doi:10.1038/nature13182

Acknowledgements

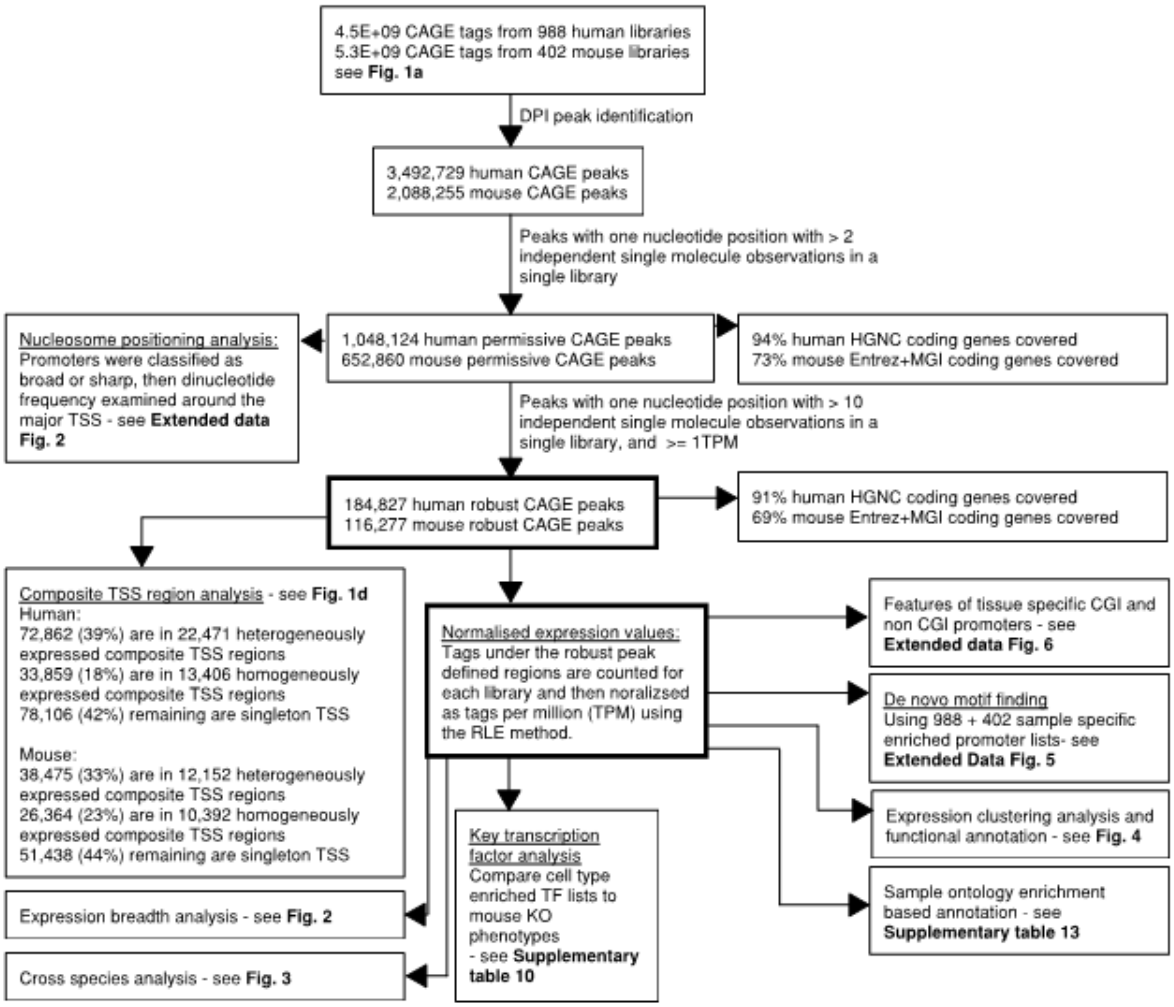
General: We would like to thank the Dutch Brain Bank for making the post mortem brain samples available. We thank the RIKEN Integrated Cluster of Clusters (RICC) for the computer resources used for the motif significance calculations. We would like to thank Fadwah Booley, Rudiger Eder, Petra Hoffmann, Alisa J. Carlisle, Rebecca Simms, Kaoru Takahashi, Noriko Yumoto, Shinji Fukuda, Takashi Kanaya, Yoshimi Tokuzawa, Yukiko Kanesaki-Yatsuka, Shinji Fukuda, Takashi Kanaya, Kaoru Takahashi, Noriko Yumoto, Mark Walker, Timothy Barnett, James Fraser, Matthew Sweet, Lisa Seymour, Nilesh Bokil, Rowland Mosbergen, Othmar Korn, Elizabeth Mason and Lars Nielsen for helping prepare samples. The CD34 cells differentiated in to the erythroid lineage were provided by Professor David Anstee and Dr Stephen Parsons, Bristol Institute for Transfusion Sciences, UK. We would also like to thank Hanna Daub, Linda Kostrencic, Hiroto Atsui, Emi Ito, Nobuyuki Takeda, Tsutomu Saito, Hiroo Inaba, Teruaki Kitakura at RIKEN Yokohama for assistance in arranging collaboration agreements, ethics applications, computational infrastructure and the FANTOM5 meetings. The authors wish to acknowledge RIKEN GeNAS for generation and sequencing of the CAGE libraries using the Heliscope (Helicos), and subsequent data processing.

Funding: FANTOM5 was made possible by a Research Grant for RIKEN Omics Science Center from MEXT to Y.Hayashizaki and a Grant of the Innovative Cell Biology by Innovative Technology (Cell Innovation Program) from the MEXT, Japan to Y.Hayashizaki. This study is also supported by Research Grants from the Japanese Ministry of Education, Culture, Sports, Science and Technology through RIKEN Preventive Medicine and Diagnosis Innovation Program to Y.Hayashizaki and RIKEN Centre for Life Science, Division of Genomic Technologies to PC. In addition the participation of consortium members was made possible through various national funding schemes listed below. CJM's work was supported by the Director, Office of Science, Office of Basic Energy Sciences, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. LMK is supported by an Australia Fellowship and a Program Grant from the National Health and Medical Research Council of Australia. BB and VO are supported by grants from Telethon Foundation (S00094TELA), Italian Institute of Technology (IIT) and Epigenomics Flagship Project EPIGEN (MIUR-CNR). BB is supported by a French Muscular Dystrophy Association (AFM) fellowship. LL has been supported by a Research Stimulation Award from Wayne State University, and by grant number 1U01-HG007031 from the ENCODE Consortium, NHGRI, NIH. This work was supported by grant numbers APP597452, APP1041294 and CDA481945 from the NHMRC Australia, Smart Futures Fellowship from the Queensland Government to CAW. DG and TH were funded by Genome British Columbia through its Strategic Opportunities Fund. This work was supported by the grant Dopaminet from the European 7th Framework Program and by the grant SEED S00094IIT1 from Italian Institute of Technology to S.Gustincich. JKB is supported by a Wellcome Trust Clinical Fellowship (090385/Z/09/Z) through the Edinburgh Clinical Academic Track. EvN acknowledges support by the Swiss National Science Foundation and the Swiss Institute of Bioinformatics. PJB and EvN are both supported by the Swiss Systems Biology Initiative SystemsX.ch within the network ‘‘Cellplasticity’’. ZT, AG, M.Thompson, JFJL, PACH, and EAS are supported by a grant from the Centre for Medical Systems Biology within the framework of the Netherlands Genomics Initiative/Netherlands Organisation for Scientific Research. EAS is supported by grants from the Concept Web Alliance and the John Templeton Foundation (The opinions expressed in this publication are those of the author(s) and do not necessarily reflect the views of the John Templeton Foundation). This work was supported by grant number 634494

