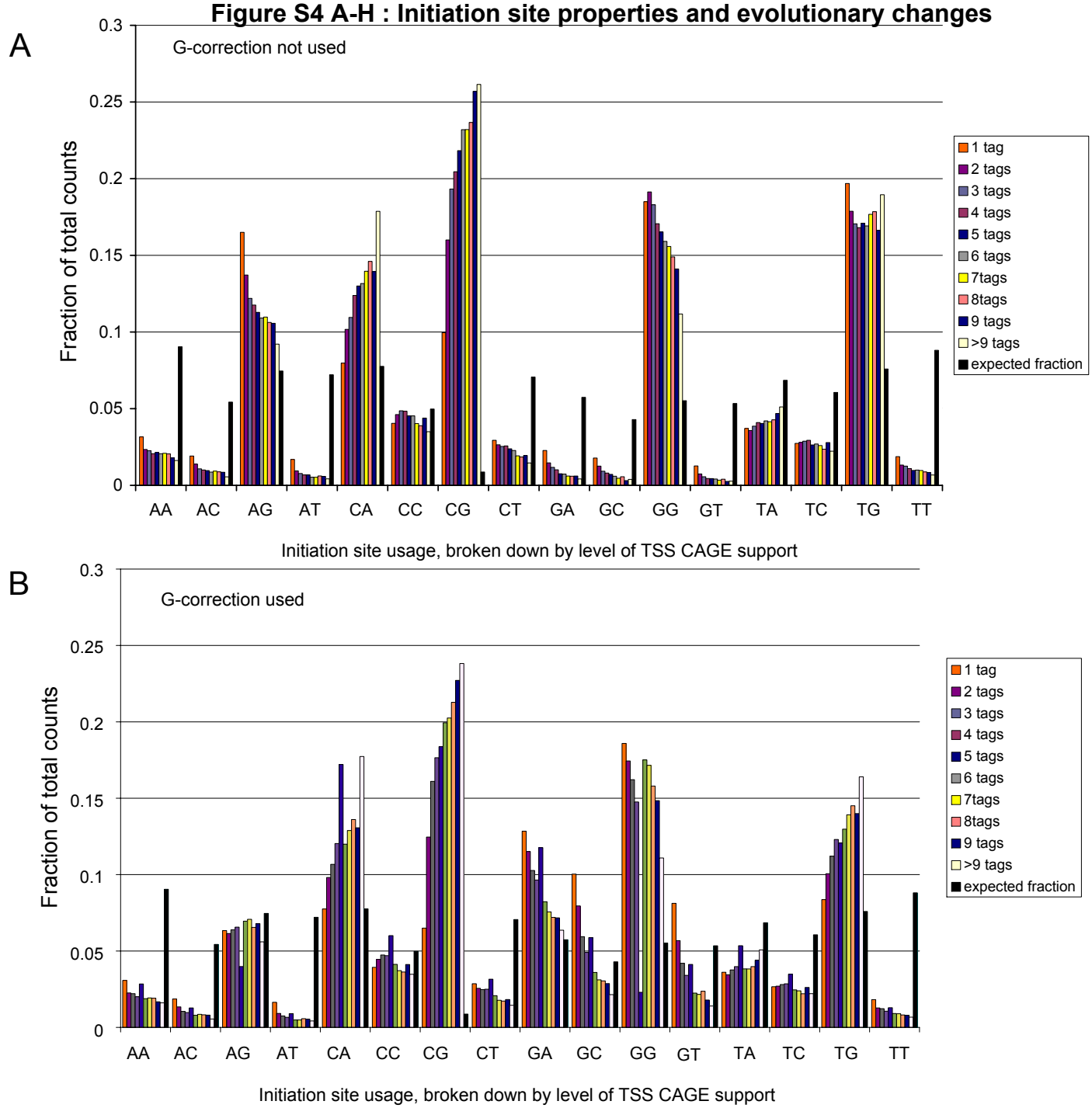
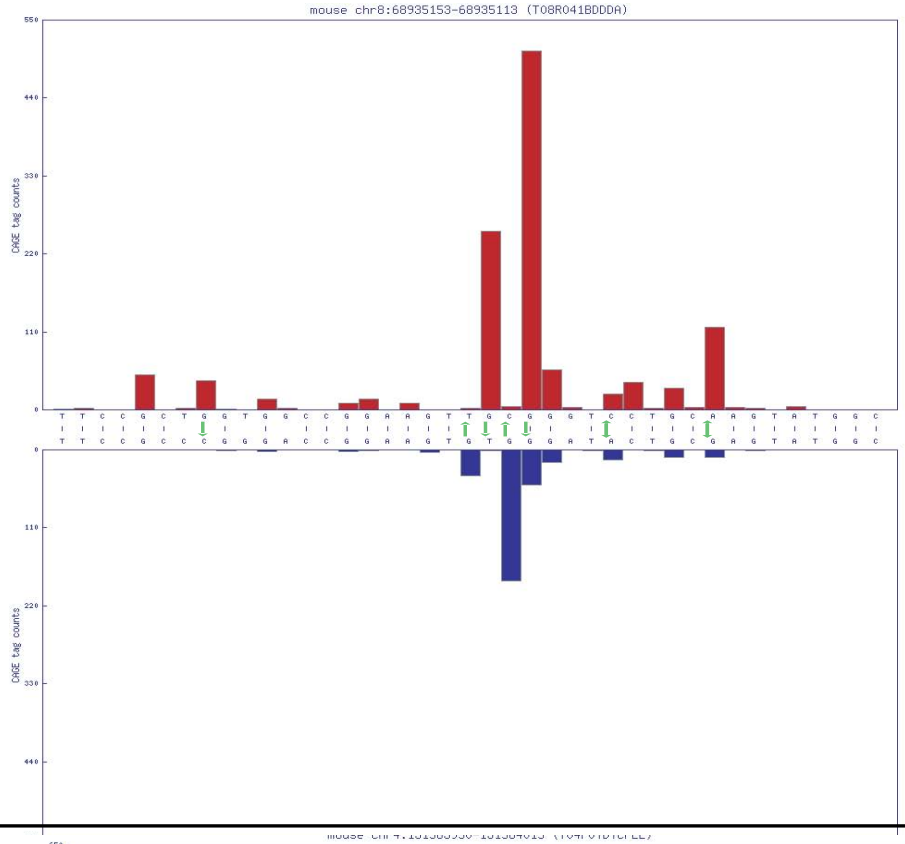


Figure 4 A-B. Dinucleotide distribution analysis of CTSS with varying CAGE tag support

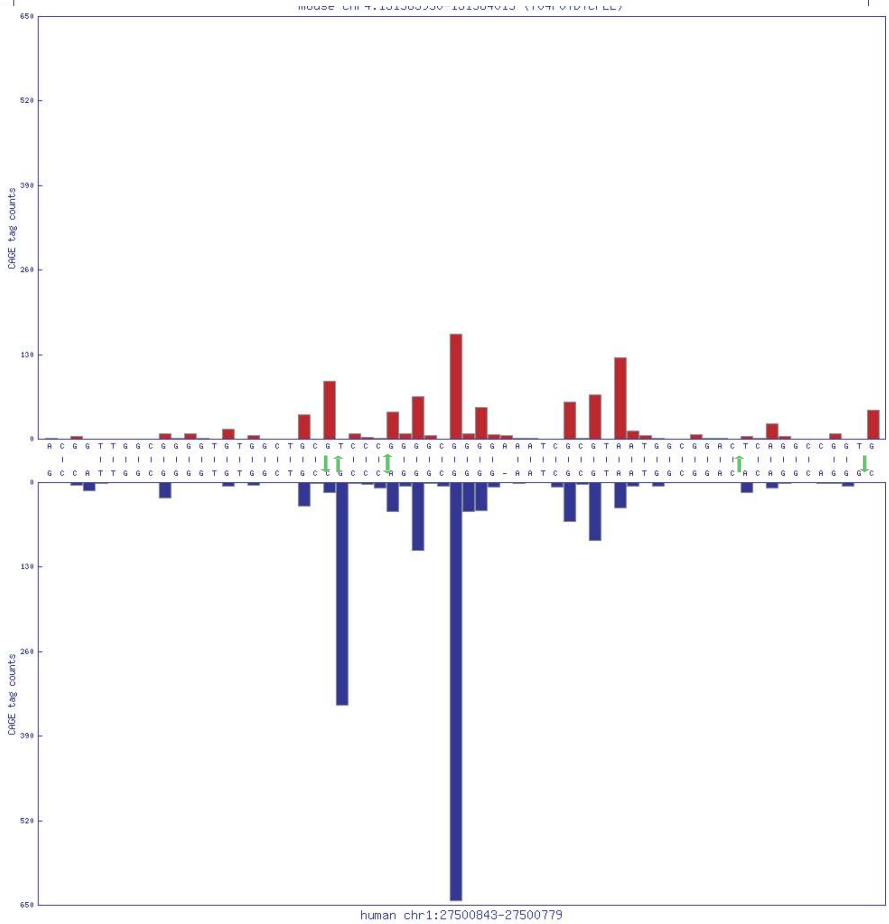
We analyzed the usage of different [-1, +1] dinucleotides relative to each CTSS in the data set (note that the -1 nucleotide is not part of the sequenced tag). We subdivided the cases in respect to how many tags the CTSS contained into 10 classes (1,2,3 to 9 tags and 10 tags). As an additional reference class, we collected 10.000 randomly selected start points in the genome (non-overlapping and not part of repetitive regions). This distribution will correspond to the expected distribution if start sites are random (noise). The frequency of all possible dinucleotides for the 11 classes is shown as a barplot, with (panel B) or without G correction (panel A). The dinucleotide distribution is dramatically different from random selection, even with single CAGE tag support. We also note that there is a higher preference for INR-like CA dinucleotides when the transcript has a higher expression (i.e. more tag counts), while AG and GG dinucleotides are more favored in rarely expressed transcripts. Part of the GG dinucleotides corresponds to the GGG motif (before G correction) we found for the novel 3'UTR transcripts.This is true regardless of whether the CTSSs are subjected to G correction or not. The difference in dinucleotide use when the tag count is 5 is a rounding artifact in the G correction algorithm (which was designed for correcting larger tag counts). Regardless of this, the overall frequency pattern as a function of number of supporting tags is indicative of very low level of noise in the CAGE dataset: otherwise the preference for TSSs supported by one tag (singletons) would be much closer to that expected by chance, and different from the preference of TSSs supported by two or three tags.

Figure S4 A-H : Initiation site properties and evolutionary changes

Fig. S4C-D Examples of pyrimidine-purine dinucleotides substitutions and effects.
Gallery of barplots of mouse and human orthologous TCs illustrating dinucleotide substitutions and their effect on the start site usage. Y-axis indicate the number of CAGE tags starting at given genomic positions(X axis). Green arrows indicate the transition from a pyrimidine-purine start site to any other base combination.

