nature genetics

Genome-wide analysis of mammalian promoter architecture and evolution

Piero Carninci[1,2,21], Albin Sandelin[1,3,21], Boris Lenhard[1,3,20,21], Shintaro Katayama[1], Kazuro Shimokawa[1], Jasmina Ponjavic[1,20], 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[3,20], 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[3,20], Edison T Liu[14], Matthias Harbers[15], Jun Kawai[1,2], Vladimir B Bajic[1,7,16], David A Hume[1,6,21] & Yoshihide Hayashizaki[1,2,17,18]

Mammalian promoters can be separated into two classes, conserved TATA box–enriched promoters, which initiate at a well-defined site, and more plastic, broad and evolvable CpG-rich promoters. We have sequenced tags corresponding to several hundred thousand transcription start sites (TSSs) in the mouse and human genomes, allowing precise analysis of the sequence architecture and evolution of distinct promoter classes. Different tissues and families of genes differentially use distinct types of promoters. Our tagging methods allow quantitative analysis of promoter usage in different tissues and show that differentially regulated alternative TSSs are a common feature in protein-coding genes and commonly generate alternative N termini. Among the TSSs, we identified new start sites associated with the majority of exons and with 3′ UTRs. These data permit genome-scale identification of tissue-specific promoters and analysis of the *cis*-acting elements associated with them.

With the completion of several mammalian genome sequences, the next challenge for mammalian genomics is to understand how transcription is controlled. Present algorithms aimed at TSS prediction have proven unsatisfactory[1]. Although many TSSs from mouse can be inferred from the 5′ ends of full-length cDNAs and 5′ ESTs[2,3], the depth of coverage is limited.

To increase the depth of coverage, we have carried out systematic 5′-end analysis of the mouse and human transcriptome using the cap analysis of gene expression (CAGE) approach[4]. Here we redefine basic promoter features and analyze the diversity, evolutionary conservation and dynamic regulation of mammalian promoters on a genome-wide scale.

[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]UK Medical Research Council (MRC) Human Genetics Unit, Western General Hospital, Crewe Road, Edinburgh, EH4 2XU, UK. [5]University of Oxford, Roosevelt Drive, Oxford, OX3 7BN, UK. [6]Australian Research Council (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, California 92093, USA. [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-Scuola Internazionale Superiore Studi Avanzati (I.S.A.S.-S.I.S.S.A.), AREA Science Park, Padriciano 99, 34012 Trieste, Italy. [12]Sector of Neurobiology, I.S.A.S.-S.I.S.S.A., AREA Science Park, Padriciano 99, 34012 Trieste, Italy, [13]Dulbecco Telethon Institute, Institute of Genetics and Biophysics, Consiglio Nazionale delle Ricerche (IGB CNR), Epigenetics and Genome Reprogramming Laboratory, Pietro Castellino Street 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.[19]The Eskitis Institute for Cell and Molecular Therapies, Griffith University, Nathan Campus, Kessels Road, Queensland 4111, Australia. [20]Present addresses: Bergen Center for Computational Science, Unifob AS, University of Bergen, Thormøhlensgate 55, N-5008 Bergen, Norway (B.L.), Scripps Florida, Jupiter, Florida 33458, USA (C.W.), Department of Molecular Medicine, National Public Health Instititute, Department of Medical Genetics, University of Helsinki, Biomedicum, FIN-00251 Helsinki, Finland (S.M.-T.) and MRC Functional Genetics Unit, Department of Physiology, Anatomy and Genetics, University of Oxford, Oxford OX1 3QX, UK (J.P.). [21]These authors contributed equally to this work. Correspondence should be addressed to D.H. (d.hume@imb.uq.edu.au) or Y.H. (yosihide@gsc.riken.jp).

Received 9 December 2005; accepted 27 March 2006; published online 28 April 2006; corrected online 5 May 2006; corrected after print 29 August 2007; doi:10.1038/ng1789

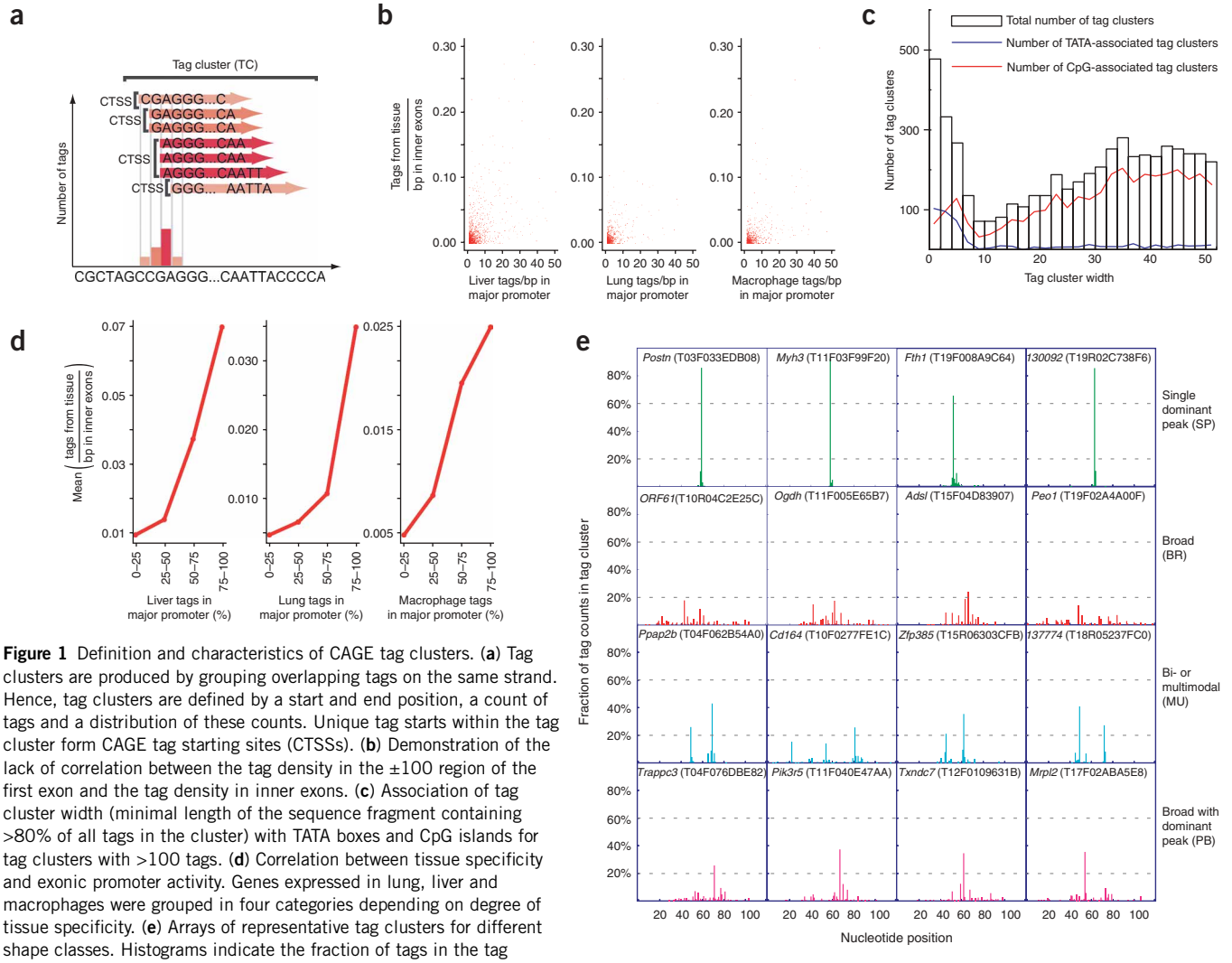

Figure 1 Definition and characteristics of CAGE tag clusters. (**a**) Tag clusters are produced by grouping overlapping tags on the same strand. Hence, tag clusters are defined by a start and end position, a count of tags and a distribution of these counts. Unique tag starts within the tag cluster form CAGE tag starting sites (CTSSs). (**b**) Demonstration of the lack of correlation between the tag density in the ±100 region of the first exon and the tag density in inner exons. (**c**) Association of tag cluster width (minimal length of the sequence fragment containing >80% of all tags in the cluster) with TATA boxes and CpG islands for tag clusters with >100 tags. (**d**) Correlation between tissue specificity and exonic promoter activity. Genes expressed in lung, liver and macrophages were grouped in four categories depending on degree of tissue specificity. (**e**) Arrays of representative tag clusters for different shape classes. Histograms indicate the fraction of tags in the tag cluster mapping into each position in a 120-bp window centered on the tag cluster. The single peak (SP) class is characterized by a sharp peak, indicative of a single, well-defined TSS. The broad (BR) shape indicate multiple, weakly defined TSSs. The bimodal/multimodal (MU) shape class implies multiple well-defined TSSs within one cluster. Combination of a well-defined TSS surrounded by weaker TSSs results in a broad with dominant peak shape (PB). HUGO gene names or transcriptional unit identifiers for cognate genes and tag cluster identifiers are shown above each tag cluster.