from the NHMRC to PK LNW and ARRF. A.Sandelin is supported by grants from the Lundbeck Foundation, the Novo Nordisk Foundation, The Danish Cancer Society and European Research Council (FP7/2007-2013/ERC grant agreement 204135). ASBE and JSK were supported by NIH grant RO1 DC007174. BD is supported by grants from the Swiss National Science Foundation, from the Japanese-Swiss Science and Technology Cooperation Program (ETHZ), and by Institutional support from the Ecole Polytechnique Fédérale de Lausanne (EPFL). BM is supported by KAUST AEA Grant of VBB and KAUST CBRC Funds. FB, S.Savvi, A.Schwegmann, and RG were supported by grants from the South African Research Chair initiative, DST, NRF and South African Medical Research Council (MRC). IVK is supported by the Dynasty Foundation Fellowship and Russian Foundation for Basic Research grant [14-04-01838_a]. J.Kere and A.Sajantila were supported by the Sigrid Jusélius Foundation and Academy of Finland. J.Kere is the recipient of a Distinguished Professor Award at Karolinska Institutet. KM was supported by Precursory Research for Embryonic Science and Technology (PRESTO) from the Japan Science and Technology Agency (JST), a Grant-in Aid for Young Scientist (A) (22689013) from the Japan Society for the Promotion of Science (JSPS), and a Grant-in-Aid for Scientific Research on Innovative Areas (23118526) from the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan. ME is supported by grants from Deutsche Forschungsgemeinschaft and BayImmuNet. MO is supported by grants from the Japan Society for the Promotion of Science (KAKENHI, #21592637, 24593129). M.Rehli is supported by grants from Deutsche Forschungsgemeinschaft and Rudolf Bartling Foundation. MST, CAS, SB, AM and JGDP are supported by the Medical Research Council of the UK. MV is supported by an IPA scholarship from RIKEN and Frankopani Fund Grant. PA is supported by grants from Swedish Research Council and Novo Nordic Foundation. S.Koyasu is supported by a grant; "Research Program of Innovative Cell Biology by Innovative Technology", from the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan. S.Koyasu was supported by a Grant-in-Aid for Scientific Research (S) (22229004) from the Japan Society for the Promotion of Science (JSPS), Japan. S.Schmeier and YAM are supported by KAUST Base Research Funds of VBB. The Roslin Institute is supported by an Institute Strategic Programme Grant from the Biotechnology and Biological Resources Council of the UK. This work was supported by a Japan Partnering Grant from the Biotechnology and Biological Resources Council of the UK to GJF, DAH, KMS and TCF. GJF acknowledges the support of a New Investigator Award from the British BBSRC (BB/H005935/1) and a C.J. Martin Overseas Based Biomedical Fellowship from the Australian NHMRC (575585). This work was supported by grant number NIH NHGRI P41 HG-002273-09S1 to JAB. This work was supported by grant number R01 DE022386-01 from the NIH to MCFC. US, BRJ, IA and JACA are supported by KAUST CBRC Funds. VBB is supported by KAUST AEA Grant and KAUST Base Research Funds. VJM is supported by grants from Presidium of the Russian Academy of Sciences Program in Cellular and Molecular Biology, Presidium of the Russian Academy of Sciences Fundamental Research Subprogram ‘Gene pools dynamics and conservation’, and Russian Ministry of Science and Education grant [11.G34.31.0008]. WWW and A.Mathelier were supported by NIH grant R01GM084875. Work from LNCIB/Laboratorio Nazionale Consorzio Interuniversiatrio Biotecnlogie is supported by a grant from AIRC Special Program Molecular Clinical Oncology ‘‘5 per mille” to CS. YY is supported by a grant from Sato Fund, Nihon University School of Dentistry. WL was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2012R1A1A2007017). This work is supported in part by a Grant of the Cell Innovation Program from MEXT to MO-H. AJK is

supported by the Wellcome Trust UK. MH, RS, SZ were funded by NIH grant RO1CA-047159. MF is supported by a grant from Portuguese Foundation for Science and Technology. AP & M.Rashid were supported by KAUST faculty baseline and KAUST-OCRF funds and a SABIC post-doctoral fellowship to M.Rashid. BK was supported by the Japan Society for the Promotion of Science (JSPS) Postdoctoral Fellowship for Foreign Researchers.

Figure S1. Flow diagram showing relationship between tags, peaks and analyses.
The analysis starts by peak calling across the human and mouse libraries. Key points are highlighted by the thick boxes. Note, the majority of analyses are carried out using the robust peak set, and RLE normalised expression values.



Methods

1. Ethics, sample collection, RNA extraction and quality control

For information on specific samples, all of the following information is summarised in **Supplementary Table 1**.

Human Ethics

All human samples used in the project were either exempted material (available in public collections or commercially available), or provided under informed consent. All non-exempt material is covered under RIKEN Yokohama Ethics applications (H17-34 and H21-14).

Mouse samples

Mouse tissue samples were collected as per RIKEN Yokohama institutional guidelines. Mouse primary cells were collected as per our collaborators Institutional guidelines and shipped as either purified RNA or as guanidinium isothyocyanate lysates (Trizol, Isogen or Qiazol) which were then purified using the miRNeasy kit (QIAGEN).

Primary cells

The majority of human and mouse primary cell samples were purchased as purified RNA from Cell Applications, 3HBiomedical or Sciencell. Additional primary cells were also purchased from Cell systems, CET, Lonza, Promocell, Sciencell, Stem cell technologies and Xenotech. These were cultured as per the manufacturer's instructions, and then RNA extracted using the miRNeasy kit (QIAGEN). The remaining primary cell samples were provided by the FANTOM5 collaborator network to the OSC as either purified RNA or as guanidinium isothyocyanate lysates (Trizol, Isogen or Qiazol) which were then purified using the miRNeasy kit (QIAGEN). Human salivary acinar cells were isolated as described previously[1]. Human sebocytes were prepared as described previously[2]. Human epithelial cell rests of Malassez (ERM)-derived epithelial cells, gingival epithelial cells, gingival and periodontal fibroblasts were prepared as described previously[3]. Mouse tracheal epithelial cells were prepared as described previously[4]. Human dermal lymphatic endothelial cells were prepared as described previously[5]. Mouse regulatory T cells were prepared as following. C57BL/6JJcl mice and Balb/cAJcl mice were purchased from CLEA Japan (Tokyo, Japan). CD4+ T-cells were isolated from splenic and lymph node as previously described[6]. CD4+CD25+ T-cells (T-reg cells) and CD4+CD25-CD44low T-cells (T-conv cells) were purified by sorting with a cell sorter (MoFlo, Beckman Coulter). For in vitro TCR stimulation of T-conv cells, plate coated anti-CD3 (1mg/ml) and anti-CD28 (1mg/ml) for 6hrs or phorbol 12-myristate 13-acetate (20ng/ml) and ionomycin (1uM) for 2hrs were used. Whole Blood, CD19+ B-cells and CD8+ T-cells were also prepared from anonymous donors over several (2 or 3) donations. RNA from whole blood was prepared using the Ribopure blood kit from Ambion. CD19+ B-cells and CD8+ T-cells were isolated using the pluriSelect bead system (huCD4/CD8 cascade and huCD19 single; pluriSelect Germany) and RNA then extracted using the miRNeasy kit (QIAGEN).

Human Post mortem tissue RNAs

The majority of human post mortem tissues were purchased from Ambion, Biochain, and Clontech. Universal RNA mixtures were also purchased from the above and SABiosciences. Human postmortem brain RNA samples from the Dutch Brain bank were collected by P. Heutink and P. Rizzu (exempted public collection). Remaining post-mortem tissue samples collected

under ethics (H17-34) were provided by J. Kere, A. Bonetti, and A. Sajantila. The tissues derived from human cadavers were snap-frozen in liquid nitrogen. The frozen tissues were transferred into Lysing Matrix D tubes (MP Biomedicals) containing 800µl of chilled Trizol (Gibco) each. The tissues were disrupted using the FastPrep Homogenizer (Thermo Savant) according to the manufacturer's instructions. After homogenization the tubes were centrifuged at 12,000g for 15 minutes at +4°C. The supernatants were transferred to a sterile 1.5 ml eppendorf tubes and kept at -90°C until shipped in dry ice to RIKEN Yokohama for further analyses.

Cell lines

The cell lines used are all available from public repositories (RIKEN BRC (http://www.brc.riken.jp/lab/cell/english/), ATCC (http://www.atcc.org/), Coriell (http://ccr.coriell.org/), ECACC (http://www.hpacultures.org.uk/collections/ecacc.jsp), and Japan Health Sciences Foundation - Health Science Research Resources Bank (http://www.jhsf.or.jp/English/index_p.html)). COBL-a[7] and HEK293-SLAM[8] cells are available on request from C. Kai. TSt-4/DLL1 feeder cells and EBF KO HPCs are available from T. Ikawa[9]. J2E cells are available on request from K.P. Klinken[10]. Aliquots of HeLa-S3, HepG2, K562 and GM12878 RNAs used by the ENCODE consortium were provided by Carrie Davis and Thomas Gingeras. Briefly, frozen cell line stocks were rapidly thawed at 37°C, diluted in 10ml 37°C PBS, pelleted, and RNA directly extracted using the miRNA easy Kit.

Quality control

Working with large numbers of samples from multiple collaborators and companies brings about 3 potential issues of QC (RNA quality, library depth and sample identity). RNA Quality: Degraded RNA can affect the quality of CAGE libraries affecting both the promoter hit rate and the complexity of transcript species measured. To address this, RNA integrity measurements were made using an Agilent Bioanalyser for samples with more than 1ug of RNA available. 97% of the samples used in the study had RIN above 6.8. For low quantity libraries this step was skipped so not to waste RNA and library quality metrics used instead. Library depth: shallow libraries can lead to false negative calls on gene expression. For the purposes of the gene expression analyses used in this paper libraries needed to contain at least 500,000 successfully aligned reads (mapping quality is 20 or more, and sequence identity is 85% or more) mapped tags. However for peak calling shallow libraries were also used (the logic being that cell specific peaks found in these shallow libraries could still be captured). In addition the fraction of mapped tags falling within the robust peak regions were used as an additional metric for library quality. Sample identity: finally sample hierarchical clustering and marker gene checks were used to confirm or refute the identity of samples. Samples where their identity was in doubt were excluded from the expression analysis and labeled as unconfirmed_sample (they were however used in the peak calling).