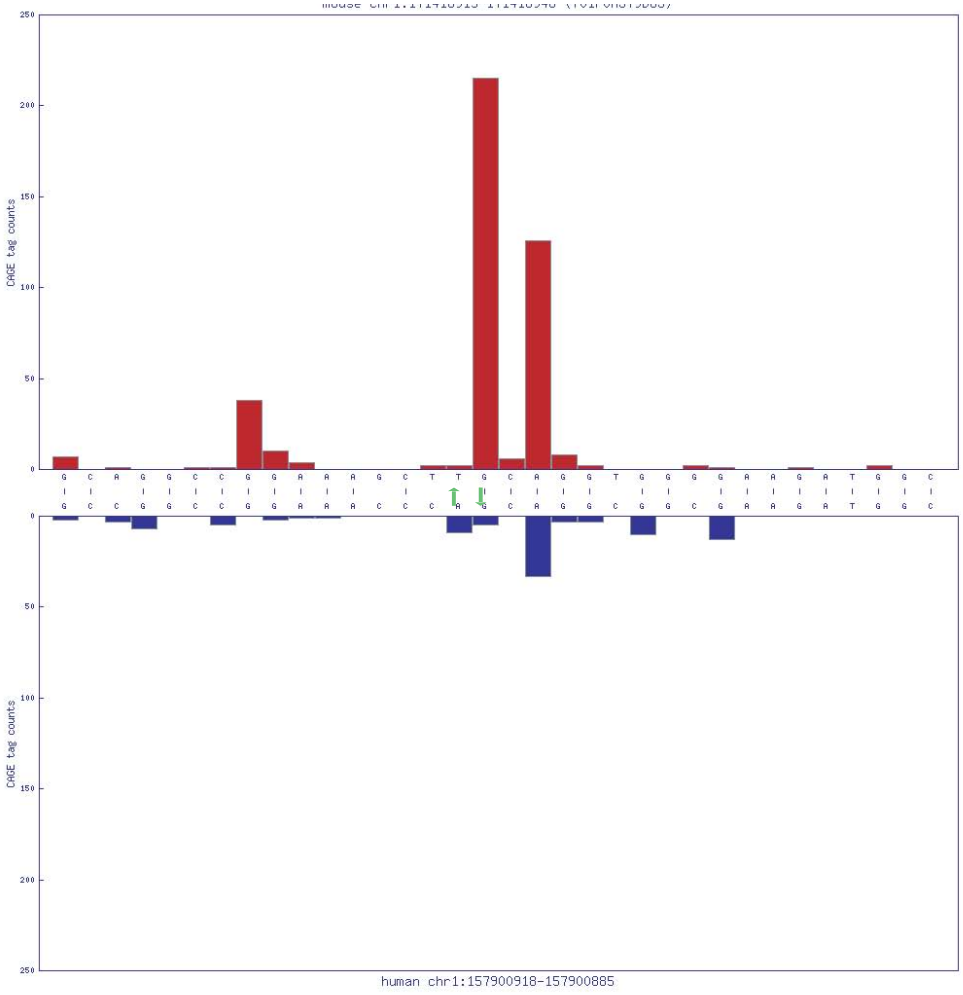


C
Ccm1 gene
Tag cluster T05F0003A1FA6

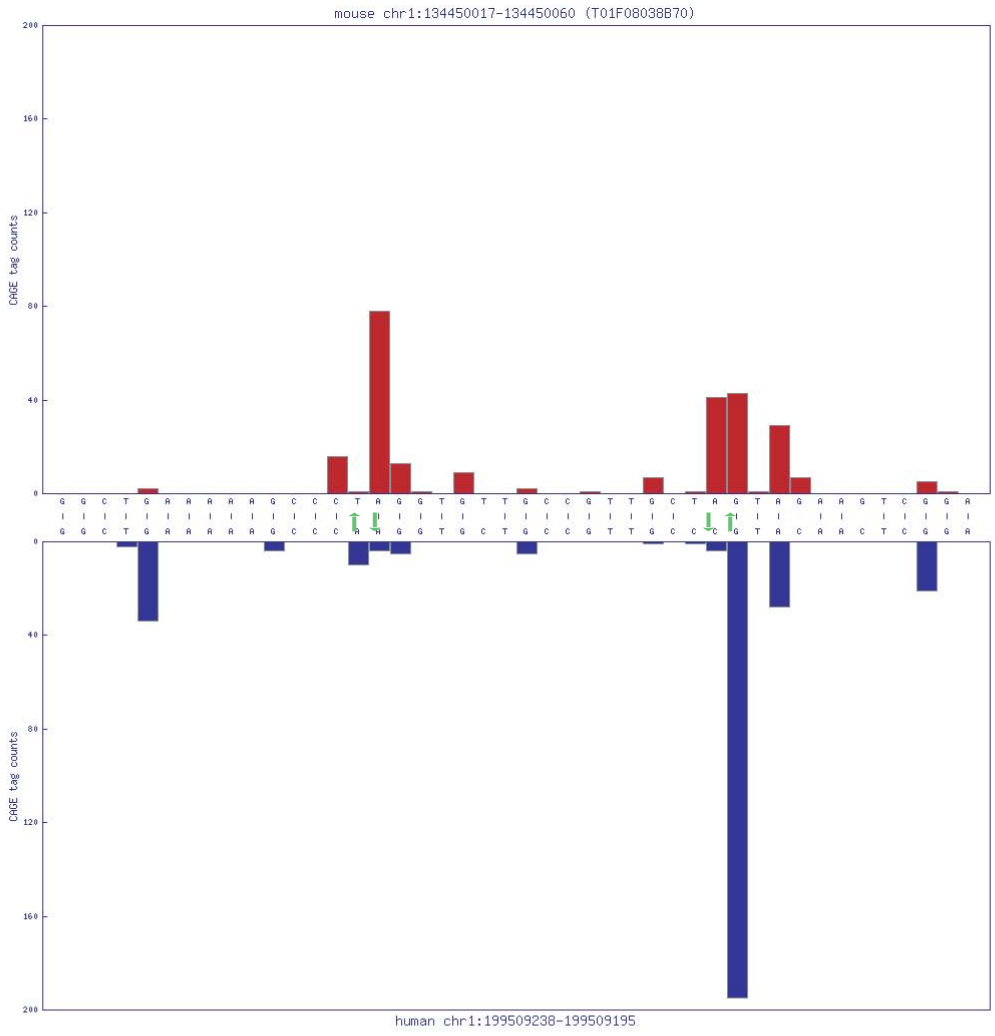


D
Wasf2 gene
Tag cluster T04F07D7XFEE

Figure S4 A-H : Initiation site properties and evolutionary changes

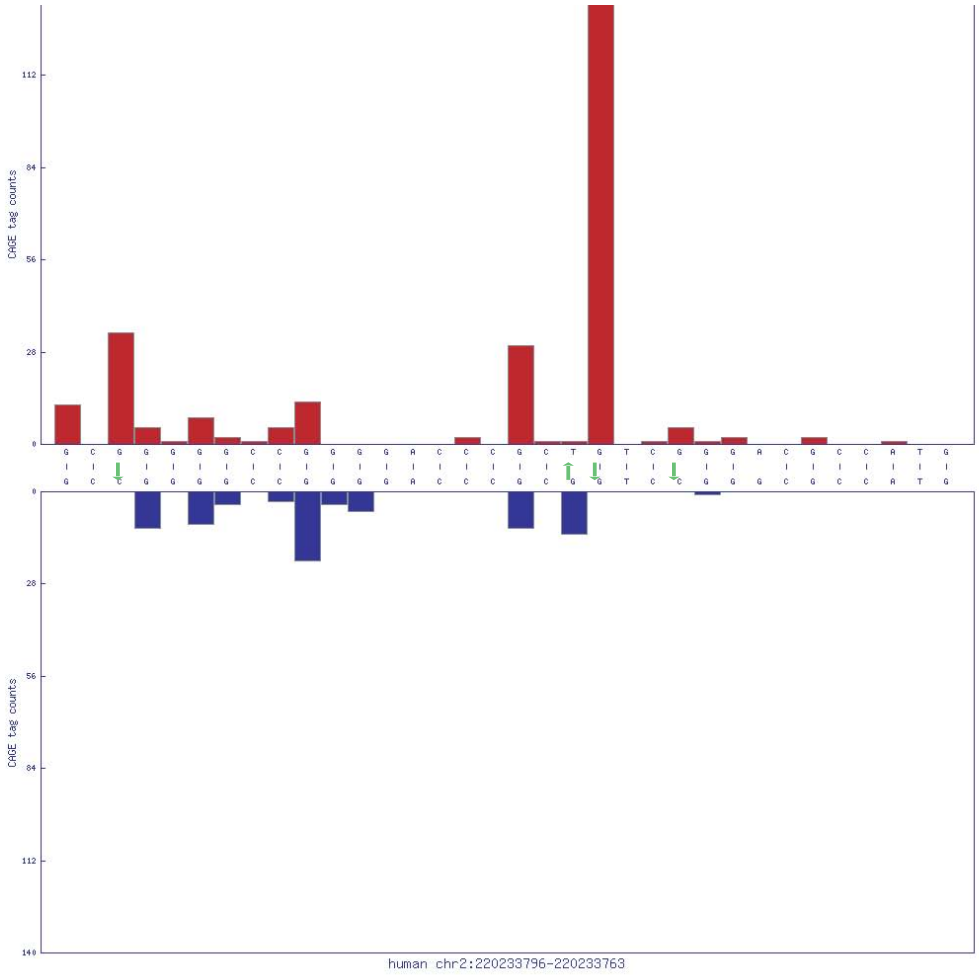


E
Pfdn2 gene
Tag cluster T01F04A379D63

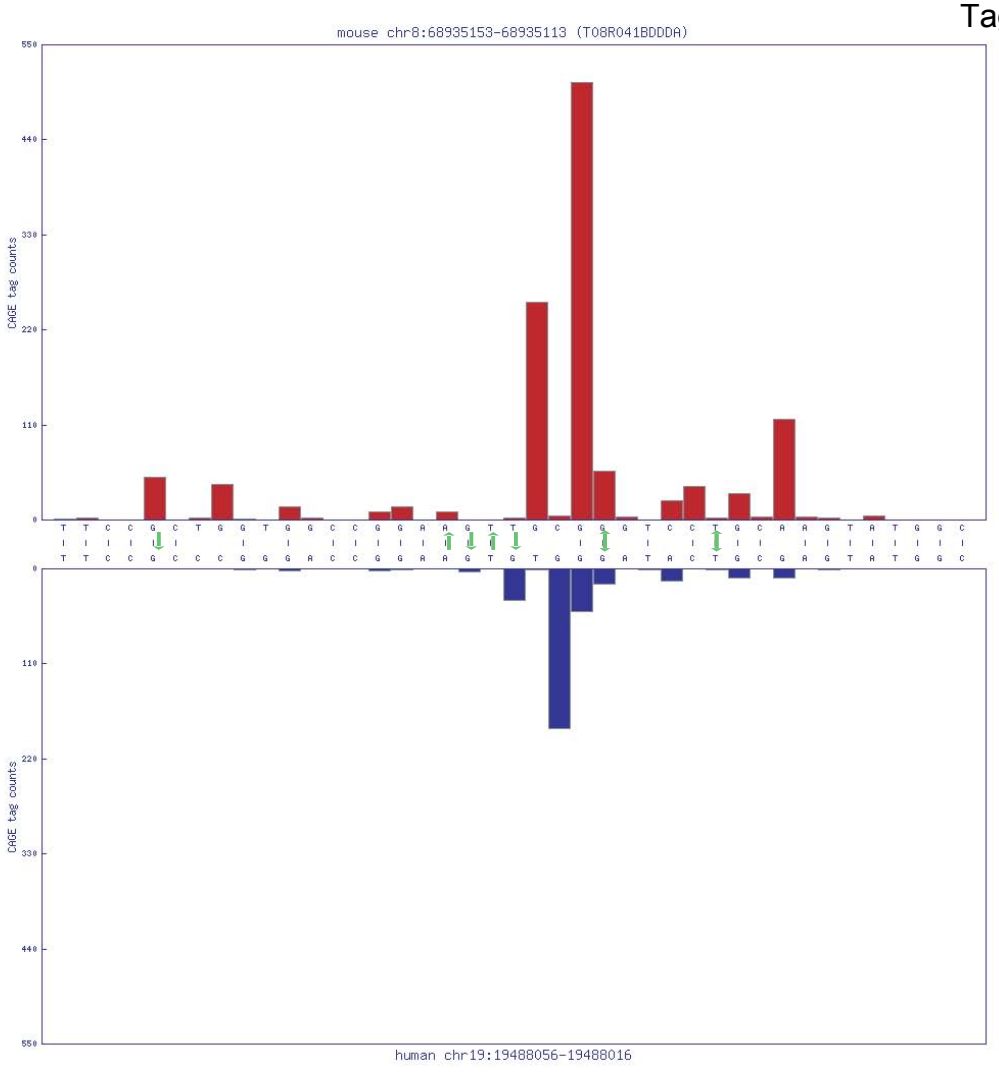


F
Jarid1b gene
TagclusterT01F08038B70

Figure S4 A-I : Initiation site properties and evolutionary changes



G
D1Bwg1363 gene
Tag cluster T01R048684BF



H
Grim19 gene
Tag cluster T08R041BDDDA

Figure S4 A-I : Initiation site properties and evolutionary changes

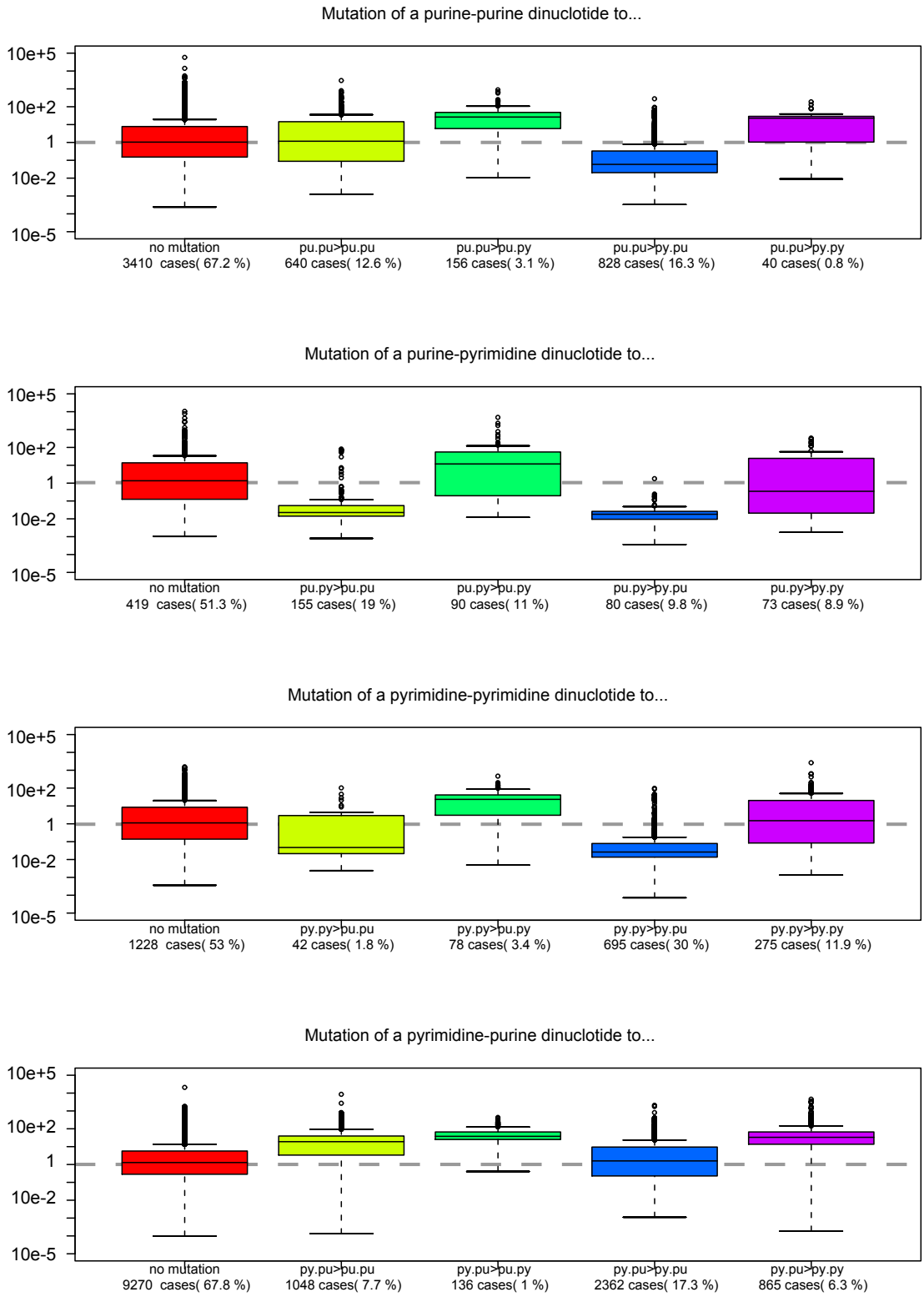


Fig. S4l Substitution effects on dinucleotides in core promoters.

Boxplots show the effects of substitutions on initiation sites for all possible base combinations. Mutations are annotated relative to mouse (i.e. mouse to human). Boxplot generation and Y axis score is described in Methods. The four sections correspond to four different reference dinucleotides (Pu-Pu, Pu-Py, Py-Pu, Py-Py).

Figure S5A-D: Sequence pattern distributions for different classes of promoters

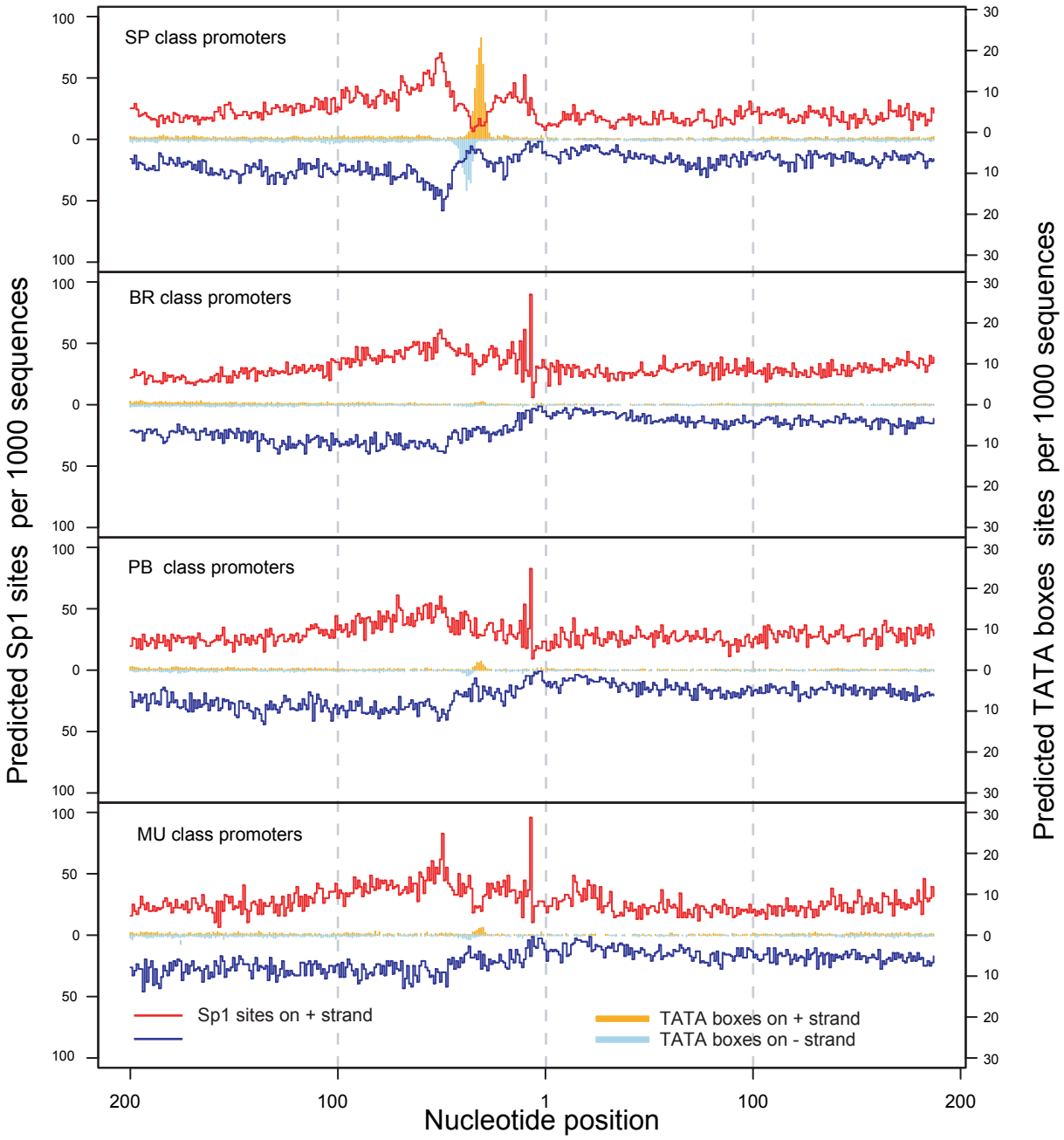


Fig. S5A Distribution of TATA and SP1 sites for different shape classes of TCs.
The [-200,+200] sequence region was extracted around each CTSS for a given TC. The density of detected TATA box and Sp1 motifs was obtained by scanning sequences with the corresponding weight matrix profiles from the JASPAR database. Densities on (+) strand are plotted on positive half of y-axis and densities on (-) strands on its negative half. SP-type TCs exhibit a typical localized preference for Sp1 and TATA boxes as previously described15. BR-type TCs have an increased density of putative Sp1 sites around -50 relative to TSSs, but the spacing is not well defined. PB and MU classes share general properties with the BR class, but with a significantly higher incidence of TATA boxes, in accord with the notion that these are ambiguous or mixed cases. A peak of Sp1 at the position -4 relative to TSS of broad type of promoters is a secondary effect due to a pyrimidine:purine preference at [-1,+1], which matches positions [4,5] of the Sp1 consensus sequence.