RESULTS

Defining TSSs by CAGE tags

CAGE tags are 20- or 21-nt sequence tags that are derived from the mRNA sequenced in the proximity of the cap site, and their mapping onto unique genomic regions identifies TSSs[4,5]. CAGE libraries are constructed from full-length cDNAs selected through a biotinylated cap. Second-strand synthesis is absolutely dependent upon the ligation, to the first-strand full-length cDNAs, of a primer that contains restriction sites allowing the cleavage of 5′ 20- to 21-bp tags from the resulting cDNA. These short fragments are concatemerized and sequenced. We applied CAGE sequencing to 145 different mouse and 41 different human libraries. We mapped tags to the mouse and human genomes using a hierarchical data structure (**Fig. 1a**). CAGE tags that had an identical 5′ start site were grouped into a CAGE-tag starting site (CTSS), whereas CTSSs that overlap on the same strand form a tag cluster.

Mapping cDNAs, ESTs and CAGE, GIS and GSC tags to mouse and human genomes allowed us to identify 729,504 potential mouse and 665,278 human TSSs (**Table 1,** and **Supplementary Fig. 1** and **Supplementary Table 1** online). Of these, 593,290 in the mouse genome and 629,716 in the human genome were defined by CAGE tag clusters.

The majority of tag clusters identified by two or more tags (159,075 mouse and 177,563 human) were derived from independent libraries (**Supplementary Note** online). Therefore, we selected these tag clusters for detailed expression and promoter analysis. We aimed to identify the major promoters of the widest possible diversity of genes by sampling at relatively low depth many tissues and conditions. Most single CAGE tags (singletons) reflect the fact that in most libraries the number of tags sequenced (~100,000) is lower than the total number of transcripts per cell[6]. Therefore, rare transcripts were sampled randomly.

We provide several lines of evidence demonstrating that CAGE identifies genuine transcription start sites, including (i) statistical analysis of reproducibility within and across species, (ii) experimental validation by distinct primer extension approaches, (iii)

Table 1 Data sets used for analysis

Tag cluster grouping

Mouse sets	Number of tags (CAGE, GIS, GSC, RIKEN 5′-EST, FANTOM3 clones)	Number of CAGE tags	Number of tag clusters	Number of tag clusters with ≥2 tags
Full set	8,892,784	7,151,511	736,403	236,498
CAGE set	8,413,283	7,151,511	594,136	177,349
CAGE set (tag clusters with ≥ 2 tags)	7,996,496	6,734,724	177,349	177,349
CAGE set (tag clusters with ≥ 100 CAGE tags)	6,403,169	5,632,183	8,242	8,242
Clustering set	7,906,938	6,714,273	159,075	159,075

Human sets	Number of tags (CAGE, Long-SAGE, dbTSS)	Number of CAGE tags	Number of tag clusters	Number of tag clusters with ≥2 tags
Full set	5,510,369	5,312,921	665,278	190,513
CAGE set	5,460,627	5,312,921	629,716	184,379
CAGE set (tag clusters with ≥2 tags)	5,015,290	4,855,717	184,379	184,379
CAGE set (tag clusters with ≥100 CAGE tags)	3,858,982	3,781,211	5,561	5,561
Clustering set	4,997,086	4,855,717	177,563	177,563

Association of tag clusters with transcriptional units

Mouse sets	Number of associated coding transcriptional units	Number of associated noncoding transcriptional units
CAGE set	25,420	14,173
CAGE set (tag clusters with ≥2 tags)	21,182	7,819
CAGE set (tag clusters with ≥100 CAGE tags)	7,172	217
Clustering set	20,732	7,370
Human sets		
CAGE set	24,248	9,655
CAGE set (tag clusters with ≥2 tags)	21,506	6,730
CAGE set (tag clusters with ≥100 CAGE tags)	4,650	159
Clustering set	21,368	6,653

historical comparison to TSSs analyzed by other methodologies, (iv) sequence bias around CAGE tags, (v) correlation with published sites of TATA-binding protein–associated binding protein 1 identified by chromatin immunoprecipitation (ChIP), (vi) conservation of start site architecture between orthologous mouse and human genes and (vii) enrichment over noncapped RNAs (**Supplementary Note**).

CAGE tags identify transcription from unconventional sites

Not all CAGE tags mapped to previously identified 5′ ends of full-length cDNAs (**Supplementary Fig. 1**). Typically, we observed peaks over the known 5′ end of the transcript and a second (generally smaller) peak within 3′ UTRs[7]. The CAGE tags that mapped to genomic regions between these two peaks mapped mostly to exons. For relatively highly expressed genes, the 'internal' CAGE tag frequency was supported by a considerable number of 5′ EST sequences[2] and has been confirmed using RACE based on the alternative oligonucleotide-capping method[8]. If we consider the overall set of 159,075 TSSs identified by CAGE, 34,229 TSSs mapping within exons would generate transcripts that truncate or eliminate the predicted protein product.

Exonic promoter activity varies between genes and is conserved across species. For example, the gene encoding albumin (*Alb1* in mice and *ALB* in humans) has high level exonic initiation in both mouse and human, whereas *Col3a1* and *COL3A1* have negligible levels in mouse and human, respectively (**Supplementary Fig. 2** online). Exonic promoter activity does not correlate with the number of tags over the major promoter(s) (**Fig. 1b**), but it is highest in tissue-specific genes (**Fig. 1d**) and correlates with a single dominant site of transcription initiation (**Supplementary Fig. 2** and **Supplementary Note**).

Promoter coverage

We found that 13,767 of the 20,639 mouse protein-coding transcriptional units (67%) were supported by one or more tag clusters at ±20 nt from the reported 5′ end of a full-length cDNA (**Table 1**). If tags aligned to the 5′ UTR and the rest of the transcript are included, 73% and 81%, respectively, of all protein-coding loci are supported. Many of the remaining protein-coding loci actually have a candidate TSS supported by individual tags. The extended distribution of TSS within CpG islands means that for weakly expressed transcripts there were multiple tags in the putative promoter region, but they did not overlap to form a larger tag cluster. Given the reliability of the CAGE technology (**Supplementary Note**), single tags can be regarded as candidate TSSs. The annotation of CAGE tags is based on the annotation of the closest gene on the same strand, which can be confirmed in specific cases by additional experiments.

Table 2 Overrepresentation and underrepresentation of transcriptional starting site sequence

Overall analysis	SP	BR	PB	MU
TATA (all)	**3.1×10^{-73}**	***1.9×10^{-16}***	***1.8×10^{-10}***	***2.4×10^{-9}***
CCAAT (all)	0.04	0.42	0.37	0.49
GC (all)	***1×10^{-4}***	0.20	0.40	0.33
CpG (all)	***1.0×10^{-137}***	**1.4×10^{-65}**	**8.7×10^{-6}**	0.02
CpG promoters versus non-CpG promoters	**SP**	**BR**	**PB**	**MU**
TATA (no CpG)	**2.6×10^{-77}**	***1.6×10^{-16}***	***2.8×10^{-16}***	***1.0×10^{-9}***
CCAAT (no CpG)	**6.8×10^{-23}**	***9.2×10^{-16}***	0.11	0.42
GC (no CpG)	**7.8×10^{-25}**	***5.9×10^{-18}***	0.48	0.35
CpG (no TATA, CCAAT or GC)	***4.8×10^{-45}***	**4.7×10^{-17}**	**3.4×10^{-5}**	0.87

For each shape class, we determined whether a TATA box (within 50 bp) or a CCAAT, GC or CpG (within 200 bp) upstream of the start site of the clusters was present. *P* values were determined using the Fisher exact test (**Supplementary Note**). *P* values in boldface and italics indicate significant underrepresentation ($P < 0.01$); *P* values in boldface alone indicate significant overrepresentation ($P < 0.01$). In the lower part of the table, we separated pure CpG-island–overlapping promoters (without TATA, CCAAT and GC elements) and TATA, CCAAT and GC promoters (without CpG islands).