Libraries with very poor quality RNA and low promoter hit rates are not listed in the supplementary however we note that one set of profiles from post-mortem donors were largely discarded due to poor RNA quality and low promoter hit rates. A few samples from the same donor were used for peak calling however they were excluded from the expression analysis. **Supplementary table 1** provides these quality metrics for all samples used in the study.

2. Single molecule CAGE

We prepared CAGE libraries for single molecule sequencing as described previously[11,12]. The

standard preparation was done using 5 ug of total RNA by manual and automated protocols. For low quantity samples (1 ug or less), we used a low quantity manual protocol. All CAGE libraries for single molecule sequencing were measured by OliGreen fluorescence assay kit (Life Technologies), and then 3 ng aliquots were subjected to poly-dA tailing reaction with terminal transferase and dATP, followed by blocking with ddATP. Poly-dA tailed libraries were then applied on HeliScope sequencers following the manufacturer's instructions (LB-016_01 and LB-017_01). Sequencing on HeliScope Single Molecule Sequencer was done according to the manufacturer's manual, LB-001_04.

3. Data Processing of Heliscope CAGE data

Sequenced Heliscope reads have a high sequencing error rate (~5%), vary in length and lack an estimation of base qualities. Combined these factors make the data processing challenging. As an initial step we removed reads corresponding to ribosomal RNA. We accomplish this by directly aligning each read against the whole human (mouse) ribosomal DNA complete repeating unit and discarding all reads with an edit distance smaller or equal to two. For this purpose we implemented Myers' bit parallel dynamic programming algorithm[13] in the program rRNAdust (author: T. Lassmann). For computational efficiency we further parallelized this algorithm using both SIMD instructions and threads. All remaining CAGE reads were mapped to the genome (hg19 and mm9) using Delve, a probabilistic mapper[14]. In brief, Delve uses a pair hidden Markov model to iteratively map reads to the genome and estimate position dependent error probabilities. After all error probabilities are estimated, individual reads are placed to a single position on the genome where the alignment has the highest probability to be true according to the pHMM model. Phred scaled mapping qualities[15], reflecting the likelihood of the alignment at a given genome position, are also reported. Reads mapping with a quality of less than 20 (<99% chance of true) were discarded. Furthermore, we discarded all reads that map to the genome with a sequence identity of less than 85%.

4. Peak analysis of the CAGE profiles

4.1 Identification and selection of CAGE peaks

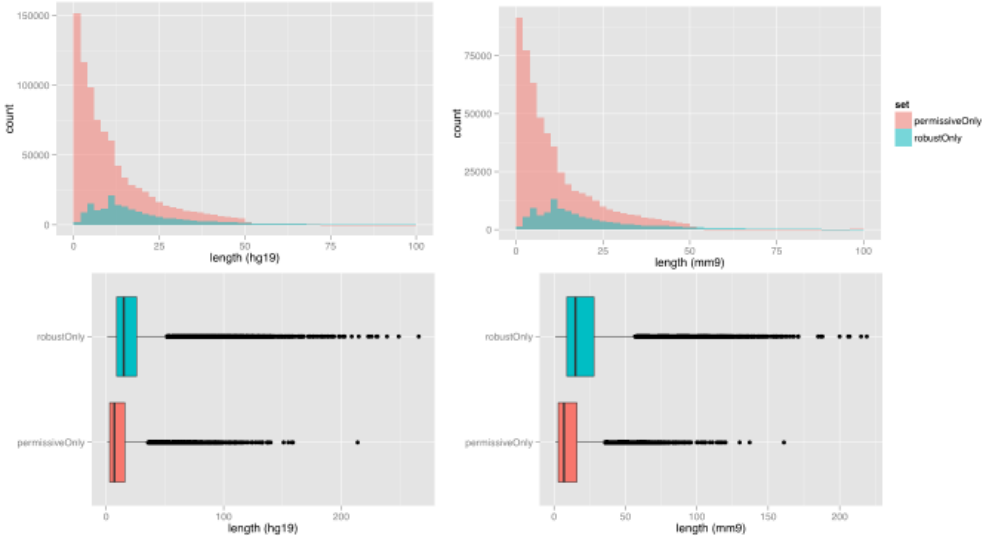
To identify peaks in the CAGE profiles, we developed a new method called decomposition peak identification (DPI, Kawaji *et al.* in prep, source code available at https://github.com/hkawaji/dpi1/). The method consists of the following steps: (i) identify local regions producing signals continuously along the genome, (ii) estimate a limited number of CAGE profiles underlying the whole observed biological states by independent component analysis[16], and (iii) determine peaks based on the estimated profiles. In the first step, we started from all the CAGE profiles (998 and 394 samples for human and mouse respectively) and selected the single nucleotides (CAGE tag starting sites; CTSS) supported by 2 or more CAGE read 5'-ends in a single profile. The selected CTSSs were grouped into tag clusters if they were neighboring within 20bp as in a previous study[17] however, since the depth of sequencing and coverage of biological states was greatly extended from the previous study[17], we found that such a simple approach tended to produce very long tag clusters which merged multiple transcription initiation events. To segregate the resulting regions into distinct or distant transcription initiation events, in the second step, we selected long (50bp or more) and abundantly observed (50 counts or more) tag clusters. We performed independent component analysis (ICA[16]) on each of the selected tag clusters, to estimate representative CTSS intensity patterns along the local genomic region underlying all the sample profiles. The number of components used in ICA, which is

bounded by up to five, is determined depending on each tag cluster by considering the number of principle components explaining 95% of variance in the CAGE signals at most and the number of resulting independent components consisting of more than 10% of the CAGE signals. We found modest but continuous CAGE signals in proximity to very active CTSSs, and we subtracted 10% of the highest CTSS signal in each of the tag clusters within individual CAGE profiles used for the estimation of representative CTSS patterns. In the third step, we identified areas where signal intensities were higher than median over each of the estimated profiles. We aggregated the identified regions on individual estimated profiles when overlapping. Lastly, we combined the aggregated regions with the tag clusters that were not selected in the second step. As a result of the three steps, we obtained ~3.5 and ~2 million peaks in human and mouse.

A substantial fraction of the peaks identified above had very limited tag support and were located in exonic regions, while the majority of known transcript 5' ends were well supported by peaks with many tags. To enrich for promoter associated signals, we examined thresholds in expression levels at individual single CTSSs, with the thesis that genuine TSS are likely to reproducibly use the same position, whereas random degradation should be spread more broadly along the transcript. We set this by examining the ratio of peaks that were near 5' ends of known transcripts (within 500bp) versus peaks that were within internal exons (but not promoter). We settled on two thresholds the first a **permissive** threshold gave a ratio of promoter to exonic peaks of ~0.7 and corresponded to the subset of peaks with a single CTSS in a single experiment supported by 3 or more observations in at least one profile, and a **robust** threshold yielding a ratio of ~2.0 and corresponding to peaks with a single CTSS in a single experiment supported by 11 or more observations and 1 or more TPM. Although the thresholds are based on single nucleotide positions, the total number of observations (reads) in each CAGE peak is substantially more (see below). We provide the permissive set to the research community for TSS exploration; however for the majority of the manuscript we use the higher trust robust set for further analysis.

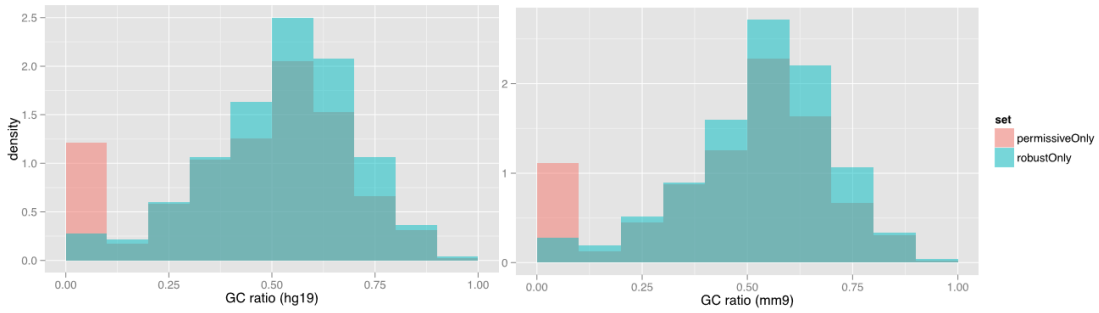
We examined several properties of the CAGE peaks below, such as peak length, GC content, low complexity regions, and maximum read count. The distribution of peak length demonstrates that the majority of the permissive peaks are very small (shorter than three base pairs), while length distribution of the robust peaks has a longer tail up to 300 base pairs. This demonstrates that the DPI peaks represent subcomponents of the broad promoters rather than broad complex promoters themselves.