Figure S5A-D: Sequence pattern distributions for different classes of promoters

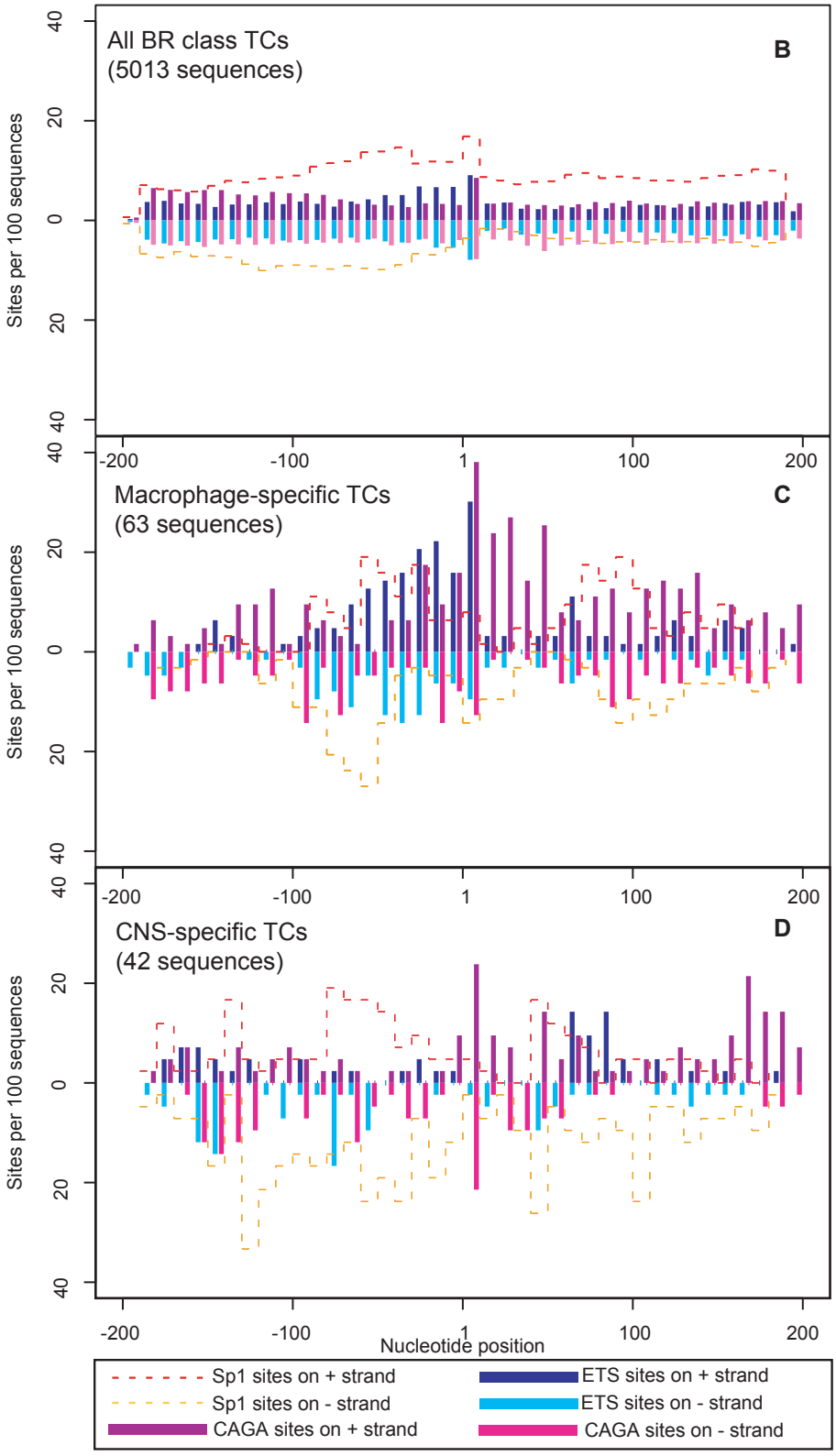


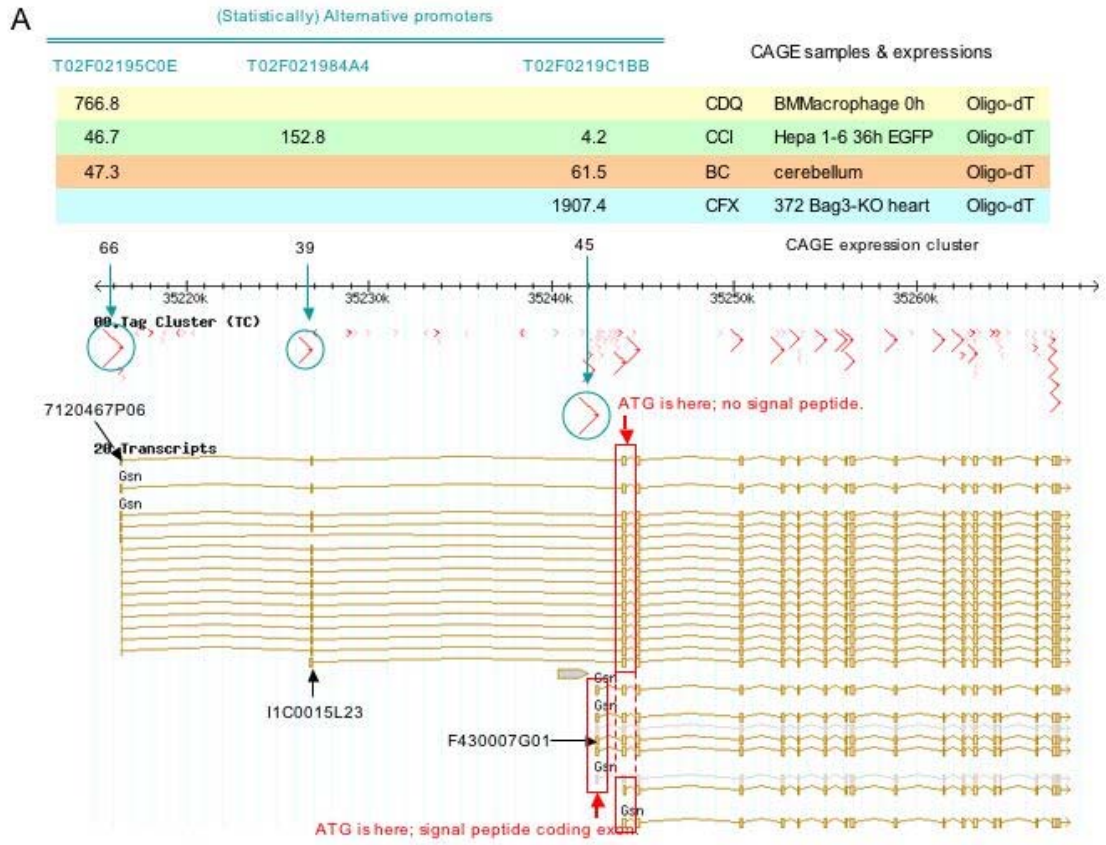
Fig. S5B-D Macrophage-specific promoters are characterized by high density of core Ets (GGAA) and CAGA motifs.
The promoter sequence (one per CTSS, relative positions -200 to +200) were aligned at the corresponding TSS positions. Densities of Sp1 elements, GGAA (core Ets) motifs and CAGA (presumed Ewing sarcoma protein binding motif) were counted in 10bp bins for (+) and (-) strand separately (See Methods). The density of the three types of binding sites are not significantly different between the set of all BR-type promoters (B) and CNS-specific BR-type promoters (D), the latter being more variable due to smaller sample size. Macrophage specific promoters (C) show a strong overrepresentation of Ets elements upstream of TSS, and a less pronounced, but significant overrepresentation of CAGA elements downstream. (CAGA sites at position -1 are overrepresented in all three cases because CA matches Py-Pu consensus of the initiator sequence).

Figure S6 A-E:Alternative promoters and transcription start sites in 3' UTRs

Fig. S6A-B Context-driven alternative promoter usage in the Gsn gene.

A) Expression and genomic overview of Gsn (gelsolin) primary three alternative promoters. 1. T02F02195C0E (left red angle with blue circle in "Tag Cluster" track, potential transcript 7120467P06 in FANTOM3 clone) 2. T02F021984A4 (center, potential transcript I1C0015L23) 3. T02F0219C1BB (right, potential transcript F430007G01). Colored rows describe the expression level (TPM normalized) in characteristic four RNA libraries. Red rectangles describe the 'ATG' (translation start site) containing exons. In case that transcription starts from the promoter 1 and 2, these transcripts do not contain the signal peptide coding exon. B) Detailed clustering based on the heatmap (Fig. 5) showing alternate Gsn promoter usage. Each of the clusters display detailed promoter usage for each TSS and the related annotated TUs, deriving from the parental supergroups (66, 39 and 45). 02F02195C0E is in the CAGE expression cluster No. 66, T02F021984A4 is 39, T02F0219C1BB is 45 respectively. We clip the most similar 6 promoters in each Gsn alternative promoters.

The Gsn gene has two alternative promoters (T02F02195C0E and T02F021984A4) potentially producing the same protein product, and a third alternative promoter, T02F0219C1BB, which directs a distinct 5'UTR encoding not only a distinct methionine but an N-terminal signal peptide permitting protein secretion. Although the T02F02195C0E and T02F021984A4 have the same cytoplasmic protein product, they belong to different TC supergroups. The promoter T02F02195C0E is in the same cluster as the core promoter of the vimentin gene (Vim), an intermediary filament, and is the dominant form of Gsn expressed in macrophages. Conversely, the promoter T02F021984A4 is in the same supergroup as the laminin beta 3 gene (Lamb3), which is a part of the basement laminins, and is the main Gsn promoter in liver cells (Hepa 1-6). We infer that T02F02195C0E is used in cellular contexts where the Gsn and Vim need to be coexpressed while T02F021984A4 is used when Gsn and Lamb3 are needed. The secreted protein product of the third alternative promoter, mainly found in cerebellum and heart libraries, encodes a plasma form of gelsolin, which has a potential role to solubilize actin molecules derived from damaged cells to prevent thrombosis.



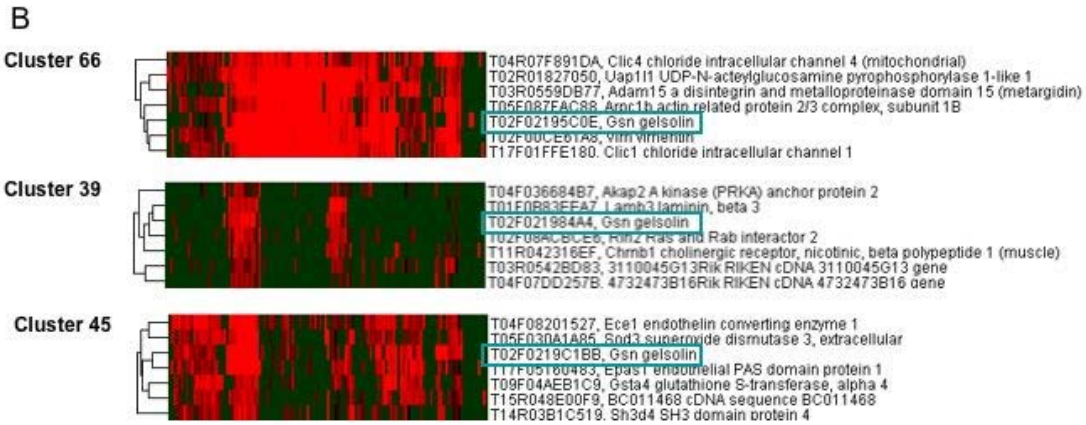


Figure S6 A-E:Alternative promoters and transcription start sites in 3' UTRs

Fig S6C Validation and analysis of 3' UTR transcription

RACE validation of the 3' UTR transcription of the A130090K04 (Oprm1 locus).

Genome analysis viewer screenshot of the RACE validation of the 3' UTR transcriptions reveals a large number of TSSs. The upper panel shows the TCs and their tag support level. The panel below shows the location of the 3' UTR of the transcript A130090K04, as well as two shorter cDNAs starting inside the 3' UTR region. Coding regions are marked in orange, UTRs in grey. The central panelshows that multiple EST sequences (colored in beige) start in the region corresponding to the 3'UTR. Five sets of RACE primers designed from the sequences corresponding to the TCs (not shown) produced ~25 discrete race products over the 3.5 Kb region, suggesting multiple initiation sites within this 3' UTR (bottom panel). The number of RACE products initiating in each position is indicated.

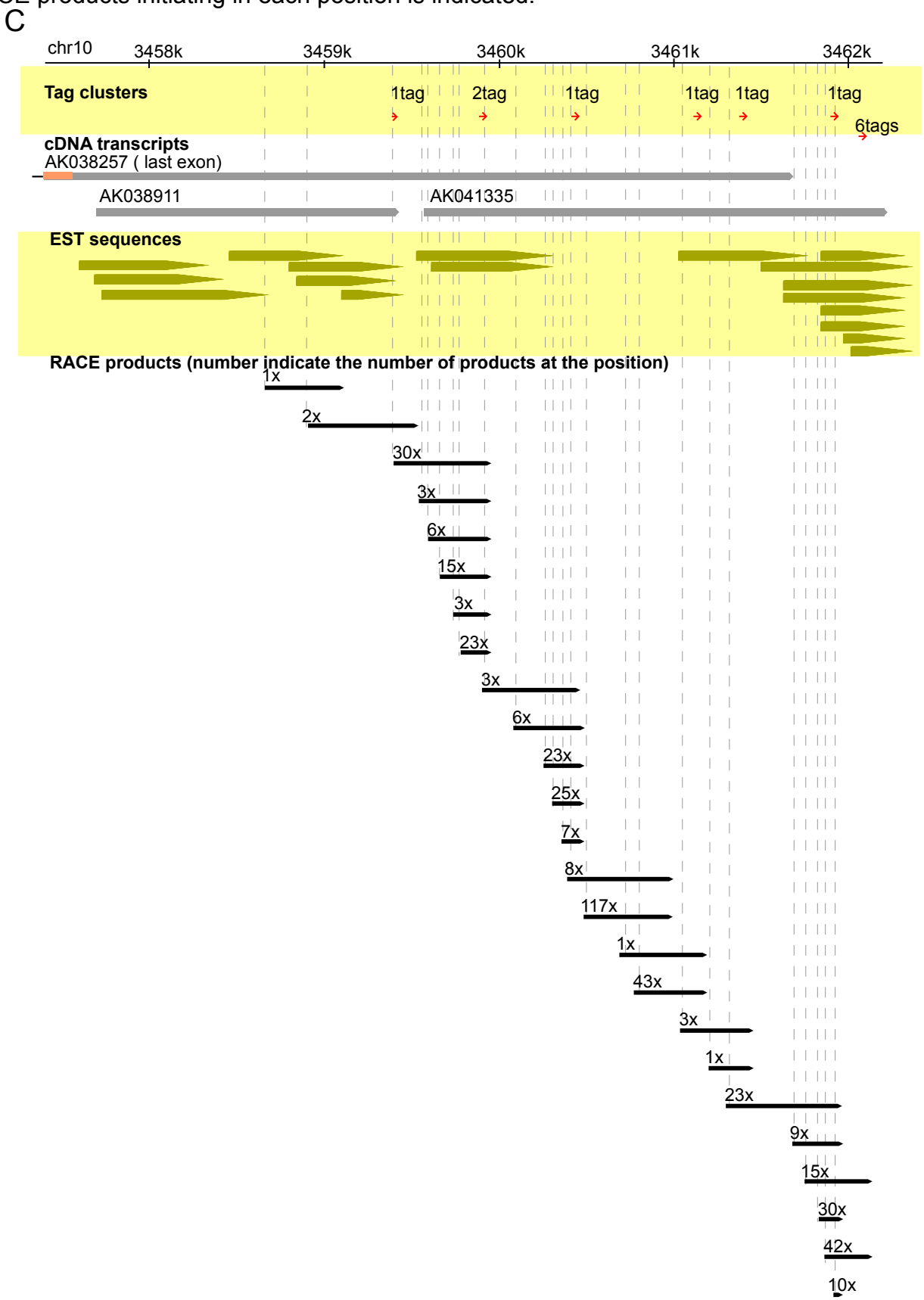


Figure S6 A-E:Alternative promoters and transcription start sites in 3' UTRs

D)

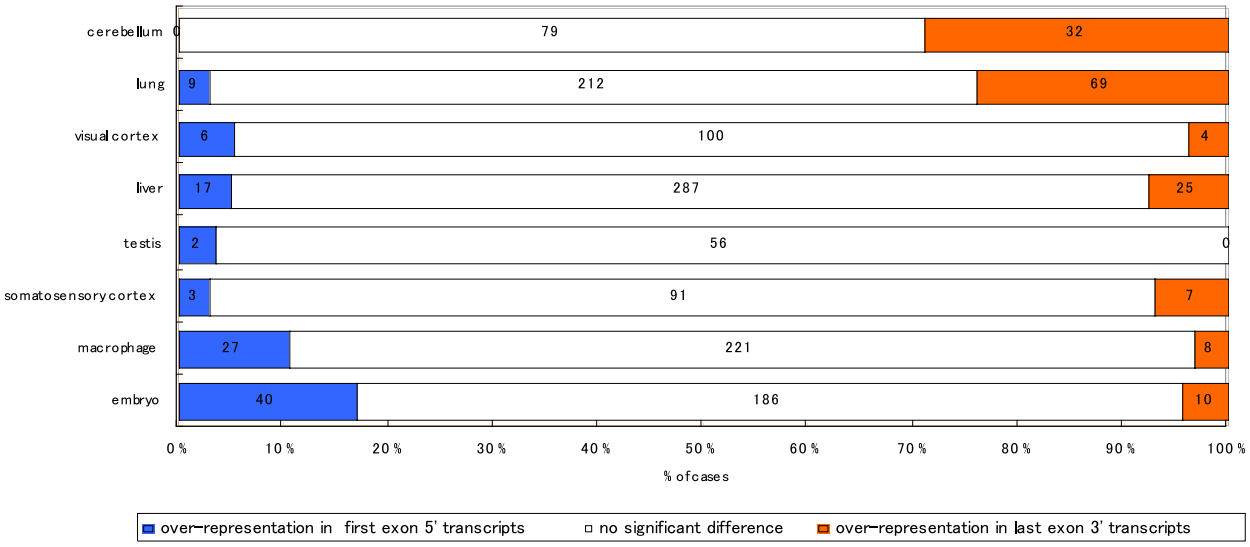


Figure S6D Tissue proportionality of transcripts in 5' and 3' ends of genes.
Numbers of significant discrepancies in tissue distribution between tags from the 5'-end of representative transcript vs 3' UTR promoters. Note that the skew in the discrepancies varies between tissues.