CAGE tag distribution shape defines distinct promoter groups

To analyze and classify TSS distributions, we chose 8,185 mouse and 5,928 human tag clusters supported by at least 100 CAGE tags. These clusters had biphasic distribution in terms of the genomic interval covered by the cluster of tag sequences (**Fig. 1c**).

We classified the landscapes defined by tags within a single cluster into four shapes. In the single dominant peak class (SP), the majority of tags are concentrated to no more than four consecutive start positions (**Fig. 1e**, top row) with a single dominant TSS. We divided the clusters spanning a broader region into three categories (**Fig. 1e**): a general broad distribution (BR), a broad distribution with a dominant peak (PB) and a bi- or multimodal distribution (MU). This classification is essential for further characterization of structural and functional difference of SP versus BR distributions; we devised the additional PB and MU classes to sequester ambiguous cases. There was a high degree of conservation of shape classes between orthologous mouse and human promoters, even at a single-nucleotide level (**Supplementary Figs. 3** and **4** online).

Shape classes identify different promoter contexts

We identified putative TATA-box, CCAAT-box and GC-box sites using position-specific weight matrices[9] and extracted positions of CpG islands from the UCSC Genome Browser database[10]. For each shape class, we compared the under- and overrepresentation of the four promoter features to the whole set of promoters (**Table 2**). TATA boxes were strongly overrepresented in promoters showing sharp TSSs, whereas broad TSS regions were strongly associated with CpG islands. In roughly 90% of the cases, TATA-independent transcription initiation occurred within a CpG island. This percentage of TATA-independent transcription initiation is much greater than previous estimates[11].

To distinguish between CpG-island and other types of transcription, we separated pure CpG island–overlapping promoters (without TATA, CCAAT and GC elements) and TATA, CCAAT and GC promoters, which are not in CpG islands (**Table 2**), and checked for the frequency of occurrence of the four TSS shape categories. CCAAT-box and GC-box sequences that were not associated with CpG islands were preferentially associated with SP-type TSSs.

Re-estimation of spacing of core TATA-promoter elements

The main function of TATA boxes is to anchor the transcription preinitiation complex guiding RNA polymerases[12] upstream of TSSs. Of the 1,315 putative TATA boxes situated within 50 bp of a tightly defined TSS, >70% start between positions –33 and –28, with positions –31 and –30 as the preferred sites (**Fig. 2a**). This spacing is consistent with the structural evidence showing that the distance from the TATA box to the active center of RNA polymerase II is 30 bp[13].

The consensus mammalian initiator sequence

Transcription preferentially starts with a purine at position +1, with a preference for pyrimidine at position –1 (58.6% of tags have a pyrimidine-purine dinucleotide at position –1,+1). This pyrimidine-purine dinucleotide corresponds in part to the Inr element[14] or Cap motif[9] (pyrimidine, pyrimidine, A(+1), N, T/A, pyrimidine, pyrimidine, where N is any nucleotide), which was determined by mutagenesis studies of specific promoters[13]. The sequence logos in **Figure 2b–e** show that adenine is not actually the preferred starting base, and only

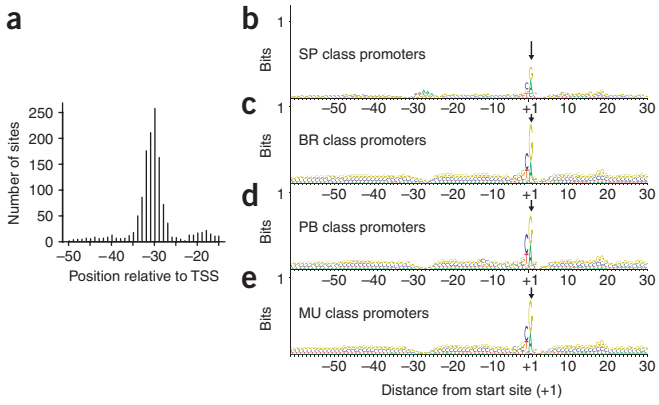

Figure 2 TATA-box and TSS spacing definition and consensus. (**a**) Accurate distribution of the spacing between TATA-box promoter and initiation sites. (**b–e**) Sequence logos[15] for promoter sequences aligned at the TSSs constructed by counting each tag and its flanking region as one sequence, divided by promoter shape class. The *y* axis shows the information content (measured in bits), reviewed in ref. 15. In all cases, there is a clear preference for a pyrimidine-purine initiation site at –1,+1. A TATA-like motif is visible around the –30 position in the SP class promoters (**b**). In the BR class promoters, as most of those promoters are overlapped by CpG islands, the entire region is GC-rich; there is anisotropy of nucleotide content: there are more guanine than cytosine nucleotides in the plus strand upstream of the TSS (**c**). The logos of PB (**d**) and MU (**e**) class promoters look similar to this, indicating that these two ambiguous two categories are more likely to share the common initiation mechanism with BR promoters than with the SP ones. The PB class has a certain proportion of mixed cases, with both a CpG island and a TATA-box.

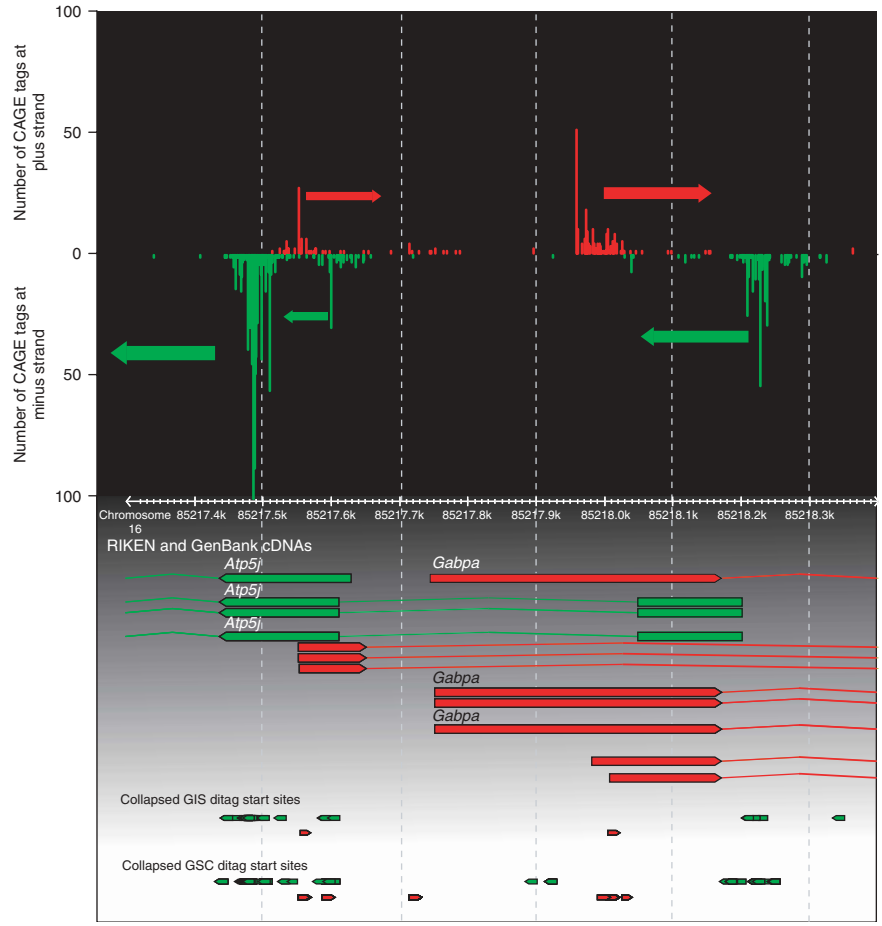

Figure 3 Bidirectional overlapping promoters of *Gabpa* and *Atp5j*. The Gabpa (on the plus strand) and Atp5j (on the minus strand) bidirectional gene pair contains two promoters in each direction, strongly supported by overlapping start sites of full-length cDNA clones and by GIS and GSC data. The distribution of CAGE tags mapping to the region is indicated in the top panel. Red and green denote plus and minus strand direction for all data types. GIS and GSC data has been collapsed into single tracks.

the central pyrimidine-purine shows any bias when assessed over the full diversity of TSSs in the mouse transcriptome. The most preferred initiators are CG, CA and TG. The comparative prevalence of CA, CG and TG increases with tag frequency class (**Supplementary Fig. 4**). Therefore, the initiator sequence is not an absolute determinant of transcription initiation, and the presence of CG, TG or CA dinucleotides is associated with more active TSSs.