Figure S2: Length distributions of the CAGE peaks



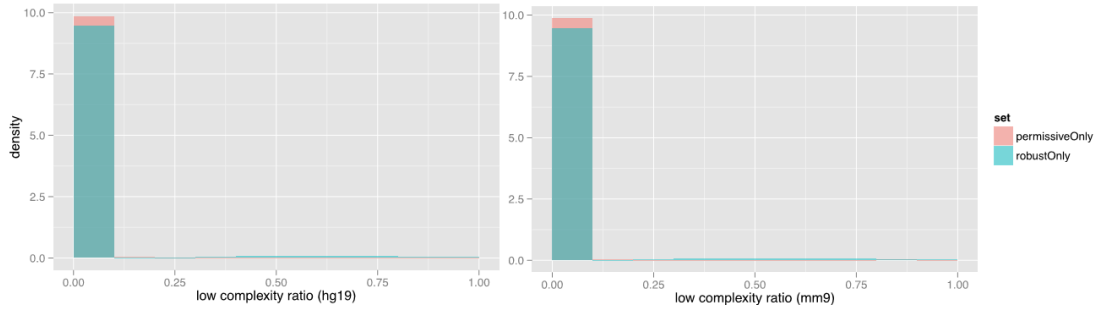
The GC content plots suggest that the GC contents are largely consistent between the robust peaks and the permissive ones, except for the ratio of peaks consisting with G or C only. The difference likely comes from the nature of individual peak sets, where true TSSs are further enriched in the robust set as shown below (4.2).

Figure S3: Ratio of G or C nucleotides within the CAGE peaks



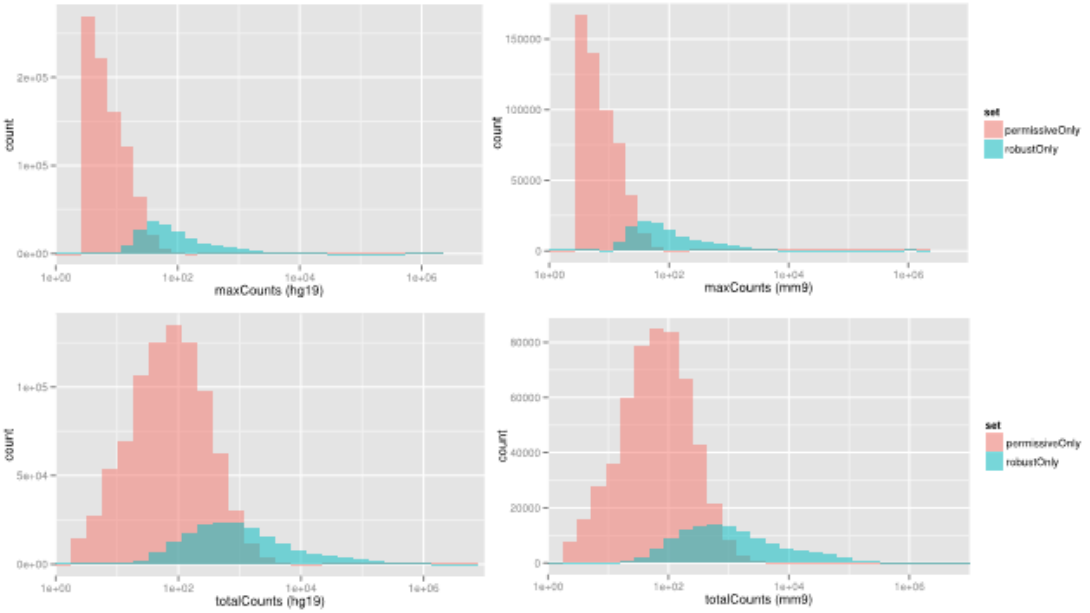
One could assume that the permissive set may consist of less reliable alignments due to low complexity of the genomic sequences. We measured the ratio of low complexity sequences in individual peak regions identified by NSEG with default parameters[18], and the result shown in the plot below suggest almost all of the peaks consist of middle- or high-complexity sequences.

Figure S4: Ratio of low complexity sequences within the CAGE peaks



Finally, we examined read counts actually observed in the spanning regions of individual peaks, (our threshold of read counts for the robust and the permissive set is applied to individual TSS within a peak but under the entire span of the peak there can be considerably more tags). The plots indicate maximum and total read counts of the peaks across all the CAGE profiles, and indicate the robust peaks are supported and quantified by substantial observations (typically more than 100 reads).

Figure S5: Maximum and total read counts of the CAGE peaks

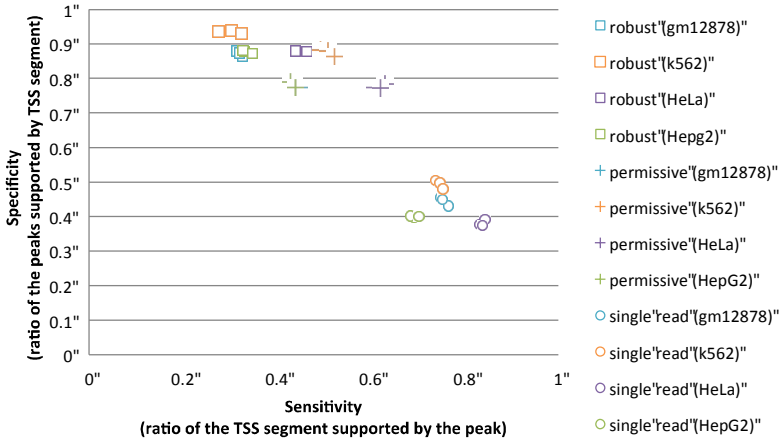


4.2 Comparison of identified CAGE peaks to chromatin state defined candidate TSS regions

Recently the ENCODE consortium have extensively used computational classifiers for genome segmentation that integrate genome-wide datasets on chromatin states to assign biological interpretations along the genome, which includes predictions of TSS segments. To examine whether the peaks identified by the FANTOM5 CAGE profiles were independently supported by these predictions based on chromatin states we compared CAGE peaks identified in four of the ENCODE cell lines (HepG2, K562, GM12878, HeLa-s3 were run as biological triplicates within the FANTOM5 dataset, using matched RNA supplied by ENCODE members) with the TSS segments identified by ENCODE[19] in the same cell lines using chromatin marks. For this analysis we selected a segmentation track integrating ChromHMM[20] and Segway[21] as reference, and considered a CAGE peak and a TSS segment as 'closely located' if they are within 1000bp (note: ChromHMM provides its results at 200bp resolution and we set the 1000bp threshold to make the two distinct datasets comparable). In addition, for this specific analysis we applied the robust and permissive thresholds on the individual CAGE profile being compared to the corresponding chromatin marks (e.g. robust in K562 replicate 1 means there is a position with 11 or more reads from K562 replicate 1). The result (Fig. below) indicates that our thresholds are very strict in general. Of the CAGE peaks supported by the robust threshold, ~90% or more peaks are supported by the TSS segments. The remaining peaks (~10%) could be explained by difference of method and/or general limitations of large-scale studies. The result indicates that the robust CAGE peaks represent true TSSs with high confidence. For the permissive threshold, ~80% or more peaks are supported by TSS segments. Even the permissive peaks represent true TSSs with substantial confidence. Conversely however only 30-40% of the TSS segments called by ENCODE in these cell lines also had CAGE peaks above the robust or permissive thresholds within the corresponding profiles. This could suggest a high false negative rate in the FANTOM5 peaks but could also suggest that the ENCODE TSS segments are not genuinely active or are transcribed at very low levels (Note: this is discussed in more detail in the main text). Up to 80% of ENCODE TSS regions are covered by at least one CAGE read, but at the expense of lost

specificity. Overall, the analysis demonstrates that our thresholds are very strict and the selected CAGE peaks represent active promoters with high confidence (~90% for the robust set, and ~80% for the permissive set).

Figure S6: Specificity and sensitivity of the CAGE peak thresholds (ENCODE integrated segmentation tracks are used as reference)



4.3 Assessment of decomposed peaks

DPI is designed to decompose larger clusters if they are composed of CAGE peaks with distinct expression profiles. To assess the performance of this we re-grouped peaks that were within 100bp of another into putative 'composite promoters' and tested whether the expression profiles were indeed distinct. The grouping identified 35,877 composite promoters consisting of two or more peaks (corresponding to 106,721: 58% of the human robust peaks). The remaining 78,106 (42%) robust peaks were 'singleton peaks' more than 100bp from another peak.

Then we asked if read counts of the different peaks in the same composite promoter arose from the same expression pattern over the profiled biological states, by using the likelihood ratio test where the read counts are modeled as a negative binomial distribution. We set its over-dispersion parameter as 0.06 (corresponding to ~25% standard error), which is a little larger than experimental estimation by edgeR[22] (0.026 for human, and 0.056 for mouse biological replicates; see the next section) to make our assessment conservative. With FDR < 1% threshold, we found 22,471 composite promoters consisting of multiple peaks with non-identical expression patterns, which corresponded to 72,862 (39%) robust peaks. The remaining 13,406 composite promoters had multiple peaks (33,859) with expression patterns too similar to discriminate using the above criteria (see main **Fig. 1d**). Running the same analysis on the mouse robust peaks, found equivalent results. Note that the peak identification method described above (DPI) considers heterogenic transcription by estimation of underlying multiple profiles, and this result confirmed that a majority (~eighty percent) of the robust peaks represent their own transcription initiation events based on a conventional statistical method applicable only after peaks are identified.