Figure S6 A-E:Alternative promoters and transcription start sites in 3' UTRs

E)

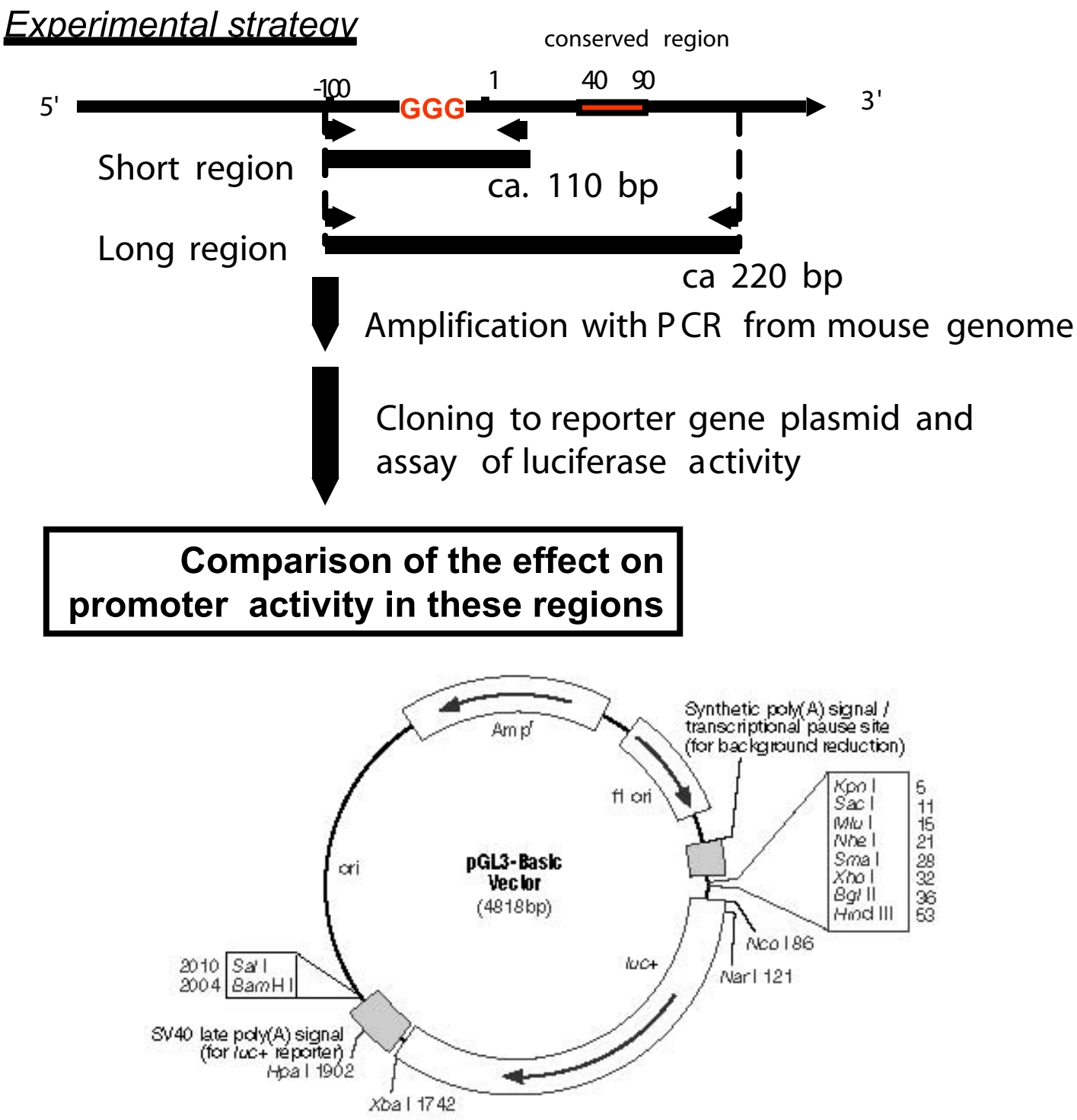


Figure S6E Reporter gene construct for validation of 3' UTR promoters
The upper part contains a schema of the reporter assay construct. Short regions and long regions represent two different regions that were subcloned and used for the reporter assay. Bottom, a schematic representation of the vector pGL3-Basic vector.

Figure S7A-U: CAGE validation examples

Fig. S7 A-U Comparison with TSS analysis annotation from literature sources

We randomly selected 19 articles encompassing 70 defined and annotated TSS using different methods (S1 nuclease protection, RNA protection assay (RPA), RACE and primer extension). The sites were mapped to the mouse genome using BLAT. In general, CAGE can confirm most sites defined by any of the methods; however, it is important to realize that a total correspondence is not expected due to fundamental methodological reasons, including i) the experimental setup is different; in particular certain TSS may be used in the specific cell lines/tissues used, which often do not overlap with our libraries. The depth of the CAGE data is many orders of magnitudes higher than the single experiments reported here, which should capture and quantify many more rarely used sites, ii) in many cases the method (especially RNase protection assay) does not provide accurate base pair-level assignment when mapped on the promoter. In fact, when CAGE data indicates many TSS, gel-based methods often indicate many weaker, broader start sites, which are not reported in corresponding papers. Promoters are indicated from 5' to 3' regardless of the directionality of the gene location in the genome. The chromosomal location is indicated above each of the panels A-S, and corresponds to the UCSC mm5 genome assembly. Red bars in the CAGE diagram indicates CAGE tags on the same strand as those TSS reported in the article in question, while blue bars indicate tags on the opposite strand.

A) PCMT[31]. The Protein Carboxyl Methyltransferase start site was mapped with both RACE and RPA. The RACE overlaps all three main peaks obtained by CAGE, while both the RPA points map several bases upstream to both the CAGE and the RACE data main peaks, suggesting potential problem in determination of length of the protected fragments. With long exposure of the RPA assay[31], minor shorter TSS became evident, which would correspond to downstream TSSs indicated by CAGE.

B) GHR/BP[32]. Mouse Growth-Hormone receptor/growth hormone-binding protein 5' UTR was determined by S1-Nuclease mapping. The main peak is accurately mapped, as well as the second strongest peak one (originating the shortest transcript), which is one nucleotide off. However, the correspondence is lower for most upstream TSS, and CAGE identifies more upstream TSS. In reference[32], additional weaker TSS are found but not annotated by the authors, including an upstream TSS transcript which would correlate with the CAGE most upstream. The authors used only liver mRNA, which may explain the absence of some TSSs.

C) RNAse1[33]. Annotation of the gene of ribonuclease gene Rib-1, where there is perfect match of the TSS determined by primer extension and the main CAGE peak. CAGE shows a few other shorter start sites, which are detected but not annotated in the publication[33].

D) LOX[34]. The mouse lysyl oxidase promoter was determined by RNA protection assay[34]. One of the two confirmed cap sites agree with the smallest tag cluster, while the other identifies a minor start but maps in proximity of another broad initiation site. The paper shows additional minor unannotated sites, which agree with a widespread distribution of start sites obtained by CAGE.

E) ITG7A[35]. Alpha-7 Integrin Primer extension and S1 protection show two main starting sites, which were annotated[35]. One of them matches perfectly with the CAGE tags main peak, the other is shorter. CAGE analysis shows other minor starting sites; in the report[35] only myoblasts and myotubes were used.

F) Munc18[36]. The Munc-18 TSS, as determined by CAGE, is not in agreement with the

Figure S7A-U: CAGE validation examples

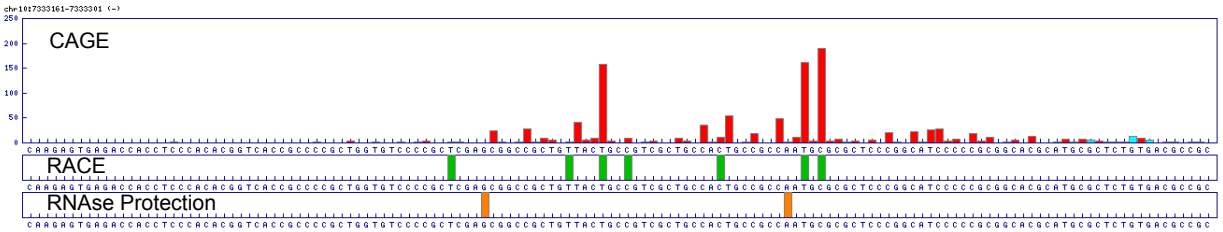
TSS determined by S1 nuclease assay[36]. S1 nuclease assay, determined with brain RNA, seems to identify a longer transcript than CAGE data, with the exception of a minor CAGE determined TSS.
G) PAX8[37]. CAGE tag coverage of this developmental gene is very low because of its overall low expression level. The two TSSs lie between two of the four sites determined by S1 nuclease and primer extension, all within 6 bp in the genome. The longest were not detected by CAGE. In the[37] a cocktail of RNA was used.
H) Tbxas[38]. The thromboxane synthase gene most upstream TSS matches perfectly the upstream CAGE tag. The most downstream RPA-determined TSSs map within other CAGE-TSS, but the match frequently in disagreement by a few bp. CAGE identifies additional downstream TSSs.
I) G-protein gamma3 subunit[39]. The two genes Gng3 and Gng3lg map head to head and share a bidirectional promoter, and Gng3lg (+ strand) has a downstream different core promoter. The Gng3 RPA mapping starts relatively close (6bp) downstream of the main CAGE peak. CAGE tags do only partially match the RPA and do poorly correlate with the primer extension, probably due to fact that only brain and testis were used[39]. The presence of TSSs on both strands strongly confirms the bidirectional nature of this promoter.
J)11beta-hydroxysteroid dehydrogenase 2 (HSD11B2)[40]. The RPA assay detects a TSS about 10 bp downstream of the CAGE-detected TSS, however the band in the gel is quite broad and molecular weight markers are not shown. Except for liver, other tissues used for the RPA (kidney and colon) are not well represented in the CAGE collection.
K) TIMP-4[41]. Timp-4 RPA-determined TSSs differ from the main CAGE peaks, which are located between the two RPA sites. CAGE did not detect the most upstream RPA-annotated TSS, which seems more evident in heart (not deeply sampled with CAGE). The width of the bands in the RPA assay are does not enable single nucleotide resolution.
L) CSF1R[1]. An additional first exon of CSF-R is not shown, as it is expressed only in trophoblasts which were not sampled by CAGE. The shortest TSS of the [1] may not be real due to a high background signal.
M) Pore forming gene[42]. CAGE and S1 nuclease protection methodology on CTLL-R8 cell line (not sampled with CAGE) disagree on the position of the TSS by about 5bp. However, the resolution of the RPA reported is not that high, and seems to suggest a slightly shorter form. CAGE coverage is low in this region because this gene is expressed exclusively in CTL and NK cells, not samples in our project.
N) Laminin B2[43]. The start sites reported by S1 nuclease protection agrees well with the CAGE data. Consistent with the broader distribution of CAGE tags, the report indications of many additional weaker sites.
O) DHFR[44]. The report indicates that three different promoters exist for the DHFR gene. CAGE and full-length cDNA data adds another explanatory layer on the historical data. Promoter II, which is the most 5' TSS, is actually a TSS for the MSH1 gene, which is on the opposite strand and forms a sense-antisense pair with DHFR. CAGE and S1 nuclease protection agrees fully on the most used start site (promoter III). The TSSs defined by CAGE and S1 nuclease protection are not in agreement: these TSSs are inside the fist exon of the MSH1 gene and are likely rarely used since no EST or cDNA evidence supports them.
P) DNA cytosine-5 methyltrasferase[45]. The RACE-defined peak from erythroleukemia

cells (not sampled by CAGE) coincides with a CAGE CTSS with 4 tags. Expression from other tissues may account for the remaining TSS. Resolution in the original Figure 3 in ref.[45] is not high.
Q) Alpha7 integrin[35]. The two closely located TSS defined by S1 nuclease protection agrees with the CAGE data; both methods agree on exact position of the dominant peak.
R) Alpha 1 (IV) collagen[46]. The primer extension-defined TSS is only 1bp off the major TSS defined by CAGE, which is likely to be the correct site since the initiation site is a PyPu. The report also indicates several minor start sites around this TSS, fully consistent with CAGE data.
S) Wnt-1[47]. The cell lines in this study (P19 teratocarcinoma and cell line 3S) are tumor cell lines that do not correspond to any of the libraries used for CAGE. None of the methods agree on the start sites in this case: it is interesting to note that the report indicates several minor start sites between the reported TSS. It is possible that all TSSs reported are real, but are used in different contexts.

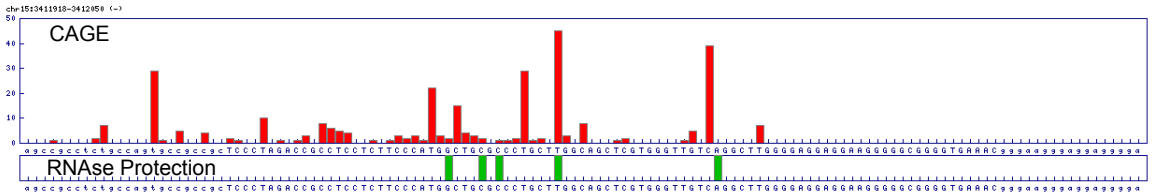
The CAGE data is inherently digital, while the signals obtained by RNase protection assay (the most commonly used experimental method) are analog. It is remarkable how easy it is to find CAGE evidence for minor bands in autoradiographs that their original authors probably disregarded as methodological noise. Data from the tissues that were not sampled by CAGE in this paper shows suggests that their CAGE sampling would be worthwhile to obtain more accurate representation of TSS usage in those expression contexts. Of special interest might be future large-scale CAGE sampling of tumor libraries, which were shown by multiple independent evidence (EST and other) to have TSS position preferences that differ significantly from normal tissues.