Sequence of TSSs for different shape classes of promoters

We constructed sequence logos[15] for promoter sequences aligned at the TSSs (**Fig. 2b–e**) by counting each tag and its flanking region as one sequence . As expected, there is a prominent initiator sequence at position –1,+1 and a TATA-box signature around the –30 position for the single dominant start class. We observed an equivalent sequence logo for broad-type promoters (**Fig. 2c**). Although there are multiple TSSs, there is still a preference for specific initiation sites, corresponding to pyrimidine-purine dinucleotides at position –1,+1. Additionally, there is evidence of GC anisotropy (more guanines on the plus strand) both upstream and downstream of the major TSSs (**Fig. 2c**). CpG-island promoters do show a preferential directionality in promoter assays[12]; the apparent prevalence of guanines on the plus strand could contribute to promoter orientation, although no clear mechanism is evident. The transcription factor Sp1 has been found to recruit TATA-binding protein in the absence of TATA boxes[16]. Consistent with this and the overall high CG content, consensus Sp1 sites were overrepresented in broad promoters, although the position relative to individual TSS was less precise than for TATA boxes (**Supplementary Fig. 5** online).

Although the logos of two 'hybrid' promoter categories (**Fig. 2d,e**) look similar to those of the BR category, both contain a higher proportion of TATA boxes; a substantial number of MU and PB promoter regions represent TATA boxes within CpG islands. The class probably represents independent functional core promoters that lie fewer than 20 bp apart and have been joined into a single tag cluster. In PB promoters, the TSS indicated by the peak has on average a significantly higher tissue specificity than the rest of the combined tags in the cluster ($P < 2.2 \times 10^{-16}$, Wilcoxon test). Moreover, the tissue specificity of the peak was significantly increased when clear TATA boxes were found ($P < 3.823 \times 10^{-7}$, Wilcoxon test; **Supplementary Note**).

Start site preference and bidirectional promoters in CpG islands

The CAGE data is derived by polling TSS use from a wide variety of tissues. The broad distribution of TSSs is a well-documented feature of CpG island promoters assayed in single cell types[13]. In some cases, the depth of coverage is sufficient to determine whether the broad distribution of TSSs in CpG islands derived from CAGE data represents overlapping distributions of TSSs used differentially by distinct tissues. Statistically significant tissue-specific TSS preference occurs for 34% of 5,607 CpG island promoters ($P < 0.01$, Kruskal-Willis test) for which we currently have sufficient tags to make assessments (H. Kawaji *et al.*, data not shown). CpG island promoter regions are also commonly associated with bidirectional promoter activity[17]. In these cases, the key issue is whether the promoters are truly bidirectional. CAGE data strongly support multiple TSSs in the bidirectional promoter region separating the adenosine kinase and *AP3m1* genes. There are two separate tag clusters in each direction within the one CpG island, and there is no absolute overlap between any of the TSS regions (data not shown). A contrary example is the bidirectional *Gabpa-Atp5j* promoter region[18] in mouse. This bidirectional promoter region contains two promoters in each direction (**Fig. 3**), strongly supported by cDNAs and by gene identification signature (GIS) and gene signature cloning (GSC) data. In such cases, there is potential for transcriptional interference as described in yeast and bacteria[19]. Such instances extend the numerous examples of sense-antisense interactions within the transcriptome[20].

Promoter evolution in mammalian genomes

Phylogenetic footprinting[21] has been used to identify conserved transcription factor binding sites (TFBSs) within the vicinity of

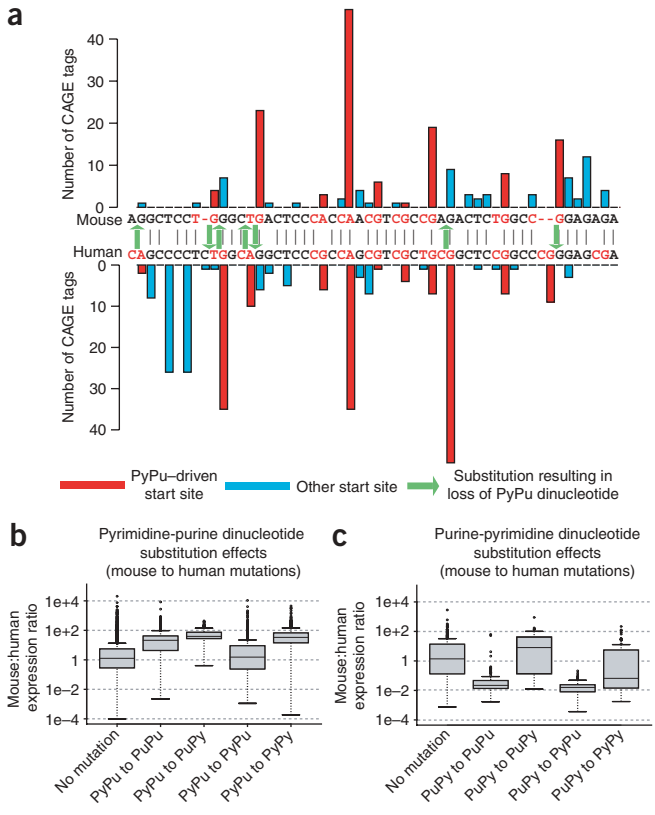

Figure 4 Pyrimidine-purine dinucleotides drive expression. (**a**) A detailed view of the core promoter of the mouse *Ptprn* gene (TC 73140) and corresponding human region illustrates the usage of pyrimidine-purine dinucleotides as dominant start sites and the expression changes resulting from mutations at these positions. (**b**,**c**) Box plots describing substitution effects in pyrimidine-purine (PyPu) and purine-pyrimidine (PuPy) dinucleotides. Expression ratios are calculated as (number of tags in mouse/number of total tags in mouse)/(number of tags in human/number of total tags in human).

mammalian genes[22]. We evaluated evolutionary divergence in mammalian promoters using a set of 30,898 mouse and 26,290 human promoter sequences validated by at least ten tags per TSS. For all categories of TSSs, the percentage identity declined considerably within 200 nt upstream of the TSS (**Supplementary Fig. 3**). There is a pronounced pattern of troughs and peaks in conservation within the 70 nt flanking the TSS where the most conserved peak corresponds to the TATA box.

We analyzed substitution rates within the four distinct promoter shape classes, based on the mouse promoter sequences versus rat, human and dog genomes and based on human promoters versus chimpanzee, dog, mouse and rat, over a 1,000-bp window upstream of the TSSs (**Supplementary Table 2** online). In all comparisons, the TATA-containing promoters had a lower substitution rate than the other three promoter types (the 'upstream' regions in all of these comparisons, except for the human-chimpanzee comparison, which is uninformative owing to the short evolutionary distance). These data support the view that human BR promoters have higher substitution rates (relative to the mutation rate) than mouse promoter regions[23].