4.4 Quantification of transcription initiation activities (peak expression profiles)

Using the robust peaks defined above we counted tags which 5' end alignments (mapping quality >=20, percent identity >= 85%) started within the boundaries of individual robust peaks. We selected 889 human profiles and 389 mouse profiles which had a minimum of a half million mapped CAGE reads for this expression analysis, since shallow profiles would not be very

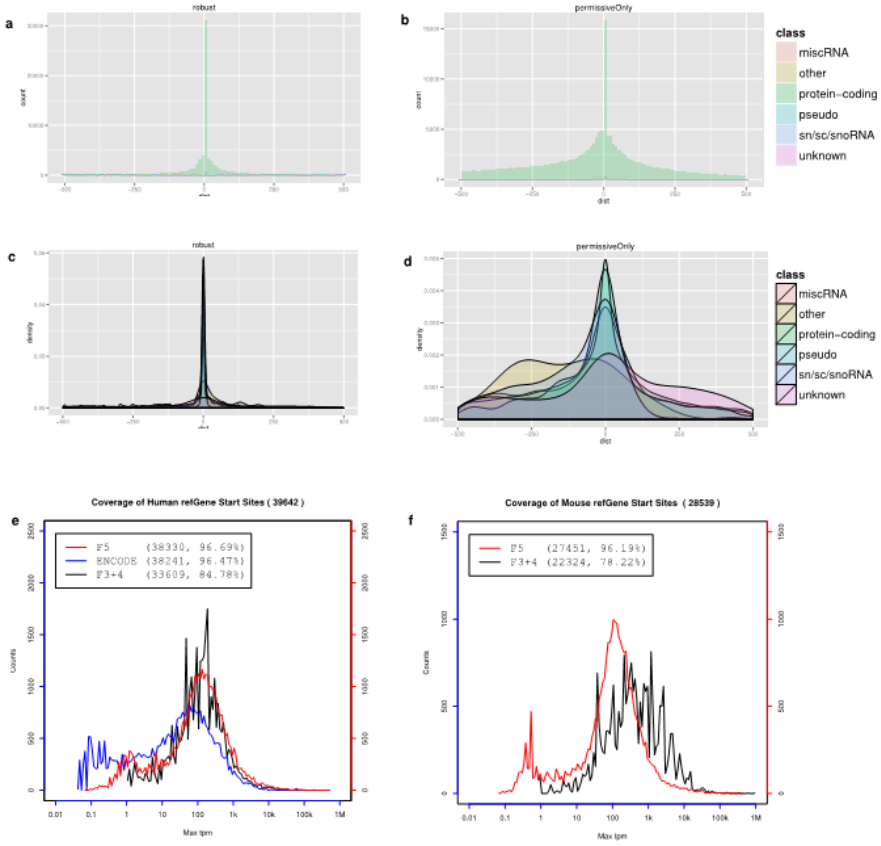
reliable for quantification of TSS activities. We counted the CAGE reads arising from individual CAGE peaks in each of the selected profiles and normalized the counts as TPM (tags per million) based on the library size and normalization factors estimated by edgeR[22] using the relative log expression (RLE) method[23]. All of the expression analyses in our paper are based on these expression values.

Based on the quantified expression above, we assessed the variability of biological replicates in our dataset (replicates from multiple donors were available for most of the primary cells). We estimated the overdispersion parameter (common dispersion) by using edgeR[22], and found 0.026 for human and 0.056 and for mouse, which corresponds to 16% ~ 24% of standard error. In comparison technical variability of HeliScopeCAGE sits at 5.3% standard error (that is, overdispersion parameter 0.003) described in another study (Kawaji et al. in press). Thus the biological variability is larger than technical variability and the variability across replicates is roughly 25% in the dataset overall.

4.5 Gene associations

As expected many of the CAGE peaks are very close to (or overlapping) known TSSs on the genome.

Figure S7: Association of CAGE peaks to annotated TSS.
A 500bp threshold was chosen for known TSS association. The plots show the distribution of the distances between CAGE peaks and annotated TSSs, according to the classes of the associated EntrezGene entries. **a**, robust count distribution. **b**, robust density distribution **c**, permissive (non-robust) count distribution, **d**, permissive (non- robust density distribution. Note, mouse has similar distributions (not shown). Coverage of Refseq 5' ends in the FANTOM3, 4 & 5 and ENCODE CAGE datasets for **e**, human, **f**, mouse



To systematically annotate them based on their relationships with known genes and transcripts, we compared them to the following gene models downloaded from the UCSC Genome database January 2012: RefSeq[24], , UCSC known gene[25], Gencode V7[26] transcripts (for human), ENSEMBL[27] transcripts (for mouse), and full-length mRNA tracks. CAGE peaks were assigned to a gene or transcript if their 5' end was on the same strand and within 500bp of the 5' end of the transcript model. In this process, gene models whose 5'-ends do not correspond to transcription starting sites (e.g. snoRNA, snRNA, and miRNA 5' ends result from cleavage of primary transcripts) were given lower priority. From the transcript and gene associations we further extended the annotation and provided HGNC gene symbols, EntrezGene IDs, and UniProt IDs (if coding) according to their association with the selected gene models. The tables below show the number of genes associated with each reference model and different Entrez gene classes.

Number of peaks associated with genes

		Transcript (RefSeq, GENCODE/ENSEMBL, UCSC Known Genes, mRNAs)	*Protein (UniProt)*	*HGNC*	*EntrezGene*
human	*permissive*	294,765	136,741	245,829	245,514
	robust	93,558	56,011	82,257	82,150
mouse	*permissive*	146,148	101,130		131,998
	robust	61,072	47,755		56,744

Number of peaks associated with EntrezGene categories

	human	*protein coding*	*pseudo*	*miscRNA (incl. miRNA)*	*snRNA scRNA snoRNA*	*other unknown*
human	*permissive*	237,424	1,254	5,808	454	818
	robust	79,735	489	1,755	126	163
mouse	*permissive*	127,445	949	4,098	64	79
	robust	55,217	435	1,356	22	16

All peaks in the dataset have persistent names consisting of chromosome, chromosomal co-ordinates, and strand (anchored to build Hg19 and Mm9 of the human and mouse genomes), however to aid researchers familiar with gene names we assigned peak names consisting of a peak number and a gene symbol where available. We discussed nomenclature of transcription starting sites in the FANTOM5 consortium extensively and reached a consensus that peak names should consist of gene symbols and numbers, allowing us to distinguish individual peaks used by a gene. We could not however find any optimal numbering scheme that would circumvent updates. For example, A) if we decided to call a peak with the highest expression level as the first TSS, it would not necessarily be the most active TSS if we change the biological states we profiled, B) if we numbered them based on proximity to the gene, new transcripts and new promoters would change the ordering. We decided that any numbering would be arbitrary, and therefore took a very simple approach: we numbered the CAGE peaks associated with the same gene according to the number of total read counts. For example, we named the peak chr10:102820579..102820585,+ as "CAGE peak 5 at KAZALD1 5end" (p5@KAZALD1 in a short form), since it is fifth peak in terms of total read counts within the peaks associated with KAZALD1. We admit this is arbitrary choice, but this is similar to the situation of annotated transcript variants (isoforms), and yet this is still useful to the scientific community for exchanging our observations on transcription starting sites. Finally peaks that were not assigned to known genes were given the short form (p@chr... as a peak name).

5. Sample ontology creation and sample ontology enrichment analysis (SOEA)

Ontology creation

The FANTOM5 Sample Ontology was generated using a combination of automated and manual methods. First, each sample description was scanned for occurrences of terms from a number of open biological ontologies: CL[28] (cell types), Uberon[29] (tissues and gross anatomy), DO[30] (diseases) and EFO[31] (treatment types). We used the Biomedical Logical Programming Toolkit (http://blipkit.org) for the entity matching. The matched terms were then used to automatically annotate each sample, creating a composite description. The annotations were validated and augmented by the ontology curators using a combination of visual inspection, and curation using the OBO-Edit[32] ontology creation tool. This was performed iteratively, with each FANTOM5 update, with the updated version released to consortium members each time. Consortium members also performed additional validation, in some cases leading to upstream fixes in the cell and anatomy ontologies.

Sample ontology enrichment analysis

To summarize promoter activities (expression profile of a TSS region) across ~1000 samples, we performed enrichment analysis based on FANTOM5 Sample Ontology (FF ontology). The question here is "in which type of samples the promoter is more active". To answer this question, we compared expressions (TPMs) in the samples associated with a sample ontology term and the rest of the samples by using the Mann-Whitney rank sum test. We iterated this test for all the ontology terms and selected only the terms with false discovery rate[33] below 1%.

To summarize ontologies enriched in a particular co-expression cluster, we ran the same analysis above except it was carried out on an averaged expression profile of all promoters that make up the cluster. The averaged expressions are calculated as followings: (i) calculate median TPM value in a promoter, (ii) produce fold changes to the median value, and convert their logarithm, and (iii) averaged the resulting values across promoters (TSS regions).