Figure S7A-U: CAGE validation examples

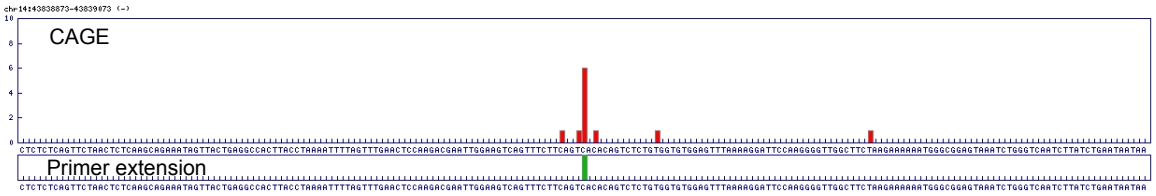
A)
GeneID□Pcmt
□



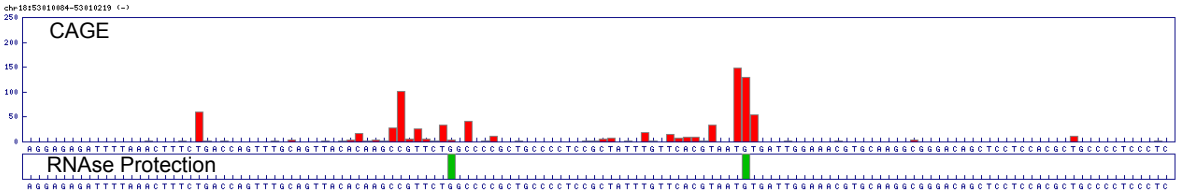
B)
GeneID□Ghr□



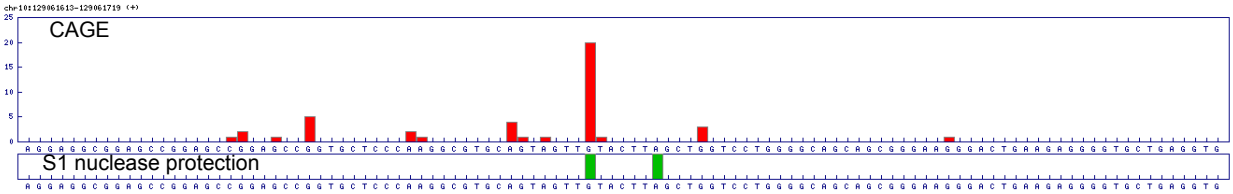
C)
GeneID□Rnase1



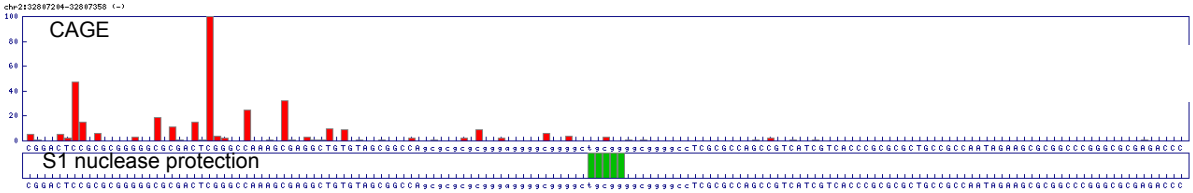
D)
GeneID□Lox□



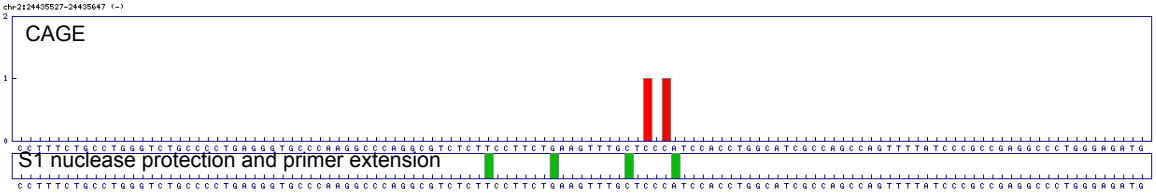
E)
GeneID□ITG7A



F)
GeneID□Stxbp1



G)
GeneID□Pax8□
□



H)
GeneID□Tbxas1

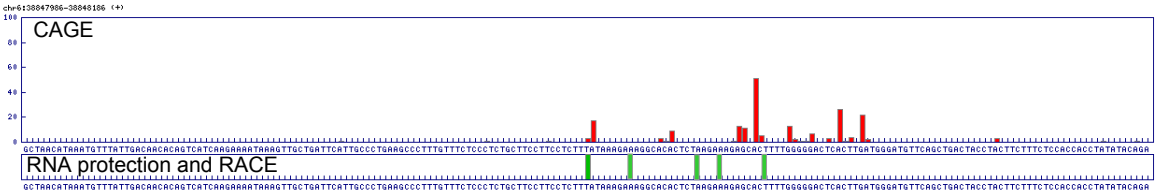
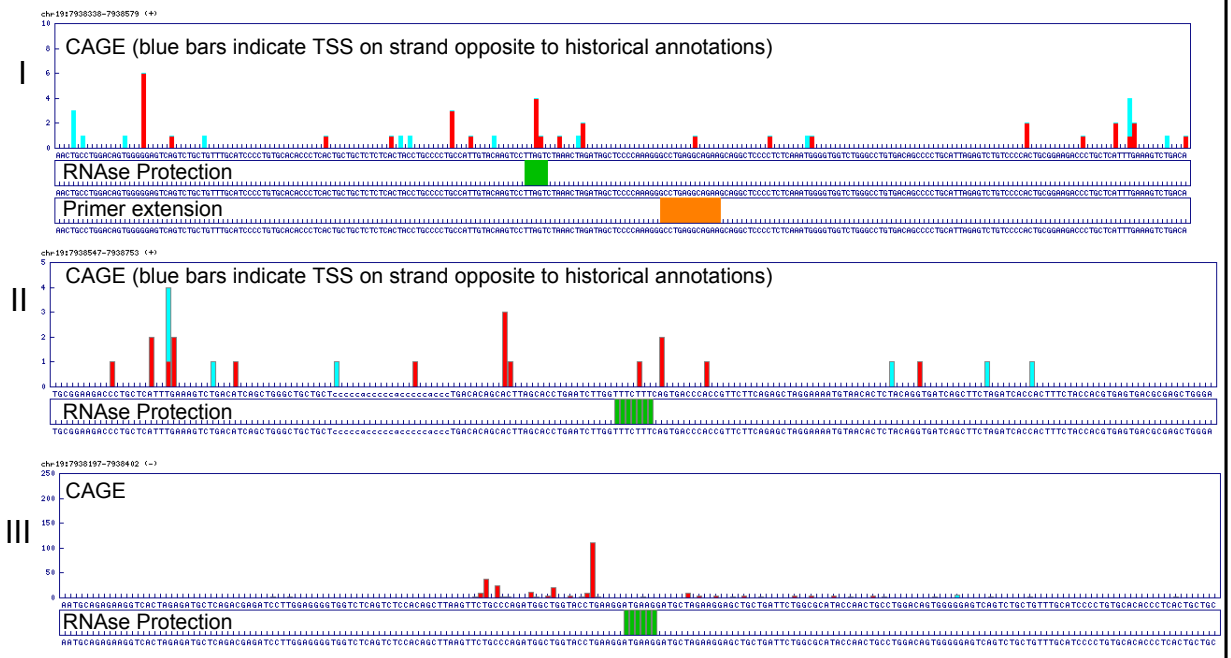


Figure S7A-U: CAGE validation examples

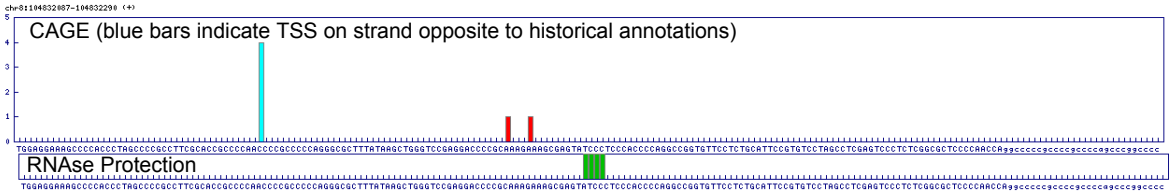
I) GeneID□Bsc12 and Gng3 (bidirectional promoter, promoters I and II direct Bsc12)

□



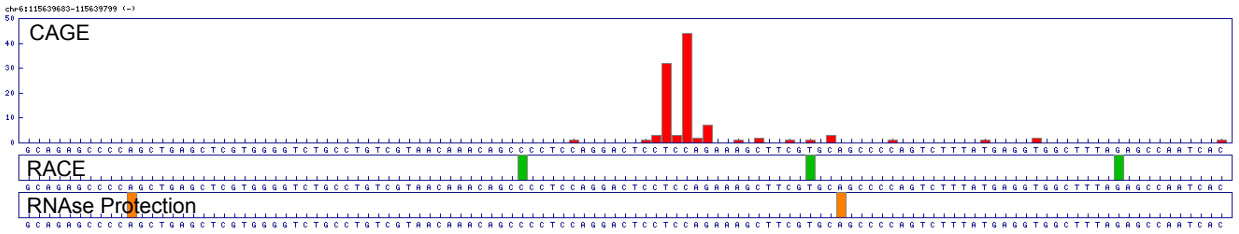
J) GeneID□Hsd11b2

□

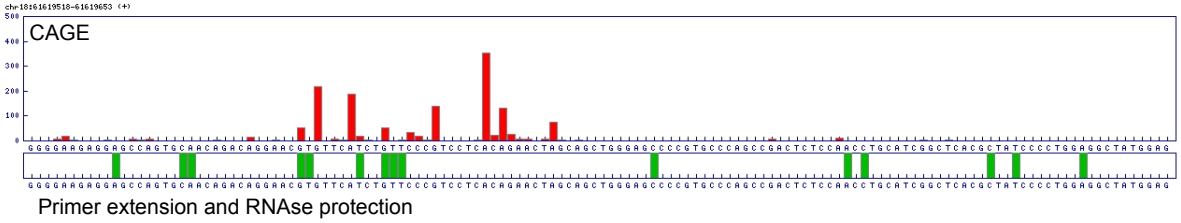


K) GeneID□Timp-4

□

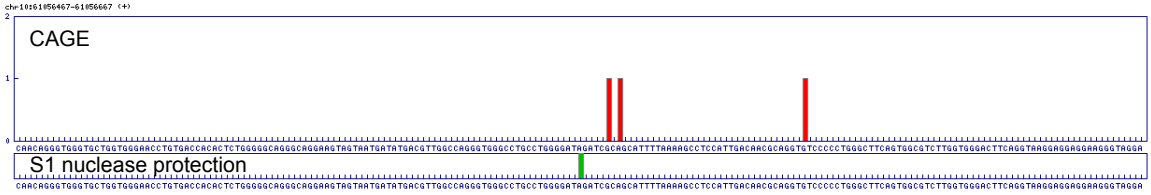


L) GeneID□Csf1r□



M) GeneID Prf1□

□



N) GeneID LamC1□

□

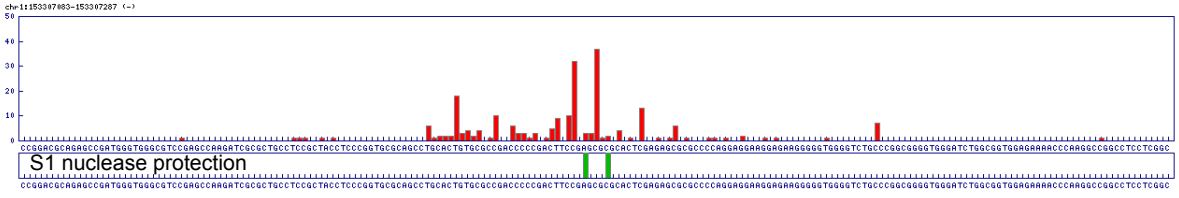
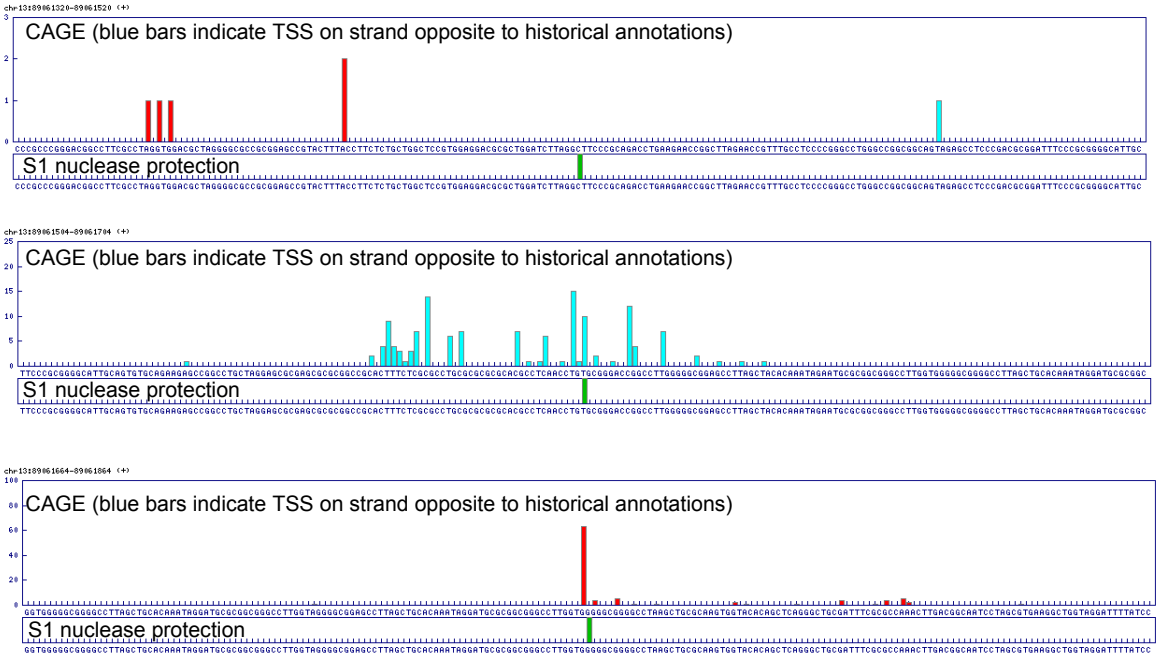


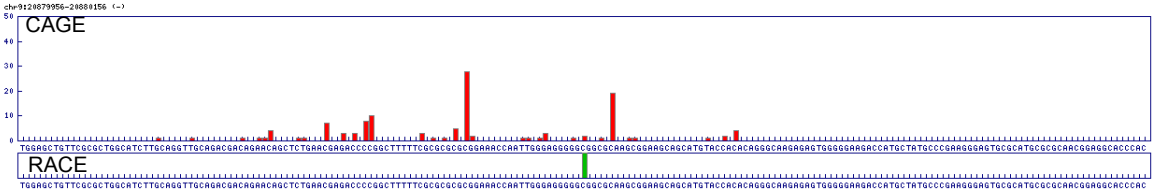
Figure S7A-U: CAGE validation examples

O)
GeneID□Dhfr□

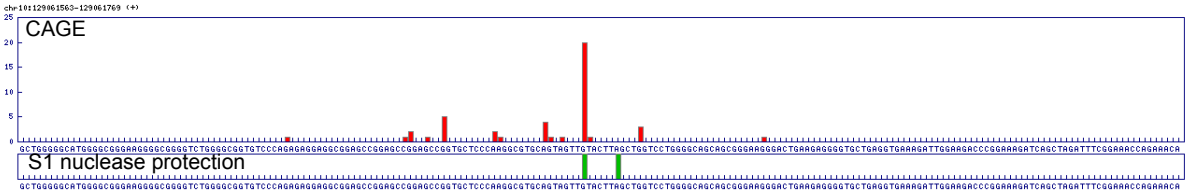
(three promoters)