Pyrimidine-purine dinucleotides drive promoter expression

As pyrimidine-purine dinucleotides are overrepresented at the position –1,+1 of TSSs (**Fig. 2**), substitutions at this Inr element are expected to represent functional changes. Indeed, transversion-type substitutions at these positions correlate with large differences in TSS usage (**Fig. 4** and **Supplementary Fig. 4**). The effect of the pyrimidine to purine substitution at position –1 ($P = 4.34 \times 10^{-5}$, two-sided *t*-test versus nonmutated pyrimidine-purine TSS) was comparable to the purine-to-pyrimidine mutations at position +1, ($P = 5.82 \times 10^{-12}$), whereas double transversions had the most severe effect ($P = 4.39 \times 10^{-12}$; **Fig. 4b**). The reverse is also true: mutations from other sequences to pyrimidine-purine dinucleotides create new TSSs (**Fig. 4c** and **Supplementary Fig. 4**). This evolutionary data further support the view that the pyrimidine-purine dinucleotide contributes to the precise TSS locations in BR promoters.

Dynamic expression and functional association of promoters

The CAGE TSS usage data constitute a quantitative profiling of relative promoter use across many tissues and cell types. Therefore, we hierarchically clustered[24] the 159,075 tag clusters from mouse according to their normalized expression values in parts per million[25] in libraries with at least 1,500 mapped tags). The 70 clusters (or supergroups) obtained distinguished tissue-specific promoters, and also clustered together promoters of broadly expressed genes according to their function or family (**Supplementary Note**).

A global heat map of the clustering contains the 70 supergroups of CAGE tag clusters (**Fig. 5**). Clustering brought together several supergroups into five larger bodies, corresponding to specific tissues and/or conditions (**Fig. 5c**). Analysis of the sequence characteristics (CAGE tag distribution shapes, CpG and TATA-box associations and densities of potential TFBSs; **Fig. 5a**,**b**,**d**,**e**) of tag clusters in each supergroup demonstrated underlying promoter features influencing global expression. In general, ubiquitous transcripts were associated with a broad TSS and CpG islands, whereas tightly regulated transcripts were associated with sharper distinct TSSs and TATA-box promoters[26] (**Table 2**, **Supplementary Note** and **Supplementary Table 3**). One exception was the central nervous system–specific promoters, which were especially CpG-rich. Using representative models[27], we measured the relative density of potential TFBSs in the immediate 300-bp upstream region of the tag clusters (**Fig. 5d**). There was a direct correlation between the global properties of promoters and TFBS density.

Impact of alternative promoter usage on the proteome

Previous analysis using limited data sets[28] has suggested that 18–20% of protein-coding genes use alternative promoters. In our data set, 58% of protein-coding transcriptional units (11,264 out of 19,142) had two or more alternative promoters, based on the presence of nonoverlapping tag clusters. Among these transcriptional units, there were 63,060 alternative tag clusters that belong to distinct clusters (**Fig. 5**; see website listed in Methods). Of the protein-coding transcriptional units that had at least two putative TSSs, 92.9% are predicted to use distinct methionine start codons. As an example, the UDP-glucuronyl transferase gene has seven promoters used preferentially by different tissues and driving six alternative ATGs (**Fig. 6a**). We found that 10,959 transcriptional units identified by clusters from oligo-dT–primed CAGE libraries (98%) and 2,668 transcriptional units identified by clusters from random primed CAGE libraries (99%) have at least one coding transcript. Among them, 5,331 transcriptional units identified by oligo-dT–primed CAGE libraries (48%) and 925 transcriptional units identified by random-primed CAGE libraries (34%) have at least one alternative promoter that overlapped the coding sequence of known or predicted transcripts.

In 97% of cases, at least two of the alternative promoters from the same transcriptional unit were located in different CAGE cluster supergroups. For example, the gelsolin gene (*Gsn*) (**Supplementary Fig. 6** online) has two alternative promoters, producing the same protein product in macrophages and liver, respectively, and one additional alternative promoter generating a protein with a different function in heart and cerebellum (**Supplementary Note** and **Supplementary Fig. 6**).

Promoters in 3′ UTRs of known protein-coding genes

As noted above, there is a considerable increase in CAGE tag incidence in the 3′ UTRs of protein-coding transcripts. These TSSs have been independently validated by a distinct RACE method (**Supplementary Fig. 6**) and are supported also by GIS and GSC ditag analysis[7,20]. These TSSs also have a distinct sequence motif. Alignments of the most tag-rich TSS derived from 3′ UTRs revealed a strong overrepresentation of three consecutive guanines, found at position –3 to–1 in 785 out of 1,327 cases, just before the TSSs (**Fig. 6c**). Analysis of cross-species conservation between sequenced vertebrate genomes in the region surrounding the 3′ UTR TSSs revealed a highly conserved region located at positions +40 to +90 relative to the TSSs (**Fig. 6b**). To confirm that these sequences can indeed initiate transcription, we performed reporter-gene analysis with four distinct 3′ UTR promoters. In each case, upstream regions of the TSS directed reporter-gene expression (**Fig. 6d**).

Transcripts initiated in 3′ UTRs might regulate downstream genes using a sense-antisense mechanism, as downstream genes on the opposite strand are located much closer than expected[7,20]. If 3′ UTR–derived transcripts function as regulatory noncoding RNAs, their transcriptional regulation might be discordant from the full-length transcript. In 43% (168/391) of testable representative transcripts, the tag distribution in 5′ and 3′ terminal exons was significantly divergent ($P < 0.05$, Bonferroni-corrected Fisher's exact test) in at least one tissue, suggesting independent regulation of the 3′ UTR promoter (**Supplementary Fig. 6**). Notably, the incidence of 3′ TSS is tissue specific: it is prevalent among libraries derived from cerebellum and lung but reduced in libraries derived from embryo.

Promoting the macrophage-specific transcriptome

As a model system to demonstrate the power of the approach, we applied CAGE to bone marrow–derived macrophages (BMM) treated with different stimuli. Based on the clustering tree (**Fig. 5**), we extracted two very tight promoter clusters: those expressed in constitutively macrophages but not in other cell types, and those inducible by macrophage activating agents such as lipopolysaccharides (LPS) and CpG DNA, but also macrophage-restricted. The first set includes many known macrophage markers such as the archetypal CSF-1 receptor (encoded by *CSF1R* in humans and *Csf1r* in mouse), and the second includes large set of known inflammatory cytokines. The global binding motif analysis of TFBS families identified an overrepresentation of the Ets and NF-κB family binding sites (**Fig. 5d**), consistent with the well-documented functions of these two gene families in macrophage transcriptional control[29,30]. A more detailed analysis of TFBS incidence in these promoter clusters can be found in the **Supplementary Note** and **Supplementary Figure 5**.

Promoter database structure links

A description of new, publicly available databases and resources integrating CAGE, ESTs, full-length cDNAs and other genomic elements is available in **Supplementary Table 4**.