6. Supervised TSS classification using random decision tree (RDT) ensembles.

A training set comprised of both positive and negative sequences was extracted from the data. Gaussian models were trained to capture the relative distribution of 4-mer occurrences surrounding annotated DPI clusters. Each sequence was scored against all models resulting in a 256 wide vector of values for each sequence. The latter together with the cluster label was used to construct a random decision tree ensemble model[34,35]. Finally, the RDT model was used to classify test sequences not used in the training of any models. To obtain final prediction scores, we performed 2,4,6,8,10-fold cross validation, each five times, and averaged the predictions for each cluster over all runs in which the cluster was not used for training. We plotted ROC curves to assess the accuracy of our classifier using the pROC software[36]. All novel TSS clusters were counted as false positives making this assessment very strict. Compared to known gene models our methods achieved an AUC of 0.93 in human (**See Fig S17** in **Supplementary note 2**) and 0.94 in mouse. To derive thresholds on our predictions we determined the prediction score at which sum of sensitivity and sensitivity is maximal. We also compared the set of TSS classified DPI clusters to ENCODE genome segmentation tracks based on chromatin marks (**See Fig S17 c**). In all four cell lines used here, the TSS classified set contained the largest fraction of clusters overlapping promoter segments as defined by ENCODE chromatin marks. The source code is available on sourceforge (http://sourceforge.net/projects/tometools/).

7. *De novo*-motif discovery.
The cell state-specific and total robust CAGE cluster sets were searched for enriched motifs using four independent *de novo* algorithms as outlined below.

Discovery of cell type- or tissue-specific sequence motifs using ChIPMunk
Cluster selection: CAGE clusters identified in all samples of the human dataset with ≥ 0.5 mio. reads by method published by Yu *et al.*[37] were used as initial data. We selected the clusters with the expression enrichment greater than 5 and the strict 1e-5 *P* value cutoff. The ChIPMunk motif discovery pipeline was applied independently to the list of TSS-clusters for each sample. For each TSS-cluster we extracted DNA regions from 300bp upstream of the cluster start to 100bp downstream of the cluster end.

Motif discovery: ChIPMunk[38] allows incorporating prior positional information to account for motif positional preferences. For each sequence we generated trapezium-shaped positional priors equal to zero on both sequence ends and having the maximal height along the TSS-cluster extent. The height of the trapezium corresponded to the TSS-cluster expression value. This procedure allowed us to search for motifs mostly associated with the highly-expressed clusters preferably located close to transcription start sites. Two background models were used: (i) the uniform nucleotide frequency distribution (0.25 for all nucleotide frequencies) and (ii) the average composition for all sequences related to the particular sample. The basic ChIPMunk procedure identifies a single motif for a sample, so the ChIPHorde add-on was used to execute ChIPMunk several times and detect up to 3 distinct motifs by masking with poly-Ns the motif hits identified in the previous run. The final results included a maximum of 6 motifs per sample detected in 2 ChIPHorde runs with 2 defined background models. The maximum motif length was fixed at 12bp; a single-box informative prior was used for motif discovery as in[39]. For each detected motif, ChIPMunk reported the Kullback Discrete Information Content and the weight of the alignment (taking into account the weights of the sequences derived from trapezium-profile heights).

Motif filtering: To produce a non-redundant motif list, all ChIPMunk motifs were sorted according to the following criteria: the masking step (see below), the motif length (starting from the longest motif and then selecting the motifs with the length decreasing), and the alignment weight (starting from the alignment with the greatest weight and then selecting the motifs with the alignment weight decreasing). The masking step could be 0, 1, or 2, where 0 corresponded to the motif discovery on the initial sequence set (thus all motifs identified from the initial data went earlier in the list), then followed all the motifs obtained in the second round of motif discovery (the masking step of 1 with all the motifs found in the first run masked with poly-Ns), and finally, the motifs obtained in the third round of motif discovery (the masking step of 2, the hits obtained in motif discovery after masking motifs found in two initial rounds). With the sorted initial list of motifs at hand we then produced the filtered list of motifs using a simple greedy approach. The top motif from the sorted list was picked and all similar motifs (having the similarity value above a threshold) were removed from the list. Then the second best motifs were selected, and this procedure was repeated until the end of the list was reached. Similarity values were computed as in Ref[39] by the MACRO-APE[40] software (http://autosome.ru/macroape/) at a fixed motif *P* value of 0.0005 and the filtering motif similarity threshold of 0.15, which resulted in the final set of 1019 non-redundant motifs.

Discovery of cell type- or tissue-specific sequence motifs using Dragon Motif Finder (DMF)
Cluster selection: Sample-specific regulatory motifs were discovered for the 184,827 robust human CAGE clusters (described in a previous section) and the 889 human CAGE libraries for which RLE-normalised expression values were available (procedure also described in a previous section). Sample-specific CAGE clusters (SSCs) were defined as having greater than or equal to 10.0 RLE-TPM and at least ten-fold higher than the median expression of this cluster across all available CAGE libraries. This approach resulted in an average of 1,411 SSCs per library. For each human CAGE library a set of matching background CAGE clusters was determined by selecting all robust CAGE clusters that were not selected as SSCs for the particular CAGE library. The genomic sequences for these SSCs and background CAGE clusters were extracted after adding 300nt upstream and 100nt downstream of each cluster. The sets of SSC sequences cover on average 614knt.

Motif discovery: An OpenMP parallelised version of the Dragon Motif Finder (DMF)[41,42]was used to identify *ab-initio* motifs in each set of SSC sequences, using the corresponding background sequences to determine statistical significance of the motifs. For each set of SSCs we determine 1,600 raw motifs of variable length. DMF is parameterised in such a way that significant core motifs or length 5, 6, 7 and 8 are determined and subsequently extended upstream and downstream by a maximum of 10nt until a considerable drop in the motifs information content is detected. This results in a maximum possible motif length of 28nt. As a matter of fact this length is almost never achieved. For speed-up the algorithm is also set to operate at 95% accuracy within a 95% confidence interval using a 50% proportion sample.

Motif filtering: For each set of SSCs the 1,600 raw motifs are post-processed in the following way. Firstly all motifs were removed that did not appear in at least 5% of SSC sequences, that did not appear in 10% more SSC sequences compared to their appearance in background sequences, or that had an uncorrected *P* value greater than 0.05 (the *P* value for each motif was calculated using a student's t-test for independent variables, given an imaginary 2x2 contingency table with the values: number of SSC sequences that show at least one occurrence of the motif, number of SSC sequences that show no occurrence of the motif, number of background sequences that show at least one occurrence of the motif, number of background sequences that show no occurrence of the motif). Secondly, redundancy was reduced by eliminating similar motifs in the remaining set of motifs. For this purpose pair-wise Pearson correlation coefficients were calculated between all motifs. This was done by sliding the shorter motifs along the longer to account for partial overlaps. A minimum overlap of 5nt or 50% of the length of the longer motif was required. The maximum correlation coefficient out of all possible mutual positions of the two motifs is selected. All motifs were removed that had a correlation coefficient greater than 0.75 with a motif that had a smaller associated *P* value. Examples of sample specific motifs that were extracted in this manner can be seen in **Extended Data Fig. 5b**, sample specific motifs extracted with DMF for all human libraries can be found online here http://cbrc.kaust.edu.sa/ft5motifs/. For inclusion in the downstream motif analysis, a second round of *ab-initio* motif discovery was performed on the mouse and human promoteromes to detect general sequence motifs. Here, genomic sequences (300nt upstream and 100nt downstream) around all 184,827 robust human CAGE and 116,277 robust mouse CAGE clusters were extracted. A copy of these sequences in which the nucleotide order was randomly shuffled

was used as a set of background sequences. 1,600 raw motifs of variable length were determined in the same manner described above under *Motif discovery* with the exception that for this run the algorithm accuracy was set to 100%. Subsequently the 1600 raw motifs were filtered as described above under *Motif filtering*, with the exception that motifs were removed that had a Pearson correlation coefficient of greater than or equal to 0.9 with a more significant motif. After this step 848 human and 837 mouse *ab-initio* motifs remained that were integrated into the downstream analysis..

Discovery of cell type- or tissue-specific sequence motifs using HOMER
Cluster selection: Sample-specific clusters were determined from 184,827 robust human or 116,277 robust mouse CAGE clusters (described in a previous section) and the 889 human and the 389 mouse CAGE libraries for which RLE-normalised expression values were available (procedure also described in a previous section). Sample-specific CAGE clusters (SSCs) were defined as having greater than or equal to 2.5 RLE-TPM and being at least eight-fold higher than the sample bias-corrected median across all libraries. The latter was determined by hierarchically clustering RLE-normalized CAGE cluster tag counts of all samples, choosing a tree cut-off that resulted in 31 human and 47 mouse clusters (representing samples with similar expression profiles), averaging tag counts across each cluster of samples showing similar expression profiles and finally calculating the median across averaged cluster tag counts. For *de novo* motif discovery, genomic coordinates of each SSC were extended by adding 300bp upstream and 50bp downstream. Extended SSC that overlapped were merged before applying motif discovery. This approach resulted in an average of 2146 merged SSCs per human and 2158 merged SSCs per mouse library.