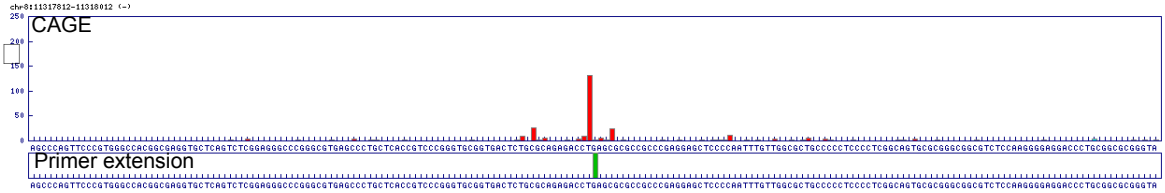
P)
GeneID□Dnmt1
□



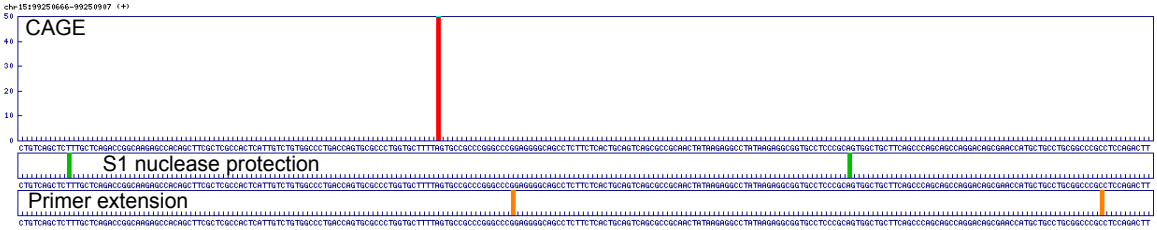
Q)
GeneID Itga7



R)
GeneID□Col14a1□



S)
GeneID□Wnt1
PMID□



T

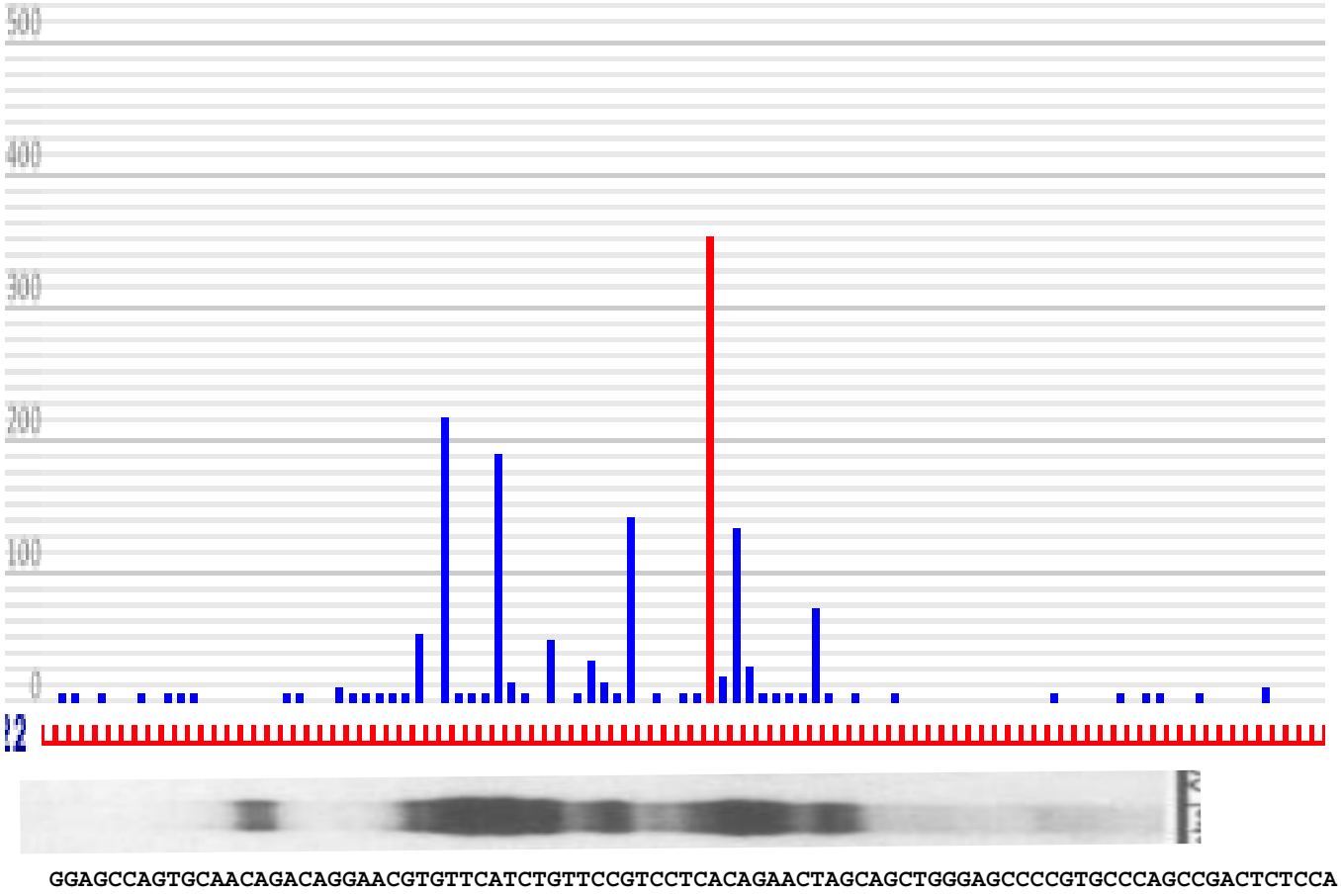


Figure S7T CAGE validation examples
A Comparison of CAGE and RNase protection assay at the Csf1R promoter
We have aligned the CAGE TSS deriving from the promoter (left to right, 5' end to 3' end of the promoter and mRNA sequence). Upper panel: CAGE derived CTSS map (red indicated the major starting site). Bottom: RNase protection assay autoradiograph image aligned to the corresponding genome position (migration: left to right). Notice that the resolution of RNA protection assay does not necessarily allow fine resolution at the single nucleotide level of the TSS. The RNase protection assay image has been sized so that the sequences over the two major protected clusters correspond to the precise base pairs of the histogram above. Since the electrophoretic mobility is not linear, the minor protected RNAs between these major protected bands do not align precisely to the sequence shown.

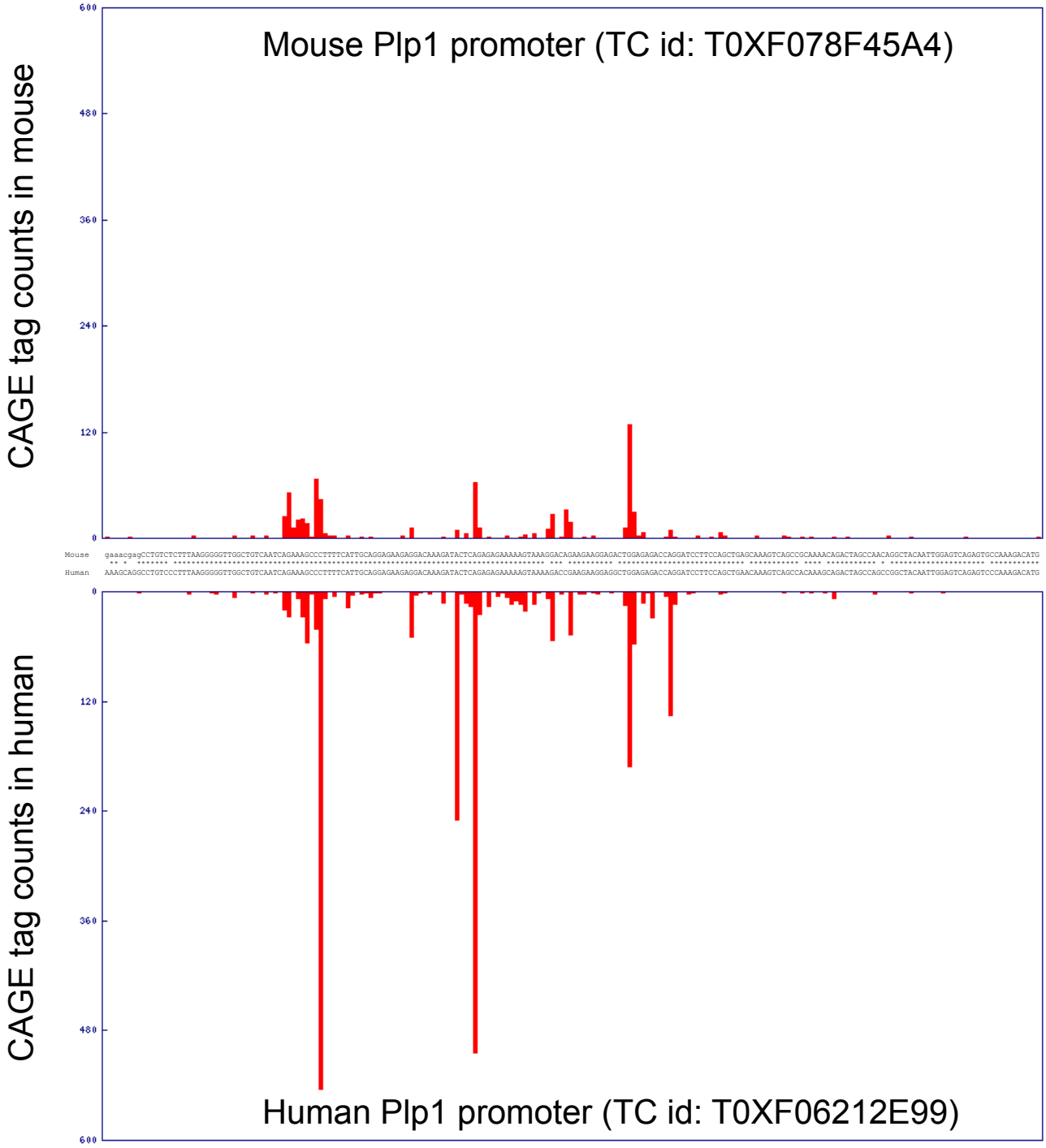
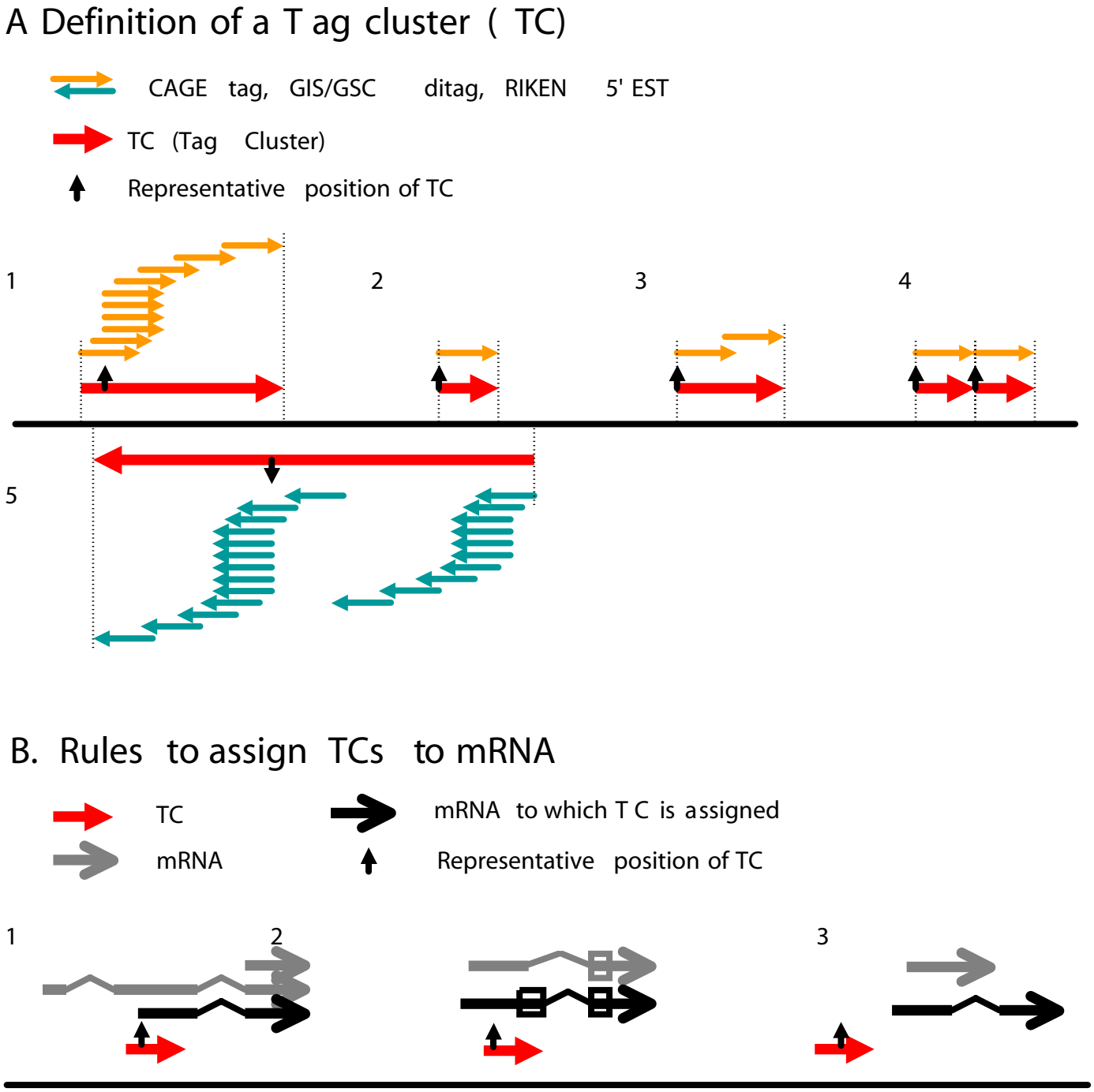


Figure S7U Example of cross-species replication of TSS: the Plp1 promoter
Alignment of the major Plp1 promoter in mouse (top panel) and human (lower panel), with corresponding CAGE tag distributions in red. The actual BLASTZ alignment is shown in the central panel. The CAGE tag peaks in human and mouse are clearly corresponding to each other in both location and relative strength, even at locations with single tags. Changes between human and mouse are correlated with observed nucleotide substitutions. Due to the difference in tissue sampling, the human promoter has been polled to a 3-fold greater extent than in the mouse.

Fig. S8 Definition of TCs and mRNA assignments of TCs.
Detailed exemplification of the rules used to assign the tags to TCs with examples. The rules for the assignments are described in the Experimental Procedures on line.



Tag origin	TC-set Types: All – All		All – Conservative		CAGE – All		CAGE – Conservative		Clustering		More than 100 CAGE tags	
Total TC numbers	736,403	100.0%	236,498	100.0%	594,136	100.0%	177,349	100.0%	159,075	100.0%	8,242	100.0%
CAGE & GIS & GSC & RIKEN 5'-EST & FANTOM3		0.6%		1.7%		0.7%		2.3%		2.5%		35.9%
CAGE & GIS & GSC & RIKEN 5'-EST		0.1%		0.3%		0.1%		0.3%		0.4%		3.1%
CAGE & GIS & GSC & FANTOM3		0.0%		0.0%		0.0%		0.0%		0.0%		0.1%
CAGE & GIS & GSC		0.1%		0.2%		0.1%		0.3%		0.2%		1.2%
CAGE & GIS & RIKEN 5'-EST & FANTOM3		0.2%		0.8%		0.3%		1.0%		1.1%		11.8%
CAGE & GIS & RIKEN 5'-EST		0.1%		0.2%		0.1%		0.2%		0.2%		0.5%
CAGE & GIS & FANTOM3		0.0%		0.0%		0.0%		0.0%		0.0%		0.0%
CAGE & GIS		0.2%		0.6%		0.2%		0.8%		0.6%		0.6%
CAGE & GSC & RIKEN 5'-EST & FANTOM3		0.8%		2.4%		1.0%		3.2%		3.4%		21.8%
CAGE & GSC & RIKEN 5'-EST		0.5%		1.6%		0.6%		2.1%		2.0%		5.0%
CAGE & GSC & FANTOM3		0.0%		0.1%		0.0%		0.1%		0.1%		0.1%
CAGE & GSC		2.2%		6.7%		2.7%		9.0%		5.1%		1.7%
CAGE & RIKEN 5'-EST & FANTOM3		1.2%		3.7%		1.5%		4.9%		3.8%		9.1%
CAGE & RIKEN 5'-EST		1.5%		4.7%		1.9%		6.3%		4.2%		2.2%
CAGE & FANTOM3		0.1%		0.4%		0.1%		0.5%		0.3%		0.1%
CAGE		73.2%		51.8%		90.8%		69.0%		76.0%		6.6%
GIS & GSC & RIKEN 5'-EST & FANTOM3		0.0%		0.0%								
GIS & GSC & RIKEN 5'-EST		0.0%		0.0%								
GIS & GSC & FANTOM3		0.0%		0.0%								
GIS & GSC		0.0%		0.1%								
GIS & RIKEN 5'-EST & FANTOM3		0.0%		0.0%								
GIS & RIKEN 5'-EST		0.0%		0.0%								
GIS & FANTOM3		0.0%		0.0%								
GIS		0.4%		0.3%								
GSC & RIKEN 5'-EST & FANTOM3		0.1%		0.4%								
GSC & RIKEN 5'-EST		0.2%		0.5%								
GSC & FANTOM3		0.0%		0.1%								
GSC		8.6%		19.2%								
RIKEN 5'-EST & FANTOM3		4.4%		1.4%								
RIKEN 5'-EST		4.7%		2.8%								
FANTOM3		0.9%		0.1%								

Table S1. Detailed description of the data sets.
A detailed breakdown of the evidence collected for the identification of mouse TSS. "All" indicates the total count of different types of evidence (type of data, left) including at least one tag, while "Conservative" indicates that at least two items of evidence (tags) were found for each position. Clustering indicates the dataset used for general promoter clustering (as in Fig. 4), consisting of CAGE tags from libraries having at least 1500 tags and having at least 2 tags per TC.