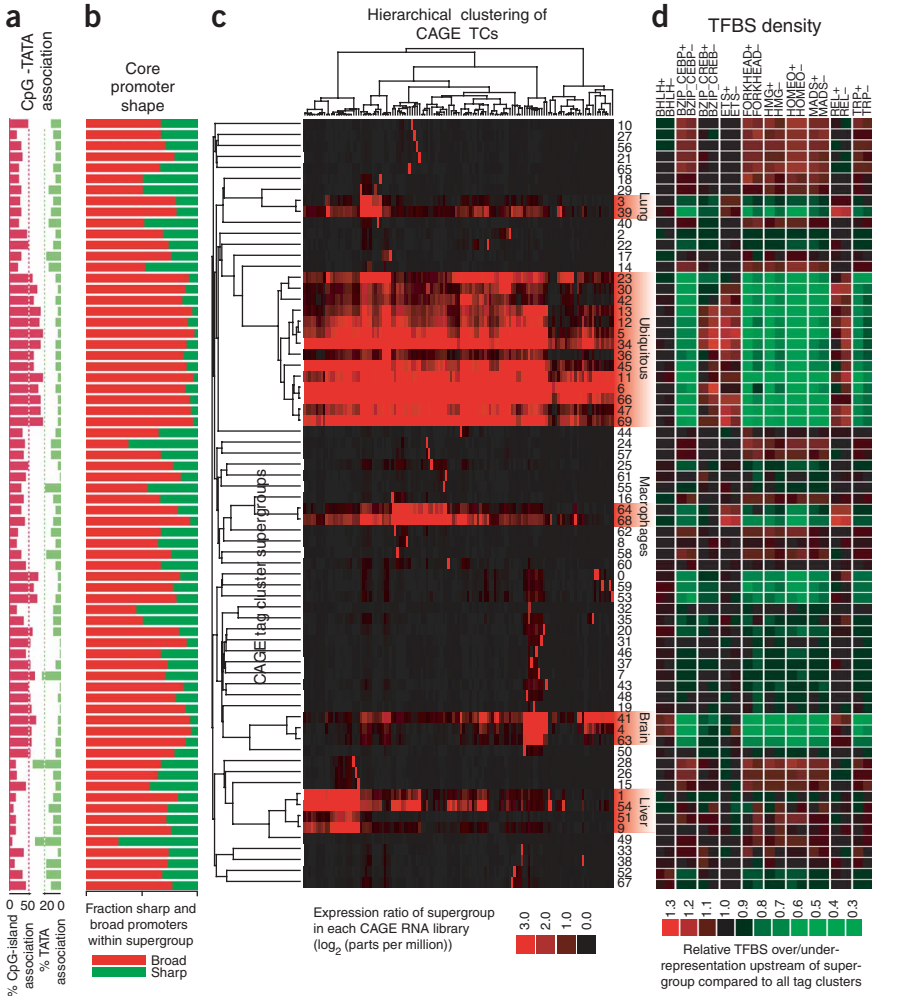

Figure 5 Promoter-based clustering reveals global features of the transcriptome. Analysis is based on hierarchical clustering the tag clusters usage using CAGE libraries containing at least 1,500 tags. (**a**) Percentage of CpG-island and TATA-box association for each supergroup. (**b**) Promoter shape propensity (broad or sharp TSS distribution for each supergroup (see **Supplementary Note** for definitions)). (**c**) The 159,075 core promoters (TCs) clustered into 70 supergroups with respect to RNA library expression ratios. *x* axis, grouping of RNA libraries (hierarchical clustering of CAGE tag clusters; tissues not indicated individually); left *y* axis, the relations among supergroups; right *y* axis, the identifier number of clusters and the tissues in which indicated tag cluster supergroups are overrepresented. (**d**) Relative density of representative TFBSs 300-bp upstream regions of the tag clusters within each supergroup; + and – indicate the strand relative to the TSS.

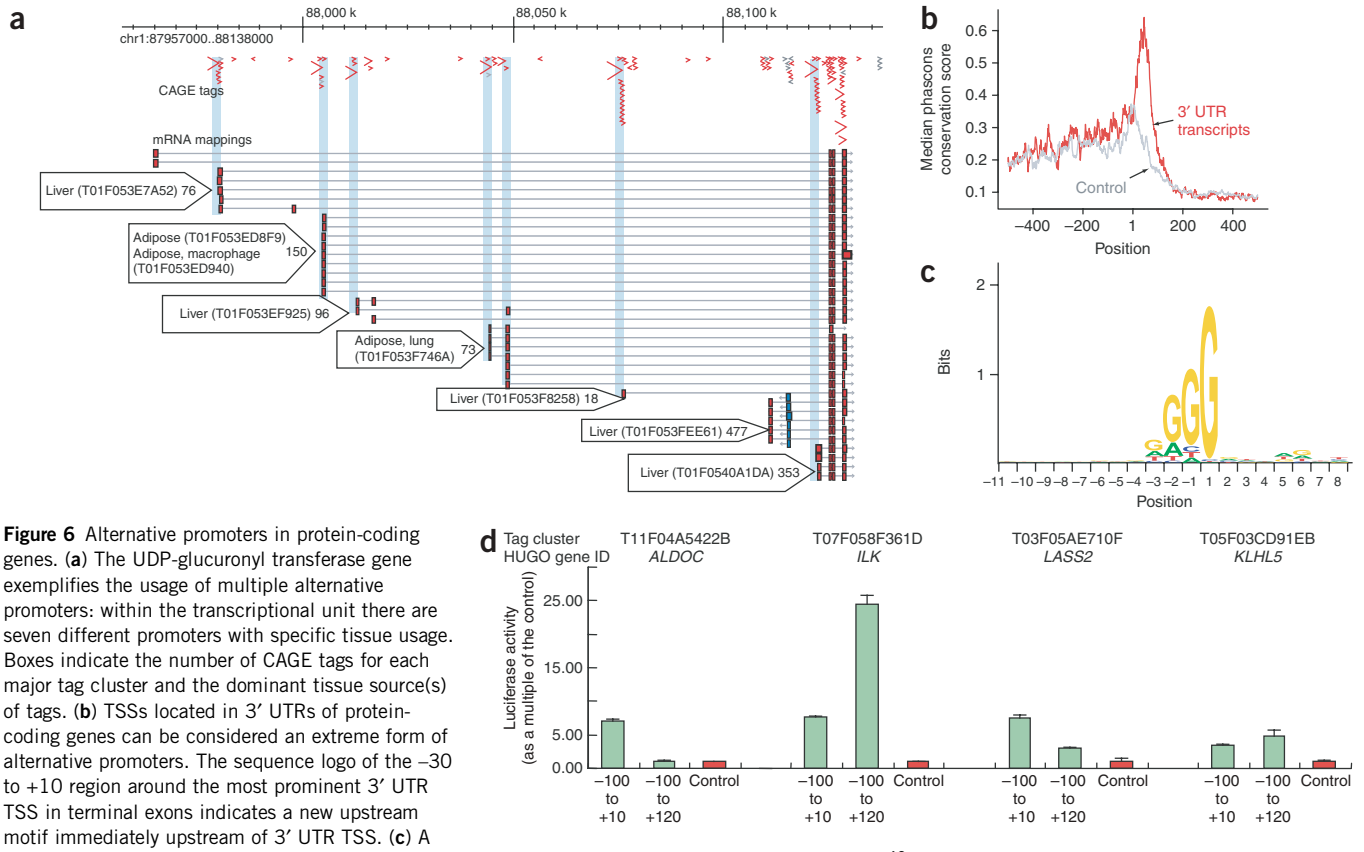

Figure 6 Alternative promoters in protein-coding genes. (**a**) The UDP-glucuronyl transferase gene exemplifies the usage of multiple alternative promoters: within the transcriptional unit there are seven different promoters with specific tissue usage. Boxes indicate the number of CAGE tags for each major tag cluster and the dominant tissue source(s) of tags. (**b**) TSSs located in 3′ UTRs of protein-coding genes can be considered an extreme form of alternative promoters. The sequence logo of the –30 to +10 region around the most prominent 3′ UTR TSS in terminal exons indicates a new upstream motif immediately upstream of 3′ UTR TSS. (**c**) A highly conserved region is located around downstream of 3′ UTR TSSs. Median phastcons conservation[40] of 1-kb region surrounding the 3′ UTR TSSs in last exons, compared to randomly selected start sites in the same exons. (**d**) Four of the 3′ UTR promoters (**Supplementary Note**) were tested with a luciferase gene reporter system. Bars indicate signal as a multiple of the signal in the empty vector control, using both a shorter upstream region (–100 to +10 relative to TSS) and a longer (–100 to +120) ahead of the reporter gene.

DISCUSSION

Broad promoters are the major class in mammals

This study contributes to the international Encyclopedia of DNA elements (ENCODE) project, which aims to identify all of the functional elements in the human genome[31] by providing a definitive survey of the classes of mammalian promoters in mouse and humans. It complements and extends published work on human systems using combinations of RACE and tiling arrays[32,33]. Our data show that the classical TATA-box promoter architecture represents a minority of the set of mammalian promoters in mouse and humans. This class is commonly associated with tissue-specific genes and high conservation across species (**Supplementary Table 2**). The BR classes, most commonly based on CpG islands, represent the majority of mammalian promoters. For the purpose of future genome and transcriptome annotation, the prevalence of the BR class of promoters in mammals means that one cannot consider the most extreme 5′ end of the longest cDNA in a cluster as the true full-length transcript, as is presently assumed in the construction of mRNA reference sequences[34]. This is an important issue in cross-species comparison, as we note that the initiator sequence is commonly subject to evolutionary change between mammals.