Motif discovery: Motif enrichment was analysed using HOMER[43] version 3, (a suite of tools for motif discovery and next-generation sequencing analysis (http://biowhat.ucsd.edu/homer/). Sequences of extended and merged SSCs were compared to ~50,000 randomly selected genomic fragments of the average SSC size, matched for GC content and auto normalized to remove bias from lower-order oligo sequences. After masking repeats in SSC and background regions, motif enrichment was calculated using the cumulative binomial distribution by considering the total number of target and background sequence regions containing at least one instance of the motif. With HOMER, *de novo* motif discovery is divided into two phases starting with a global, exhaustive scan of all oligos for their enrichment, followed by a second local optimization of motif probability matrices using best oligos from the first phase as the initial seeds for the optimization. As motifs are discovered their instances are masked from the input sequence to avoid convergence of multiple motifs on the same highly enriched sequence elements. Twenty-five motifs were searched for a range of motif lengths (7-14 bp) resulting in a set of 200 *de novo* motifs per sample.

Motif filtering: To create non-redundant motif collections for SSCs, each set of sample-specific *de novo* motifs was ranked and reduced as follows. All motifs were removed that had an uncorrected enrichment *P* value above 10^{-18}, did not appear in at least 50 SSCs, and that had a limited information content (< 1.5). Motifs were then checked for redundancy by aligning each pair of motifs at each position (and their reverse opposites) and scoring their similarity to determine their best alignment (matrices are compared using Pearson's correlation coefficient by converting each matrix into a vector of values; neutral frequencies (0.25) are used in positions

where the motif matrices do not overlap). All motifs were removed that had a correlation coefficient greater than 0.75 with a motif that had a smaller associated *P* value.

Discovery of cell type- or tissue-specific sequence motifs and modules using ScanAll
Cluster selection: Sample-specific clusters were determined from 184,827 robust human or 116,277 robust mouse CAGE clusters (described in a previous section) and the 889 human and the 389 mouse CAGE libraries for which RLE-normalised expression values were available (procedure also described in a previous section). Sample-specific CAGE clusters (SSCs) were defined as having greater than or equal to 10.0 RLE-TPM and at least ten-fold higher than the median expression of this cluster across all available CAGE libraries. In order to use ScanAll for *ab-initio* motif discovery, the genomic sequences for these SSCs clusters were extracted after adding 300nt upstream and 100nt downstream of each cluster.

Motif discovery: ScanAll (*Dalla et al.* in preparation) aims at finding structured substrings common to a significant portion of the sequences in the input set, allowing a fixed layout for mismatches in the input itself. The general strategy is based on the introduction of a data structure encoding 'a la Karp-Rabin' substrings of the strings in the SSCs. ScanAll started by outputting all the positions of the common subsequences of length ℓ=6 and with d=1 variations, with the addition of the constraint for the variations, if occurring, to be in the same location and never occurring in the first position of the element. During this phase we identified 4198 unique conserved elements with the required layout.

Module selection: ScanAll then encoded and manipulated these motifs, introducing a distance constraint to identify groups of motifs located within a given range (<40,90> minimum-maximum nucleotide distance). This allowed on the one hand to find higher levels structures corresponding to putative regulatory modules, while on the other hand to reduce the size of the motifs discovered, since only the module-composing motifs were kept. The background model was built maintaining the sequence-specific dinucleotide composition of each sequence related to every FANTOM5 sample. Two shuffled background datasets were generated and were analyzed using ScanAll, as described above.

Motif filtering: Sequential filtering steps were applied to each sample, during motif and module discovery, in order to obtain a significant and non-redundant regulatory elements list. First of all we introduced two thresholds (that we called "*quorum*") to the motif- and module-discovery phases. Newly-discovered motifs, with the aforementioned layout, were only retained if present in at least 150 different sequences of SSCs. Subsequently, motif-derived modules were preserved only if conserved in at least 60 different promoters. During the module-building phase we introduced another parameter, that we called "*complexity*", that corresponds to the number of different nucleotides required to appear in every motif, and we fixed this value to 4 (that is, for motif "AACnG" the only acceptable solution is "AACTG") in order to prevent ScanAll from taking into account any low complexity, highly-conserved genomic element. Afterwards, overlapping motifs were merged into consensus sequences ending with the best layouts selection and the generation of a non-redundant list of modules. Finally, Z-scores and their associated *P* values ($p<0.05$) were calculated with a continuity correction[44] comparing the results obtained for each sample to the relative background sequences. In the end, 1370 human and 1277 mouse non-redundant motifs were obtained.

8. Clustering and assessment of novel motifs

Overview of the procedure

We first clustered the known motifs together with motifs discovered *de novo* in the vicinity of CAGE clusters to estimate the optimal threshold for cutting a hierarchical clustering tree of motifs. Next, we removed all *de novo* motifs similar to the already known motifs to arrive at the set of novel motifs, which were then clustered using the previously selected threshold. For each cluster, one representative motif was selected, thus forming the non-redundant set of novel motifs. These non-redundant novel motifs were assessed for the statistical significance of their correlation with promoter expression across samples.

1. TFBS motif sets.

We used the following collections as sources of known motifs: HOCOMOCO[39] integrated TFBS models (426 motifs); HOMER known motifs, based on ChIP-Seq analysis (138 motifs); JASPAR[45] core vertebrate (130 motifs); SwissRegulon[46] collection (190 motifs); UniPROBE[47] collection of matrices for mouse and human TFs (413 motifs); and the human regulatory LEXICON[48] collection obtained from DNase I hypersensitive footprints (683 motifs). To remove known motifs from the *de novo* motifs, we also filtered against the human and mouse motifs in TRANSFAC release 2012.2[49].

The *de novo* motif resultsfor each method were then combined(human/mouse respectively): ChIPMunk[38] (1619 / 630 motifs); DMF[42] (848 / 837 motifs); HOMER[43] (1426 / 692 motifs); ScanAll(1370 / 1277 motifs). The combined set consisted of 10679 motifs. Position count matrices from each collection were transformed to weight matrices using the log-odds transformation with a pseudocount of 0.5.

2. Constructing the hierarchical tree

We used the UPGMA[50] approach to produce a linkage tree using pair wise similarities calculated for all motif pairs following the strategy described in[39]. The MACRO-APE[40] software (http://autosome.ru/macroape/) was used to obtain the similarity value for all pairs of motif models (the model being a combination of a positional weight matrix and a score threshold). MACRO-APE computes a variant of Jaccard measure for two motif models: the similarity is defined as the number of words recognized as TFBSs by the both models, divided by the number of words recognized by any of them. The threshold levels of motif models were selected corresponding to the *P* value level of 0.0005 referred to the random sequence with uniform nucleotide composition (i.e. 5 out of 10000 random words are recognized as positive hits; this approximately corresponds to 1 PWM hit per 1000bp of a random double-strand DNA sequence).

2a. Producing clusters based on the hierarchical tree

To produce clusters from the hierarchical tree the branches were cut at the level corresponding to the given threshold for the link length. Each cluster corresponded to a particular branch.

2b. Estimating the link length threshold for clustering

To estimate the link length threshold we clustered all *de novo* and all known motifs together and plotted the number of clusters that contained known as well as *de novo* motifs, or “the annotated

clusters", versus the link length threshold value. The curve reached a clear extreme at a link length threshold of around 0.95. The maximum of 218 annotated clusters corresponded to the link length threshold equal to 0.9586. The same number of annotated clusters was observed for two close link length threshold values. We selected the higher link length threshold value and thus the lesser overall number of clusters.

3. Using TomTom to identify novel motifs among the de novo motifs

To assess the similarity of the 8699 *de novo* motifs to known motifs, we ran the TomTom[51] motif comparison software for each of the *de novo* motifs against the HOCOMOCO, HOMER, JASPAR, SwissRegulon, UniPROBE, TRANSFAC, and ENCODE Lexicon databases separately. A *de novo* motif was considered similar to a known motif if the E-value as calculated by TomTom was less than 0.5, corresponding to less than 1 hit being expected at random. A known motif was found for 7478 of the 8699 *de novo* derived motifs, while the remaining 1221 *de novo* motifs were deemed novel.

For the purpose of evaluating the coverage of databases of known motifs, we ran TomTom for each of the known motifs against the combined set of *de novo* motifs. However, simply merging the *de novo* motifs generated by the four *de novo* motif finding methods would give rise to a certain degree of redundancy among motifs in the merged set. This would disproportionately inflate the E-value as reported by TomTom, as it depends on the size of the database against which it is run. We therefore first ran TomTom for each known motif against all *de novo* motifs and identified the *de novo* motif that best matches the known motif according to the *P* value reported by TomTom. In total, we found 1105 *de novo* motifs that were the best hit for at least one known motif. We then created a database of these 1105 best matching *de novo* motifs and ran TomTom for each known motif against it, applying a threshold of 0.5 on the E-value. This revealed that the *de novo* motifs cover the vast majority of motifs in the known motif databases (**Extended Data Fig. 5c**).