Type	Region	Mm vs Rn	Mm vs Hs	Mm vs Cf
All	Core	0.1126(0.0010)	0.3348(0.0021)	0.3775(0.0027)
	Upstream	0.1181(0.0007)	0.3525(0.0015)	0.4028(0.0019)
	Downstream	0.1077(0.0009)	0.3175(0.0020)	0.3621(0.0026)
MU	Core	0.1010(0.0040)	0.3243(0.0085)	0.3755(0.0121)
	Upstream	0.1169(0.0028)	0.3767(0.0061)	0.4370(0.0077)
	Downstream	0.0923(0.0037)	0.3031(0.0084)	0.3498(0.0112)
BR	Core	0.1123(0.0029)	0.3675(0.0069)	0.4073(0.0089)
	Upstream	0.1201(0.0020)	0.3902(0.0046)	0.4481(0.0059)
	Downstream	0.0925(0.0029)	0.3085(0.0062)	0.3539(0.0081)
PB	Core	0.1063(0.0038)	0.3422(0.0081)	0.3864(0.0105)
	Upstream	0.1191(0.0026)	0.3851(0.0057)	0.4405(0.0072)
	Downstream	0.0947(0.0033)	0.3221(0.0078)	0.3675(0.0104)
SP	Core	0.1176(0.0044)	0.3419(0.0087)	0.3801(0.0112)
	Upstream	0.1274(0.0033)	0.3777(0.0062)	0.4307(0.0080)
	Downstream	0.1125(0.0039)	0.3238(0.0076)	0.3706(0.0102)
CpG	Core	0.0998(0.0013)	0.3369(0.0031)	0.3817(0.0041)
	Upstream	0.1107(0.0009)	0.3697(0.0021)	0.4294(0.0027)
	Downstream	0.0901(0.0012)	0.3105(0.0029)	0.3558(0.0041)
TATA	Core	0.1150(0.0030)	0.3169(0.0065)	0.3508(0.0079)
	Upstream	0.1227(0.0027)	0.3321(0.0049)	0.3729(0.0060)
	Downstream	0.1160(0.0029)	0.3115(0.0060)	0.3492(0.0075)
High	Core	0.1103(0.0019)	0.3473(0.0040)	0.3899(0.0053)
	Upstream	0.1212(0.0013)	0.3836(0.0028)	0.4400(0.0035)
	Downstream	0.0983(0.0017)	0.3145(0.0037)	0.3602(0.0049)
Low	Core	0.1133(0.0011)	0.3307(0.0025)	0.3736(0.0031)
	Upstream	0.1171(0.0009)	0.3414(0.0018)	0.3902(0.0023)
	Downstream	0.1107(0.0011)	0.3185(0.0024)	0.3627(0.0030)
ATG	Core	0.1280(0.0013)	0.3808(0.0027)	0.4389(0.0036)
	Upstream	0.1312(0.0011)	0.3888(0.0021)	0.4486(0.0031)
	Downstream	0.0999(0.0012)	0.2779(0.0023)	0.3068(0.0028)
Random		0.1609(0.0010)	0.4134(0.0022)	0.4633(0.0022)

Table S2A: Substitution rate estimates for mouse promoters.
Substitution rate estimates (95% confidence intervals in parentheses) based upon mouse-based alignments for core promoter, upstream and downstream regions flanking mouse TSSs and start codons (ATG) and randomly sampled mouse genome sequences (Random). Estimates are given for all promoters considered together (All) and various categories of promoters: the promoter classes discovered in the CAGE data (MU, BR, PB, SP), those possessing CpG islands (CpG) or TATA boxes (TATA) and those promoters with TSSs supported by >100 tags (High) and those supported by <100 tags (Low). Comparisons are between the mouse sequence and those from rat (Rn), human (Hs) and dog (Cf). Green background indicates that the promoters' evolution has been significantly slower than that of randomly sampled sequence (yellow).

Type	Region	Hs vs Pt	Hs vs Cf	Hs vs Mm	Hs vs Rn
All	Core	0.0148(0.0004)	0.2901(0.0024)	0.3263(0.0023)	0.3458(0.0026)
	Upstream	0.0136(0.0003)	0.3085(0.0020)	0.3465(0.0020)	0.3667(0.0022)
	Downstream	0.0146(0.0004)	0.2748(0.0024)	0.3115(0.0023)	0.3314(0.0026)
MU	Core	0.0146(0.0020)	0.2707(0.0109)	0.3013(0.0104)	0.3245(0.0118)
	Upstream	0.0132(0.0016)	0.3216(0.0090)	0.3503(0.0078)	0.3753(0.0089)
	Downstream	0.0150(0.0023)	0.2471(0.0112)	0.2844(0.0101)	0.3050(0.0117)
BR	Core	0.0139(0.0010)	0.2956(0.0099)	0.3366(0.0087)	0.3513(0.0094)
	Upstream	0.0126(0.0008)	0.3320(0.0081)	0.3709(0.0071)	0.3875(0.0076)
	Downstream	0.0134(0.0013)	0.2538(0.0103)	0.2993(0.0089)	0.3152(0.0100)
PB	Core	0.0150(0.0019)	0.2841(0.0116)	0.3262(0.0110)	0.3425(0.0117)
	Upstream	0.0125(0.0013)	0.3311(0.0093)	0.3643(0.0086)	0.3849(0.0093)
	Downstream	0.0143(0.0020)	0.2497(0.0118)	0.2875(0.0101)	0.3048(0.0116)
SP	Core	0.0183(0.0020)	0.2764(0.0097)	0.3175(0.0098)	0.3350(0.0109)
	Upstream	0.0164(0.0016)	0.3105(0.0081)	0.3453(0.0080)	0.3663(0.0087)
	Downstream	0.0209(0.0023)	0.2702(0.0097)	0.3093(0.0095)	0.3298(0.0108)
CpG	Core	0.0144(0.0004)	0.3017(0.0032)	0.3328(0.0029)	0.3520(0.0032)
	Upstream	0.0129(0.0003)	0.3259(0.0025)	0.3564(0.0023)	0.3761(0.0025)
	Downstream	0.0138(0.0004)	0.2755(0.0033)	0.3095(0.0029)	0.3270(0.0032)
TATA	Core	0.0151(0.0016)	0.2600(0.0077)	0.3071(0.0080)	0.3266(0.0089)
	Upstream	0.0146(0.0014)	0.2821(0.0062)	0.3328(0.0066)	0.3546(0.0072)
	Downstream	0.0162(0.0016)	0.2659(0.0077)	0.3114(0.0078)	0.3323(0.0087)
High	Core	0.0156(0.0009)	0.2822(0.0052)	0.3211(0.0050)	0.3388(0.0055)
	Upstream	0.0138(0.0007)	0.3235(0.0043)	0.3578(0.0039)	0.3783(0.0043)
	Downstream	0.0162(0.0010)	0.2564(0.0053)	0.2964(0.0048)	0.3149(0.0055)
Low	Core	0.0146(0.0004)	0.2919(0.0027)	0.3276(0.0026)	0.3476(0.0030)
	Upstream	0.0135(0.0003)	0.3048(0.0022)	0.3437(0.0023)	0.3638(0.0026)
	Downstream	0.0142(0.0004)	0.2790(0.0027)	0.3153(0.0026)	0.3355(0.0029)
ATG	Core	0.0124(0.0002)	0.2901(0.0019)	0.3467(0.0018)	0.3676(0.0021)
	Upstream	0.0125(0.0002)	0.2936(0.0016)	0.3472(0.0016)	0.3676(0.0018)
	Downstream	0.0110(0.0002)	0.2186(0.0017)	0.2665(0.0017)	0.2791(0.0019)
Random		0.0132(0.0002)	0.3070(0.0017)	0.3953(0.0022)	0.4188(0.0024)

Table S2B Substitution rate estimates for human promoters.
Substitution rate estimates (95% confidence intervals in parentheses) based upon human-based alignments for upstream and downstream regions flanking human TSSs and start codons (ATG) and randomly sampled human genome sequences (Random). Estimates are given for all promoters considered together (All) and various categories of promoters: the promoter classes discovered in the CAGE data (MU, BR, PB, SP), those possessing CpG islands (CpG) or TATA boxes (TATA) and those promoters with TSSs supported by >100 tags (High) and those supported by <100 tags (Low). Comparisons are between the human sequences and those from chimp (Pt), dog (Cf), mouse (Mm) and rat (Rn). Green background indicates that the promoters' evolution has been significantly slower than that of randomly sampled sequence (yellow), while red indicates significantly more rapid evolution than randomly sampled sequence.

Tissue	SP	BR	PB	MU
adipose	1.98 P=0.14	0.27 P=0.11	1.58 P=0.29	0.44 P=0.47
cns	1.02 P=0.86	0.69 P=0.0020	1.22 P=0.10	1.23 P=0.10
embryo	4.11 P=1.21e-22	0.00 P=6.22e-08	0.30 P=0.0099	0.00 P=8.096e-05
liver	2.15 P=3.56e-21	0.41 P=1.14e-14	0.71 P=0.0053	1.07 P=0.56
lung	2.41 P=1.37e-10	0.23 P=1.42e-08	1.11 P=0.61	0.58 P=0.049
macrophage	1.39 P=0.024	0.64 P=0.0041	0.89 P=0.59	1.26 P=0.14
testis	4.36 P=7.70e-06	0.00 P=0.058	0.00 P=0.21	0.00 P=0.21

Overrepresented	1e-10	1e-06	0.0001	0.01	1.00
Underrrepresented	1e-10	1e-06	0.0001	0.01	1.00

Table S3A. Over-representation of shape classes within tissue specific libraries. Observed/expected ratios and the associated P-values are shown for each combination of tissue and promoter shape.

Table S3B Gene Ontology (GO) terms preferentially associated with SP (red rows) and NM type (green rows) promoters.

GO annotation was compared between all 1216 GO-annotated TUs with SP promoters and all 2033 GO-annotated TUs with NM promoters. TUs without GO annotation and a small number of TUs that had both NM and SP promoters were not used in the comparison. The GO terms had been assigned in the FANTOM3 annotation pipeline. For each of the 735 GO terms that were associated with at least 10 of the 1216 + 2033 TUs, we carried out a two-sided Fisher's exact test to assess preferential association with the SP as compared to the NM set or vice versa. The resulting p-values were corrected for multiple testing with the conservative Bonferroni method. Results significant at $p < 0.01$ are shown.

For example, the majority of genes encoding structural muscle-specific (myofibril) proteins in the observed set (18/19, corrected P-value P=1.7x10-4), polysaccharide binding proteins (20/20, P=1.9x10-6), and chemokines (12/12, P=5.4x10-3) have strong association with SP-type promoters. Only a handful of general categories exist where broad peaks are significantly overrepresented, including RNA processing (corrected P=5.0x10-3) and the ubiquitin cycle (P=5.2x10-3), both of which can be considered constitutive cellular processes. A note of caution is required for certain general categories. For example, this analysis suggests that developmental genes use SP-type promoters. The result is due to highly expressed, event-driven developmental effectors (e.g. Notch1, Socs1, Socs3 or Rtn4), secreted proteins (Spp1, Ifgbp7), or enzymes involved in specific differentiation processes (e.g. xanthine dehydrogenase), which show strong preference for SP-type promoters. This contrasting information can be explained by the fact that in our set of highly expressed genes there are few key developmental regulatory transcription factors overlapping CpG islands both in promoters and along much of their length17, because their overall expression level is too low and embryonic libraries were not deeply sampled.

Ontology	Level	Term id	Parent term	Term	Number of SP TUs	Number of BR TUs	Fraction of SP TUs	Fraction of BR TUs	Fisher test p-value	Bonferroni-corrected p-value
Cellular component	1	GO:0005576		extracellular region	287	248	23.6%	12.2%	8.16e-17	6.00e-14
	2	GO:0005615	extracellular region	extracellular space	262	228	21.5%	11.2%	5.56e-15	4.09e-12
		GO:0043227	organelle	membrane-bound organelle	413	875	34.0%	43.0%	3.09e-07	0.000227
	3	GO:0005886	membrane	plasma membrane	135	104	11.1%	5.1%	6.31e-10	4.64e-07
		GO:0043231	membrane-bound organelle	intracellular membrane-bound organelle	413	875	34.0%	43.0%	3.09e-07	0.000227
	4	GO:0030484	cytoplasm	muscle fiber	19	1	1.6%	0.0%	9.22e-08	6.77e-05
		GO:0042598	membrane fraction	vesicular fraction	27	9	2.2%	0.4%	5.48e-06	0.00403
	5	GO:0030016	muscle fiber	**myofibril**	18	1	1.5%	0.0%	2.36e-07	0.000174
		GO:0005792	vesicular fraction	microsome	27	9	2.2%	0.4%	5.48e-06	0.00403
	6	GO:0030017	myofibril	sarcomere	17	1	1.4%	0.0%	6.05e-07	0.000444
Molecular function	1	GO:0005198		structural molecule activity	93	64	7.6%	3.1%	1.52e-08	1.12e-05
	2	GO:0030246	binding	carbohydrate binding	34	11	2.8%	0.5%	1.98e-07	0.000146
		GO:0001871	binding	pattern binding	23	0	1.9%	0.0%	1.34e-10	9.83e-08
		GO:0004857	enzyme regulator activity	enzyme inhibitor activity	35	12	2.9%	0.6%	3.16e-07	0.000233
		GO:0005102	signal transducer activity; binding	receptor binding	43	20	3.5%	1.0%	9.90e-07	0.000728
		GO:0005200	structural molecule activity	structural constituent of cytoskeleton	24	3	2.0%	0.1%	3.89e-08	2.86e-05
	3	GO:0004497	oxidoreductase activity	monooxygenase activity	20	4	1.6%	0.2%	4.93e-06	0.00362
		GO:0030247	pattern binding; carbohydrate binding	**polysaccharide binding**	20	0	1.6%	0.0%	2.64e-09	1.94e-06
		GO:0005125	receptor binding	cytokine activity	24	7	2.0%	0.3%	6.54e-06	0.00481
		GO:0001664	receptor binding	G-protein-coupled receptor binding	12	0	1.0%	0.0%	7.30e-06	0.00537
	4	GO:0008009	cytokine activity	**chemokine activity**	12	0	1.0%	0.0%	7.30e-06	0.00537
		GO:0042379	G-protein-coupled receptor binding	chemokine receptor binding	12	0	1.0%	0.0%	7.30e-06	0.00537
		GO:0005539	polysaccharide binding	glycosaminoglycan binding	19	0	1.6%	0.0%	7.11e-09	5.23e-06
	5	GO:0008201	glycosaminoglycan binding	heparin binding	16	0	1.3%	0.0%	1.39e-07	0.000102
Biological process	1	GO:0007275		development	189	151	15.5%	7.4%	7.14e-13	5.25e-10
	2	GO:0009653	development	morphogenesis	144	101	11.8%	5.0%	2.04e-12	1.50e-09
		GO:0048513	development	organ development	120	77	9.9%	3.8%	6.13e-12	4.50e-09
		GO:0050793	development	regulation of development	34	14	2.8%	0.7%	2.91e-06	0.00214
		GO:0050874	physiological process	organismal physiological process	113	80	9.3%	3.9%	1.26e-09	9.23e-07
	3	GO:0007155	cell communication	cell adhesion	56	37	4.6%	1.8%	9.65e-06	0.00709
		GO:0009887	morphogenesis; organ development	organogenesis	120	76	9.9%	3.7%	4.85e-12	3.57e-09
		GO:0006955	organismal physiological process	immune response	69	37	5.7%	1.8%	5.00e-09	3.67e-06
		GO:0009607	response to stimulus	response to biotic stimulus	94	53	7.7%	2.6%	3.73e-11	2.74e-08
	4	GO:0007517	organogenesis	muscle development	29	9	2.4%	0.4%	1.44e-06	0.00106
		GO:0006952	response to biotic stimulus	defense response	81	41	6.7%	2.0%	4.90e-11	3.60e-08
		GO:0043207	response to biotic stimulus	response to external biotic stimulus	45	23	3.7%	1.1%	1.50e-06	0.00110
		GO:0009613	response to stress	response to pest, pathogen or parasite	41	21	3.4%	1.0%	4.20e-06	0.00308
	5	GO:0001525	blood vessel morphogenesis	angiogenesis	20	6	1.6%	0.3%	4.98e-05	0.0366
		GO:0006954	response to pest, pathogen or parasite	inflammatory response	25	9	2.1%	0.4%	2.10e-05	0.0155
		GO:0006935	taxis	chemotaxis	22	8	1.8%	0.4%	7.89e-05	0.0580
		GO:0006464	protein metabolism	protein modification	84	246	6.9%	12.1%	1.43e-06	0.00105
		GO:0006396	RNA metabolism	**RNA processing**	14	75	1.2%	3.7%	6.78e-06	0.00498
	6	GO:0006512	protein modification	**ubiquitin cycle**	25	105	2.1%	5.2%	7.02e-06	0.00516

Resource	URL	Contents
CAGE Basic Viewer	http://gerg01.gsc.riken.jp/cage/	All basic information about CAGE tags, mappings and tissue information
Mouse	http://gerg01.gsc.riken.jp/cage/mm5	
Human	http://gerg01.gsc.riken.jp/cage/hg17	
Genome Element Viewer	http://gerg02.gsc.riken.jp/gev-promoter/gbrowse/	Promoter elements, cDNA, EST and other data are mapped to respective genome
Mouse	http://gerg02.gsc.riken.jp/gev-promoter/gbrowse/mm5	
Human	http://gerg02.gsc.riken.jp/gev-promoter/gbrowse/hg17	
Promoter datasets	http://fantom31p.gsc.riken.jp/cage/download/	CAGE tag data and mapping information
Mouse	http://fantom31p.gsc.riken.jp/cage/download/mm5/	
Human	http://fantom31p.gsc.riken.jp/cage/download/hg17/	
Expression tree viewer	http://gerg01.gsc.riken.jp/expr_tree/	Promoter clustering tree with associated annotations
Mouse	http://gerg01.gsc.riken.jp/expr_tree/mm5/	
Human	http://gerg01.gsc.riken.jp/expr_tree/	
CAGE analysis viewer	http://gerg01.gsc.riken.jp/cage_analysis/	User-friendly interface to CAGE data on promoter level
Mouse	http://gerg01.gsc.riken.jp/cage_analysis/mm5	
Human	http://gerg01.gsc.riken.jp/cage_analysis/hg17/	
3D viewer		Visualization of genome location, development stages and tissue specificity
Mouse	http://gerg01.gsc.riken.jp/expr_3D/mm5/	
Alternative promoter sets	http://gerg01.gsc.riken.jp/altp/	

Table S4 Internet links to publicly available resources and datasets.