Promoters on exons

Using CAGE technology, we have found that the exons in a specific subset of highly expressed, multiexon genes contain putative promoters supported by CAGE tags, either singly or in clusters, around initiator-like motifs. We found that the patterns of exonic promoter activity were gene-specific, were conserved across species and were prevalent amongst TATA-containing tissue-specific genes. Notably, a recent study provides independent support for exonic promoter activity, showing by ChIP that hypophosphorylated RNA polymerase II is selectively concentrated over exons, but not introns, in a subset of human genes[35]. We can only speculate on the importance of exonic promoters. A function could be envisioned in RNA processing. The cotranscriptional recruitment of mRNA processing factors is dependent on their binding to RNA polymerase II (ref. 36). The binding of RNA polymerase II to exons could serve to recruit the entire transcribed gene to the concentrated transcriptional processing machinery in so-called 'transcription factories'. Additionally, exonic transcription initiation sites might have some relationship to so-called exonic splicing enhancers[37], either influencing recruitment of the SR proteins SF2/SAF, SC35, SRp40 and SRp55, or being influenced by them. In any case, the truncated transcripts generated from exonic promoters constitute a major new class of noncoding RNAs.

Evolutionary implications

The CpG island–associated category of promoters seems to be particularly rapidly evolving in mammals, whereas TATA box–containing promoters are more constrained. The transcriptional regulation and evolutionary plasticity of CpG island–associated promoters is also linked to epigenetic control of transcriptional activity. CpG

island–associated methylation events, including imprinting, have been related to the existence of noncoding RNA, including sense-antisense transcription[38]. Given the extensive use of both broad, CpG-rich promoters and the widespread occurrence of antisense transcripts[20], it is possible that newly evolved CpG promoters are to a larger extent epigenetically controlled. We noted that among the so-called 'bidirectional' CpG island promoters, the two opposing promoters actually frequently generate transcripts that overlap to form potential sense-antisense pairs. One transcript might also influence the epigenetic state of the promoter of the other in such pairs. The development of a promoter structure with multiple TSSs whose expression is regulated at a locus level by epigenetic events and fine-tuned by the actual initiation signals might have been an important component of the adaptive evolution of vertebrates.

Importance of proximal promoters as regulatory determinants

In general, searches for shared patterns of motifs among coregulated genes have been based on the comparison of arbitrary lengths of DNA sequences upstream of the longest known cDNA (the presumptive promoter)[39]. The availability of precise TSSs in two species will increase the accuracy of these approaches by narrowing the window to the actual phylogenetically conserved promoter used in the tissue, and excluding overlapping promoter regions. Our clustering analysis of promoters based on CAGE data and a detailed examination of the macrophage-expressed and LPS-inducible gene classes (**Supplementary Note**) shows that there is a strong correlation between core promoter sequences and tissue-specific promoter use. Hence, although proximal promoters may not contain all of the information required to precisely control transcription of individual genes in time and space during development, analysis of promoters alone can generate meaningful models of transcriptional regulatory networks.

Implications for future research

The results presented here provide a platform for future approaches to genome-wide analysis of transcription and transcriptional gene regulation. Linking transcription to defined genomic regions and better understanding of the functional landscape of different classes of core promoters will fundamentally affect all future developments in promoter identification, regulatory determinant pattern detection and analysis of clusters of coexpressed genes. Because technologies like CAGE are scalable to whole organisms, these approaches pave the way for 'systematic' systems biology.

METHODS

See **Supplementary Note**, **Supplementary Figures 6–8** and **Supplementary Tables 5–7** online for details of methods and extended biological findings. The supplementary information is also available as a single file online (http://fantom3.gsc.riken.jp or http://www.macrophages.com).

URLs. See http://gerg01.gsc.riken.jp/altp/ for a detailed description of alternative promoters.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank the following individuals for discussion, encouragement and technical assistance: H. Atsui, A. Hasegawa, K. Hayashida, H. Himei, F. Hori, C. Kawazu, M. Kojima, K. Waki, M. Aoki, K Murakami, M. Murata, M. Nishikawa, H. Nishiyori, K. Nomura, M. Ohno, H. Sato, Y. Shigemoto, N. Suzuki, Y. Takeda and K. Yoshida. We especially thank A. Wada, T. Ogawa, M. Muramatsu, A. Kira and all the members of RIKEN Yokohama Research Promotion Division for supporting and encouraging the project. We also thank the Laboratory of Genome Exploration Research Group for secretarial and technical assistance, and Yokohama City University, who provided human samples and computational resources of the RIKEN Super Combined Cluster (RSCC). This work was mainly supported by Research Grant for the Genome Network Project from the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT), the RIKEN Genome Exploration Research Project from the Japanese Ministry of Education, Culture, Sports, Science and Technology of the Japanese Government (to Y.H.), Advanced and Innovational Research Program in Life Science (to Y.H.), National Project on Protein Structural and Functional Analysis from MEXT (to Y.H.), Presidential Research Grant for Intersystem Collaboration of RIKEN (to P.C. and Y.H.) and a grant from the Six Framework Program from the European Commission (to P.C.).

AUTHORS' CONTRIBUTIONS

P.C., R.K., M.K.-K., Y.K., C.K., M.N., J.K., S.M.-T., P.A., A.C., S.G., F.P., H.S., S.M.G. and C.W. produced data and resources. P.C. and Y.H. designed the experiments. P.C., A.S., B.L. and D.A.H. wrote the paper. A.S., B.L., S.K., K.S., J.P., C.A.M.S., M.S.T., W.B.A., P.G.E., M.C.F., A.R.R.F., W.B.A., S.L.T., C.P., R.K., T.R., T.K., S.F., H. Kawaji, H. Konno, K.N., C.A.W., V.O., E.T.L., M.H., V.B.B., D.A.H. and Y.H. performed bioinformatic analysis and interpreted data.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/naturegenetics
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Bajic, V.B., Tan, S.L., Suzuki, Y. & Sugano, S. Promoter prediction analysis on the whole human genome. *Nat. Biotechnol.* **22**, 1467–1473 (2004).
2. Carninci, P. *et al.* Targeting a complex transcriptome: the construction of the mouse full-length cDNA encyclopedia. *Genome Res.* **13**, 1273–1289 (2003).
3. Okazaki, Y. *et al.* Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420**, 563–573 (2002).
4. 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. USA* **100**, 15776–15781 (2003).
5. Kodzius, R. *et al.* CAGE: cap analysis of gene expression. *Nat. Methods* **3**, 211–222 (2006).
6. Jackson, D.A., Pombo, A. & Iborra, F. The balance sheet for transcription: an analysis of nuclear RNA metabolism in mammalian cells. *FASEB J.* **14**, 242–254 (2000).
7. Carninci, P. *et al.* The transcriptional landscape of the mammalian genome. *Science* **309**, 1559–1563 (2005).
8. Suzuki, Y., Yoshitomo-Nakagawa, K., Maruyama, K., Suyama, A. & Sugano, S. Construction and characterization of a full length-enriched and a 5′-end-enriched cDNA library. *Gene* **200**, 149–156 (1997).
9. Bucher, P. Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J. Mol. Biol.* **212**, 563–578 (1990).
10. Karolchik, D. *et al.* The UCSC Genome Browser database. *Nucleic Acids Res.* **31**, 51–54 (2003).
11. Suzuki, Y. *et al.* Identification and characterization of the potential promoter regions of 1031 kinds of human genes. *Genome Res.* **11**, 677–684 (2001).
12. Kadonaga, J.T. The DPE, a core promoter element for transcription by RNA polymerase II. *Exp. Mol. Med.* **34**, 259–264 (2002).
13. Smale, S.T. & Kadonaga, J.T. The RNA polymerase II core promoter. *Annu. Rev. Biochem.* **72**, 449–479 (2003).
14. Burke, T.W. & Kadonaga, J.T. The downstream core promoter element, DPE, is conserved from *Drosophila* to humans and is recognized by TAFII60 of *Drosophila*. *Genes Dev.* **11**, 3020–3031 (1997).
15. Schneider, T.D. & Stephens, R.M. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* **18**, 6097–6100 (1990).
16. Butler, J.E. & Kadonaga, J.T. The RNA polymerase II core promoter: a key component in the regulation of gene expression. *Genes Dev.* **16**, 2583–2592 (2002).
17. Trinklein, N.D. *et al.* An abundance of bidirectional promoters in the human genome. *Genome Res.* **14**, 62–66 (2004).
18. Patton, J., Block, S., Coombs, C. & Martin, M.E. Identification of functional elements in the murine Gabp alpha/ATP synthase coupling factor 6 bi-directional promoter. *Gene* **369**, 35–44 (2005).
19. Prescott, E.M. & Proudfoot, N.J. Transcriptional collision between convergent genes in budding yeast. *Proc. Natl. Acad. Sci. USA* **99**, 8796–8801 (2002).
20. Katayama, S. *et al.* Antisense transcription in the mammalian transcriptome. *Science* **309**, 1564–1566 (2005).
21. Lenhard, B. *et al.* Identification of conserved regulatory elements by comparative genome analysis. *J. Biol.* **2**, 13 (2003).
22. Xie, X. *et al.* Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature* **434**, 338–345 (2005).
23. Keightley, P.D. & Gaffney, D.J. Functional constraints and frequency of deleterious mutations in noncoding DNA of rodents. *Proc. Natl. Acad. Sci. USA* **100**, 13402–13406 (2003).

24. Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **95**, 14863–14868 (1998).
25. 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**, 22–26 (2004).
26. Schug, J. *et al.* Promoter features related to tissue specificity as measured by Shannon entropy. *Genome Biol.* **6**, R33 (2005).
27. Sandelin, A. & Wasserman, W.W. Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. *J. Mol. Biol.* **338**, 207–215 (2004).
28. Landry, J.R., Mager, D.L. & Wilhelm, B.T. Complex controls: the role of alternative promoters in mammalian genomes. *Trends Genet.* **19**, 640–648 (2003).
29. Rosmarin, A.G., Yang, Z. & Resendes, K.K. Transcriptional regulation in myelopoiesis: Hematopoietic fate choice, myeloid differentiation, and leukemogenesis. *Exp. Hematol.* **33**, 131–143 (2005).
30. Bonizzi, G. & Karin, M. The two NF-kappaB activation pathways and their role in innate and adaptive immunity. *Trends Immunol.* **25**, 280–288 (2004).
31. ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306**, 636–640 (2004).
32. Kapranov, P. *et al.* Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays. *Genome Res.* **15**, 987–997 (2005).
33. Cheng, J. *et al.* Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* **308**, 1149–1154 (2005).
34. Pruitt, K.D., Tatusova, T. & Maglott, D.R. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **33**, D501–D504 (2005).
35. Brodsky, A.S. *et al.* Genomic mapping of RNA polymerase II reveals sites of co-transcriptional regulation in human cells. *Genome Biol.* **6**, R64 (2005).
36. Bentley, D.L. Rules of engagement: co-transcriptional recruitment of pre-mRNA processing factors. *Curr. Opin. Cell Biol.* **17**, 251–256 (2005).
37. Wu, Y., Zhang, Y. & Zhang, J. Distribution of exonic splicing enhancer elements in human genes. *Genomics* **86**, 329–336 (2005).
38. Imamura, T. *et al.* Non-coding RNA directed DNA demethylation of Sphk1 CpG island. *Biochem. Biophys. Res. Commun.* **322**, 593–600 (2004).
39. Bluthgen, N., Kielbasa, S.M. & Herzel, H. Inferring combinatorial regulation of transcription in silico. *Nucleic Acids Res.* **33**, 272–279 (2005).
40. Siepel, A. & Haussler, D. Combining phylogenetic and hidden Markov models in biosequence analysis. *J. Comput. Biol.* **11**, 413–428 (2004).

Genome-wide analysis of mammalian promoter architecture and evolution

Piero Carninci, Albin Sandelin, Boris Lenhard, Shintaro Katayama, Kazuro Shimokawa, Jasmina Ponjavic, Colin A M Semple, Martin S Taylor, Pär G Engström, Martin C Frith, Alistair R R Forrest, Wynand B Alkema, Sin Lam Tan, Charles Plessy, Rimantas Kodzius, Timothy Ravasi, Takeya Kasukawa, Shiro Fukuda, Mutsumi Kanamori-Katayama, Yayoi Kitazume, Hideya Kawaji, Chikatoshi Kai, Mari Nakamura, Hideaki Konno, Kenji Nakano, Salim Mottagui-Tabar, Peter Arner, Alessandra Chesi, Stefano Gustincich, Francesca Persichetti, Harukazu Suzuki, Sean M Grimmond, Christine A Wells, Valerio Orlando, Claes Wahlestedt, Edison T Liu, Matthias Harbers, Jun Kawai, Vladimir B Bajic, David A Hume & Yoshihide Hayashizaki
***Nature Genetics*; doi: 10.1038/ng1789; corrected 5 May 2006**

In the version of this article initially published online, the x-axis of Figure 4b was mislabeled. Specifically, the five groups on the x-axis should be labeled:

No mutation
PyPu to PuPu
PyPu to PuPy
PyPu to PyPu
PyPu to PyPy

The error has been corrected for all versions of the article.

Corrigendum: Spontaneous DNA breakage in single living *Escherichia coli* cells

Jeanine M Pennington & Susan M Rosenberg

***Nat. Genet.* 39, 797–802 (2007); published online 27 May; corrected after print 29 August 2007**

In the version of this article initially published, our estimate of the rate of formation of spontaneous DNA double-strand breaks (DSBs) in *E. coli* proportional to DNA content in humans should read that it differs from that of Vilenchik and Knudson (*Proc. Natl. Acad. Sci. USA* **100**, 12871–12876; 2003) by fourfold, not "approximately tenfold" (page 800, line 3, and page 800, line 59). We estimated that there are 0.01 DSBs per *E. coli* genome replication. Because *E. coli* has approximately 4.7×10^6 bp per genome (Blattner, F.R. *et al.*, *Science* **277**, 1453–1474; 1997), we estimate that approximately 2×10^{-9} DSBs per bp are replicated, or about fourfold fewer than the estimate of about 0.8×10^{-8} DSBs per bp replicated in human somatic cells (or 50 DSBs per diploid human genome replication) from Vilenchik and Knudson (*Proc. Natl. Acad. Sci. USA* **100**, 12871–12876; 2003). This would bring the number of DSBs per human genome replication down to approximately 13, if it were proportional to that in *E. coli*. Our error arose from calculating the human equivalent based on haploid, not diploid, human genome size. This error has been corrected in the HTML and PDF versions of the article.

Corrigendum: Genome-wide analysis of mammalian promoter architecture and evolution

Piero Carninci, Albin Sandelin, Boris Lenhard, Shintaro Katayama, Kazuro Shimokawa, Jasmina Ponjavic, Colin A M Semple, Martin S Taylor, Pär G Engström, Martin C Frith, Alistair R R Forrest, Wynand B Alkema, Sin Lam Tan, Charles Plessy, Rimantas Kodzius, Timothy Ravasi, Takeya Kasukawa, Shiro Fukuda, Mutsumi Kanamori-Katayama, Yayoi Kitazume, Hideya Kawaji, Chikatoshi Kai, Mari Nakamura, Hideaki Konno, Kenji Nakano, Salim Mottagui-Tabar, Peter Arner, Alessandra Chesi, Stefano Gustincich, Francesca Persichetti, Harukazu Suzuki, Sean M Grimmond, Christine A Wells, Valerio Orlando, Claes Wahlestedt, Edison T Liu, Matthias Harbers, Jun Kawai, Vladimir B Bajic, David A Hume & Yoshihide Hayashizaki

***Nat. Genet.* 38, 626–635 (2006); published online 28 April 2006; corrected online 5 May 2006; corrected after print 29 August 2007**

In the version of this article initially published, two of the smaller bar plots in Figure 1e were mistakenly duplicated. Specifically, the *Zfp385* plot is an erroneous copy of the *137774* plot, and the *Txndc7* plot is an erroneous copy of the *Pik3r5* plot. See below for the corrected version of the figure. This error does not change the conclusions of the study in any way, as the bar plots are just a few visual examples of more than 5,000 tag clusters, and the correct plots follow the same distribution patterns as the erroneous ones. This error has been corrected in the HTML and PDF versions of the article.