Ctss size	Observed number	Expected number	observed/ expected	p-value
2	144129	9707.84	14.85	<1E-324
3	60762	6.59	9221.78	<1E-324
4	35026	2.98E-003	1.17E+7	<1E-324
5	23479	1.01E-006	2.32E+10	<1E-324
10	6720	3.08E-025	2.18E+28	<1E-324
25	1259	7.66E-087	7.13E+88	<1E-324
50	348	7.48E-198	9.28E+198	<1E-324
100	90	<1E-324	NA	<1E-324

Table S5A Observed and expected number of multitag CTSS in mouse
The probability of the observed number of multitag CTSS(broken up by the number of tags in the CTSS) by random selection is shown (all p-values are below the underflow limit of standard computers)

number of tags in TC	number of TCs with tags from > 1 library	number of TCs with tags from a single libray	% reproducibility
2	52662	14998	77.83
3	23752	1824	92.87
4	12559	611	95.36
5	7870	296	96.38
6	5401	161	98.11
7	3964	116	97.16
8	2992	44	97.55
9	2421	29	98.82
10	1934	27	98.62
15	939	7	98.26
20	535	0	100
50	134	0	100
100	57	0	100

Table S5B Tag cluster reproducibility
Inter-library reproducibility of multitag tag clusters (TCs), broken up by number of tags in the TC. A TC is considered reproduced if it is composed of tags from more than one library. Since different CAGE libraries are distinct experiments, TCs that contain tags from more than one library are experimentally reproducible

	Ctss from classified tag clusters			All ctss		
number of tags in ctss	number of ctss with tags from >1 library	number of ctss with tags from a single library	% reproducibility	number of ctss with tags from >1 library	number of ctss with tags from a single library	% reproducibility
2	41528	5955	87.46	112262	31867	77.89
3	26761	843	96.94	56878	3884	93.61
4	18151	212	98.88	33803	1223	96.51
5	13396	130	99.03	22853	626	97.33
6	10420	74	99.29	16664	327	98.07
7	8137	52	99.36	12530	205	98.39
8	6620	16	99.75	9877	99	99.01
9	5602	13	99.76	8034	54	99.33
10	4794	8	99.83	6682	38	99.43
15	2538	1	99.86	3171	7	99.78
20	576	1	99.93	1929	2	99.89
50	325	0	100	348	0	100
100	89	0	100	90	0	100

Table S5C CTSS reproducibility

Inter-library reproducibility of multitag CTSS, broken up by number of tags in the CTSS. A CTSS is considered reproduced if it is composed of tags from more than one library.

number of liver tags in human ctss	*number of cases*	*number of human liver ctss validated by mouse liver ctss*	*% human liver ctss that are supported by liver ctss*
1	20492	12078	58.9
2	8268	5325	64.4
3	4530	3102	68.5
4	2894	2078	71.8
5	2076	1495	72.0
10	654	509	77.8
15	328	262	79.9
20	215	176	81.9

Table S5D Cross-species validation of CTSS

Extent of human liver CTSS supported by mouse liver ctss within alignable promoters, broken up by number of liver tags in the CTSS. A liver CTSS in human is considered repreduced if a liver CTSS occurs in mouse with the same strand at the corresponding aligned position, plus/minus 1 bp

Table S6 Over-representation index of TFBS in macrophage promoters

A) List of TFBS over-represented in the constitutive macrophage set (450 TSSs) compared to background. B) List of TFBS over-represented in the induced set (295 TSSs) compared to the background. TFBS names refer to cognate profile model name. A more comprehensive list including counts and likelihood is available in Table S7

A Constitutive macrophages TC versus 40,000 random TC comparison having frequencies ≥ 10% and ORI ≥ 1.6

Elk-1	GC box	IRF1
PEA3	PU.1	SOX-9
c-Ets-1	c-Ets-2	

B Induced macrophages TC versus 40,000 random TC comparison having frequencies ≥ 10% and ORI ≥ 1.6

IRF	IRF-7	PU.1
SOX-9	c-Ets-1	c-Ets-2

Table S7. Over-representation and under-representation index of TFBS in macrophage promoters, detailed view.
A, over-representation/under-representation of TFBS in the constitutively expressed macrophages promoters. ORI indicates the ratio of over-representation of compared to the control set (40K, consisting of 40,000 randomly selected promoters). B, over-representation/under-representation of TFBS in promoters induced in macrophages after 7 hours from addition of LPS, versus the 40,000 randomly selected promoters. An ORI value >1 indicates over-representation in the target, while <1 indicates under-representation.

A Constitutive macrophages TC versus 40,000 random TC comparison having frequencies ≥ 10% and ORI ≥ 1.6

TFBS pattern	ORI	% in target	% in background	probability in TARGET	probability in BACKGROUND
+1 c-Ets-1	3.3796+	15.35	8.82	2.5914E-04	1.3346E-04
+1 c-Ets-2	3.3325+	13.49	7.64	2.1595E-04	1.1446E-04
-1 ABI4	3.3	14.65	8.03	2.4585E-04	1.3577E-04
+1 E2F-1/DP-1	2.79	13.72	8.33	2.2259E-04	1.3153E-04
+1 ABI4	2.57	11.16	6.86	1.7940E-04	1.1340E-04
+1 MAZR	2.36	21.86	13.89	6.3123E-04	4.2068E-04
-1 PCF2	2.31	21.86	14.5	3.9203E-04	2.5604E-04
-1 c-Ets-1	2.2200+	13.26	8.87	1.9934E-04	1.3412E-04
+1 ELF-1	2.2	20.7	13.93	3.3887E-04	2.2918E-04
+1 c-Ets-1 68	2.2	18.14	12.76	3.1229E-04	2.0228E-04
+1 PEA3	2.1499+	23.95	16.26	3.7874E-04	2.5947E-04
+1 MAZ	2.12	46.51	32.94	1.2193E-03	8.1314E-04
-1 CDC5	2.03	11.4	8.12	1.8272E-04	1.2648E-04
+1 E12	1.98	11.4	8.02	1.6944E-04	1.2188E-04
+1 Nrf2	1.96	10.23	7.47	1.5947E-04	1.1132E-04
-1 PEA3	1.9457+	23.02	16.87	3.8538E-04	2.7033E-04
-1 ELF-1	1.86	18.6	14.14	3.2226E-04	2.2805E-04
-1 c-Ets-2	1.8470+	10.93	8.16	1.6944E-04	1.2283E-04
+1 E2F	1.83	64.65	50.18	5.4784E-03	3.8674E-03
+1 GC box	1.8215+	70	57.05	3.8904E-03	2.6208E-03
-1 c-Ets-1 68	1.8	17.21	12.92	2.7575E-04	2.0389E-04
+1 Sp-1	1.79	75.58	62.5	7.2558E-03	4.8926E-03
-1 GCR1	1.77	26.51	19.78	4.3854E-04	3.3231E-04
-1 ETF	1.76	52.33	40.96	2.0498E-03	1.4850E-03
-1 Elk-1	1.7514+	44.88	35.67	1.0698E-03	7.6859E-04
-1 Hairy	1.75	21.63	16.59	4.1860E-04	3.1188E-04
+1 ETF	1.73	58.37	44.52	2.4053E-03	1.8203E-03
+1 EGR	1.73	58.6	46.37	2.2259E-03	1.6254E-03
-1 PU.1	1.7221+	47.91	38.56	1.0299E-03	7.4294E-04
-1 Sp-1	1.72	75.81	61.73	6.4352E-03	4.5977E-03
-1 E2F	1.7	60.23	49.72	4.8904E-03	3.4836E-03
+1 Sp3	1.7	36.05	28.08	7.1096E-04	5.3839E-04
+1 GCR1	1.69	24.19	19.36	4.3522E-04	3.2160E-04
-1 E2F-1/DP-1	1.69	11.63	8.53	1.6944E-04	1.3686E-04
+1 Elk-1	1.6705+	43.02	34.78	1.0066E-03	7.4535E-04
+1 CAC-binding prc	1.67	41.63	33.05	9.8007E-04	7.3968E-04
+1 PU.1	1.6533+	48.14	38.75	1.0133E-03	7.6132E-04
+1 Adf-1	1.65	18.6	13.03	6.5449E-04	5.6668E-04
+1 c-Rel	1.64	36.51	29.66	7.1429E-04	5.3550E-04
-1 EGR	1.62	51.86	41.37	1.7741E-03	1.3726E-03
+1 STRE	1.62	22.09	17.14	3.6545E-04	2.9163E-04
-1 GC box	1.6119+	63.26	52.73	2.9967E-03	2.2301E-03
+1 Dfd	0.62	50.93	59.37	1.3023E-03	1.7877E-03
+1 IRF1	0.6248+	15.81	19.92	2.6578E-04	3.3779E-04
-1 dri	0.62	30.23	35.3	5.4485E-04	7.5379E-04
+1 FOX	0.62	27.91	36.59	1.1229E-03	1.3834E-03
-1 COMP1	0.62	16.28	20.24	2.5249E-04	3.2884E-04
-1 TATA	0.61	28.37	34.94	7.5083E-04	9.9744E-04
-1 HP1 site factor	0.61	12.79	15.84	1.9269E-04	2.5619E-04
-1 Nkx6-2	0.61	26.98	33.26	5.7475E-04	7.7002E-04
-1 Cdc5	0.6	36.51	43.96	6.9435E-04	9.5549E-04
-1 Cdx-2	0.6	26.28	32.72	4.9169E-04	6.5706E-04
-1 CCAAT box	0.6	16.05	20.89	2.7243E-04	3.4952E-04
+1 Evi-1	0.6	21.63	27.39	4.3189E-04	5.7037E-04
-1 TBP	0.6	36.98	43.93	9.3355E-04	1.3154E-03
+1 Nkx6-2	0.6	28.14	33.48	5.5482E-04	7.8259E-04
-1 Zen	0.59	41.63	52.77	9.8339E-04	1.3224E-03
+1 dri	0.58	27.67	34.67	5.3488E-04	7.3029E-04
-1 Nkx2-5	0.58	18.84	23.41	2.8571E-04	3.9393E-04
-1 Pbx-1	0.58	47.44	57.26	1.3289E-03	1.8880E-03

-1 FOXO4	0.57	14.19	18.09	2.2259E-04	3.0508E-04
-1 Ovo	0.57	10.7	13.96	1.6279E-04	2.1876E-04
+1 Cdx-2	0.55	25.58	33.4	4.8837E-04	6.8435E-04
-1 SBF-1	0.54	14.65	19.14	2.4252E-04	3.4196E-04
+1 BR-C Z1	0.5	16.28	23.15	3.1894E-04	4.5298E-04
+1 NIT2	0.49	13.26	18.54	2.1927E-04	3.2062E-04
-1 SOX-9	0.4238+	13.49	19.64	2.0598E-04	3.3377E-04
-1 Sox-5	0.4	10.47	15.4	1.4950E-04	2.5286E-04
-1 MRF-2	0.33	10.7	17.71	1.6279E-04	3.0150E-04

B Induced macrophages TC versus 40,000 random TC comparison having frequencies ≥ 10% and ORI ≥ 1.6 a

TFBS pattern	ORI	% in target	% in background	probability in TARGET	probability in BACKGROUND
-1 IRF	8.0819+	16.38	6.18	3.0363E-04	9.9587E-05
+1 IRF	4.4706+	12.2	6.54	2.5386E-04	1.0595E-04
+1 c-Ets-2	3.5717+	13.94	7.64	2.2399E-04	1.1446E-04
-1 STATx	3.36	10.8	5.78	1.5431E-04	8.5846E-05
-1 c-Ets-1	2.6228+	13.94	8.87	2.2399E-04	1.3412E-04
+1 c-Ets-1	2.5288+	13.59	8.82	2.1901E-04	1.3346E-04
+1 ELF-1	2.4	22.3	13.93	3.4345E-04	2.2918E-04
+1 CHOP-C/EBPal	2.36	10.45	7.16	1.7422E-04	1.0792E-04
+1 IRF-7	2.2961+	33.45	22.3	6.0727E-04	3.9667E-04
-1 Poly A downstrea	2.25	10.1	6.81	1.5431E-04	1.0185E-04
+1 c-Ets-1 68	2.17	18.12	12.76	3.0861E-04	2.0228E-04
-1 Nrf2	2.14	11.15	7.68	1.6924E-04	1.1490E-04
-1 IRF-7	2.0651+	30.66	22.93	6.3216E-04	4.0931E-04
-1 CAT8	2.03	11.15	7.95	1.7422E-04	1.2024E-04
+1 c-Rel	2.03	39.72	29.66	8.1135E-04	5.3550E-04
-1 ELF-1	2.02	19.86	14.14	3.2852E-04	2.2805E-04
+1 GCN4	1.97	10.45	7.42	1.6924E-04	1.2111E-04
-1 c-Ets-2	1.9407+	11.5	8.16	1.6924E-04	1.2283E-04
-1 GCR1	1.82	25.78	19.78	4.6292E-04	3.3231E-04
+1 NF-kappaB	1.81	50.17	40.46	1.2842E-03	8.7819E-04
-1 dl	1.8	27.87	20.58	4.4798E-04	3.3724E-04
+1 c-Ets-1(p54)	1.79	67.6	52.79	1.8915E-03	1.3507E-03
-1 core-binding fact	1.77	11.5	8.54	1.7422E-04	1.3237E-04
-1 PCF2	1.76	19.86	14.5	3.2852E-04	2.5604E-04
-1 AML1	1.7	24.39	18.66	3.9821E-04	3.0614E-04
+1 Osf2	1.7	28.92	22.53	4.9776E-04	3.7664E-04
+1 NF-Y	1.69	16.72	13.59	5.6247E-04	4.1048E-04
-1 NF-kappaB	1.68	47.04	41.85	1.3639E-03	9.1456E-04
+1 NF-AT	1.67	60.63	48.58	2.0458E-03	1.5261E-03
-1 c-Rel	1.67	35.54	28.2	6.6700E-04	5.0345E-04
+1 Ik-1	1.66	22.3	17.61	3.6834E-04	2.8082E-04
+1 PU.1	1.6392+	49.83	38.75	9.7063E-04	7.6132E-04
-1 ABI4	1.62	10.45	8.03	1.6924E-04	1.3577E-04
+1 HNF-3alpha	0.61	32.75	39.16	7.4664E-04	1.0156E-03
-1 BR-C Z1	0.61	21.6	28.06	4.6789E-04	5.9387E-04
-1 TBP	0.6	37.28	43.93	9.3081E-04	1.3154E-03
+1 Ubx	0.6	35.54	44.11	7.3171E-04	9.8429E-04
+1 FOXJ2	0.6	16.03	21.36	3.6336E-04	4.5580E-04
-1 LEF1	0.6	16.72	21.85	2.8372E-04	3.6403E-04
-1 AGL3	0.59	10.45	13.67	2.2399E-04	2.9244E-04
-1 TATA	0.58	28.92	34.94	6.9686E-04	9.9744E-04
-1 dri	0.57	26.13	35.3	5.8238E-04	7.5379E-04
+1 MAZR	0.57	11.5	13.89	2.8870E-04	4.2068E-04
-1 HP1 site factor	0.55	11.85	15.84	1.8915E-04	2.5619E-04
-1 SOX-9	0.5472+	15.33	19.64	2.3395E-04	3.3377E-04
+1 Cdx-2	0.54	25.09	33.4	4.9278E-04	6.8435E-04
-1 Sox-5	0.44	10.8	15.4	1.5928E-04	2.5286E-04
-1 MRF-2	0.42	11.85	17.71	1.8915E-04	3.0150E-04