4. Creating a non-redundant motif set by UPGMA clustering

We used Biopython[52] to calculate the position-weight matrix scores for each of the 1221 novel motifs in the -300..+100 base pair region around the representative position of each of the robust promoters in both human and mouse. Here, the representative position of a promoter is defined as the position within the promoter that has the highest total number of CAGE tags across all samples. Using a prior probability Pr(T) equal to 5×10^{-4}, we calculated the posterior probability $\Pr(\mathrm{T}|S_\mathrm{F}, S_\mathrm{R})$ of a predicted TFBS as

$$\Pr(\mathrm{T}|S_\mathrm{F}, S_\mathrm{R}) = \frac{\Pr(\mathrm{T})\left(\frac{1}{2}\exp(S_F) + \frac{1}{2}\exp(S_R)\right)}{\Pr(\mathrm{T})\left(\frac{1}{2}\exp(S_F) + \frac{1}{2}\exp(S_R)\right) + 1 - \Pr(\mathrm{T})},$$

where S_F and S_R are the position-weight matrix score on the forward and reverse strand, respectively. Retaining all predicted TFBSs with a posterior probability larger than 0.25, for each motif separately, we averaged over the robust promoters the posterior probabilities of TFBSs predicted at a distance d with respect to the representative position of the promoter to arrive at the probability $\Pr(\mathrm{T}|d)$ of detecting a TFBS as a function of the distance d. The profile of a motif along the -300..+100 base pair search region R is then calculated as

$$f(d) \equiv |R| \cdot \Pr(d|\mathrm{T}) = |R| \cdot \frac{\Pr(\mathrm{T}|d)\Pr(d)}{\sum_{d' \in R} \Pr(\mathrm{T}|d')\Pr(d')} = \frac{\Pr(\mathrm{T}|d)}{\frac{1}{|R|}\sum_{d' \in R} \Pr(\mathrm{T}|d')} = \frac{\Pr(\mathrm{T}|d)}{\langle \Pr(\mathrm{T}|d') \rangle_{d' \in R}},$$

$|R|$ = 401 base pairs being the size of the search region R, as a priori Pr(d) is independent of d. For each predicted TFBS, the posterior probability of predicting a TFBS with position-weight matrix scores S_F and S_R at a position d with respect to the promoter is

$$\Pr(\mathrm{T}|S_F, S_R, d) = \frac{\Pr(\mathrm{T}|S_F, S_R) f(d)}{\Pr(\mathrm{T}|S_F, S_R) f(d) + 1 - \Pr(\mathrm{T}|S_F, S_R)}.$$

For each promoter and each motif, we summed over the search region R the posterior probabilities Pr(T|S_F, S_R, d) exceeding the threshold of 0.25 to arrive at the predicted number of binding sites for each motif at each promoter.

We then clustered the 1221 novel motifs using MACRO-APE[40], and cut the tree at the previously determined link length threshold selected, arriving at 172 clusters of novel motifs. We calculated the Pearson correlation between the CAGE expression of the human robust promoter set in each sample and their associated TFBSs for the 1221 novel motifs. For each of the 172 clusters, we selected the motif with the largest squared correlation summed across samples as the representative motif. We discarded 3 clusters for which the motifs were too weak to generate TFBSs at any of human robust promoters at the 0.25 threshold on the posterior probability. We thus arrive at a non-redundant set of 169 novel motifs.

5. Evaluating the non-redundant novel motifs for significance.

To assess the statistical significance of the association of motifs with expression in particular samples, for each novel motif we randomized the order of the positions of the position-weight matrix, predicted TFBSs at each promoter as described above for each randomized matrix, and calculated the correlation between promoter expression and associated TFBSs. Using 1000 randomizations for each of the 169 novel motifs, we calculated the mean and standard deviations of these correlations, and expressed the correlation found for the novel motif as a Z-score with respect to this mean and standard deviation. We then calculated a P value for each novel motif in each sample as the two-sided tail probability corresponding to this Z-score under the normal distribution. We apply the Bonferroni correction for multiple testing by multiplying the P value by the number of samples. Requiring a significance level of 0.05 on the corrected P value in either human or mouse yielded 37 significant novel motifs shown in **Supplementary Table 12** (sequence logos generated by WebLogo[53]), together with the samples in which they were found significant. The frequency matrices of these novel motifs are available online at http://fantom.gsc.riken.jp/5/data/.

6. Evaluating the significance of the binding profiles of the non-redundant novel motifs

For each of the significant novel motifs, we calculated the Kolmogorov distance between the calculated binding profile $f(d)$ and a uniform binding profile, and expressed this distance as a Z-score with respect to the Kolmogorov distances calculated for the 1000 randomized motifs. The corresponding P value was then calculated as the one-sided tail probability for this Z-score under the normal distribution. These P values, calculated separately for human and mouse, are shown in **Supplementary Table 12**. As comparison, we ran the same test on the JASPAR collection of motifs. For 112 we can calculate the significance (the remaining 18 motifs, the information content was too low to predict TFBSs anywhere with a position-weight matrix model). Of these 112 motifs, 81 were significant in human or mouse (72 were significant in human, 74 were significant in mouse; of these with 65 were significant in both human and mouse (P value = 3.6e-13, Fisher's exact test)).

7. Evaluating the overrepresentation of the novel motifs in co-expression clusters
For each novel motif, we calculated the number of TFBS predictions in promoters included each of the co-expression clusters by summing the posterior probabilities of predicted TFBSs. We also calculated the expected number of TFBS predictions by averaging the sum of the posterior probabilities over all promoters. For each of the co-expression clusters, for each motif we multiplied this average by the number of promoters in the co-expression cluster to arrive at the expected number of TFBS predictions under the null hypothesis. We then calculated the tail probability of achieving at least the observed number of TFBS predictions under the Poisson distribution with a mean equal to the expected number of TFBS predictions under the null hypothesis. For each of the 37 novel motifs, the most significant cluster in human and in mouse, together with the corresponding *P* values, are shown in **Supplementary Table 12**.

8. Genomic Regions Enrichment of Annotations Tool (GREAT) analysis
For each of the 169 novel motifs, we considered the -300 to +100 base pair genomic region with respect to each of the human robust peaks, and assigned it to the foreground set if any predicted TFBSs for the motif were associated with the peak, and to the background set otherwise. We then ran GREAT[54] to discover Biological Process gene ontology terms that are enriched in the foreground set compared to the background set, and then performed the same analysis independently for mouse. For each of the 169 novel motifs, we calculated the overlap between the top-500 gene ontology terms found in human and the top-500 gene ontology terms found in mouse. To evaluate the statistical significance of this overlap, for each of the 169 novel motifs we also calculated the overlap between the top-500 gene ontology terms found in human and the top-500 gene ontology terms found for a different novel motif in mouse. Using the set of values for the overlap between non-matching motifs in human and mouse as the background distribution, we then calculated the statistical significance of the overlap for a novel motif as the tail probability of finding at least the same overlap in the background distribution (Fig. A below).

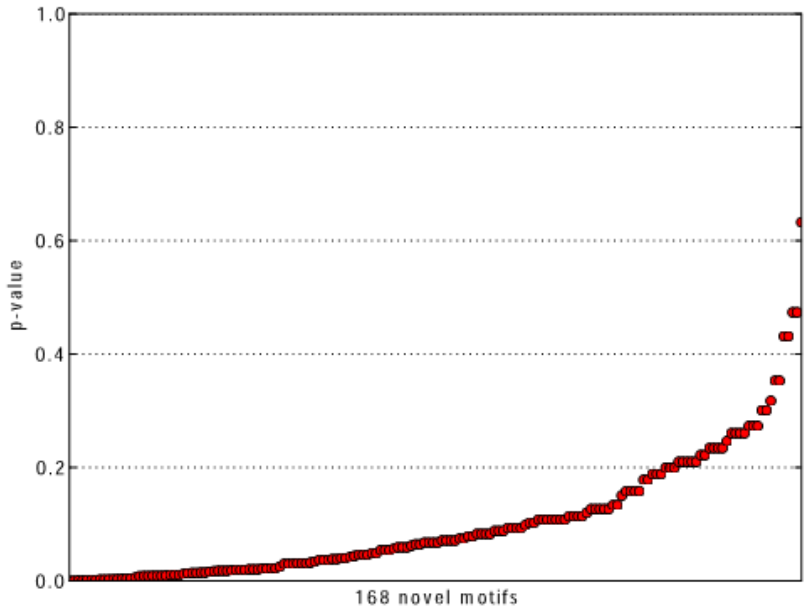


Figure S8: Significance of overlap in GREAT enrichment results for human and mouse on the same motif. For

each of the 169 novel motifs, we applied the Genomic Enrichment of Annotations Tool GREAT[54] to identify, both in human and in mouse, the gene ontology terms of biological processes enriched given the predicted TFBSs, and evaluated the overlap in the top-500 gene ontology terms between human and mouse. For each novel motif, the *P* value for the overlap was then evaluated by calculating its relative rank with respect to this background distribution. This Fig. shows the *P* values thus obtained, sorted by significance. As one of the novel motifs did not yield predicted TFBSs anywhere in the mouse genome at the thresholds we employed, its *P*-value could not be calculated and is therefore not shown in this Fig.

9. MCL clustering of sample and CAGE promoter co-expression graphs

The analysis of correlation networks has been used extensively to explore these data.

Promoter correlations: In main **Fig. 6** three samples of pooled RNA were excluded from the co-expression clustering since they were not expected to contribute useful information. A Pearson correlation matrix was constructed consisting of pair wise comparisons of expression across the remaining 886 tissues, primary cell types and cell lines. Correlations with r < 0.75, corresponding to the 99.7th percentile, were ignored.

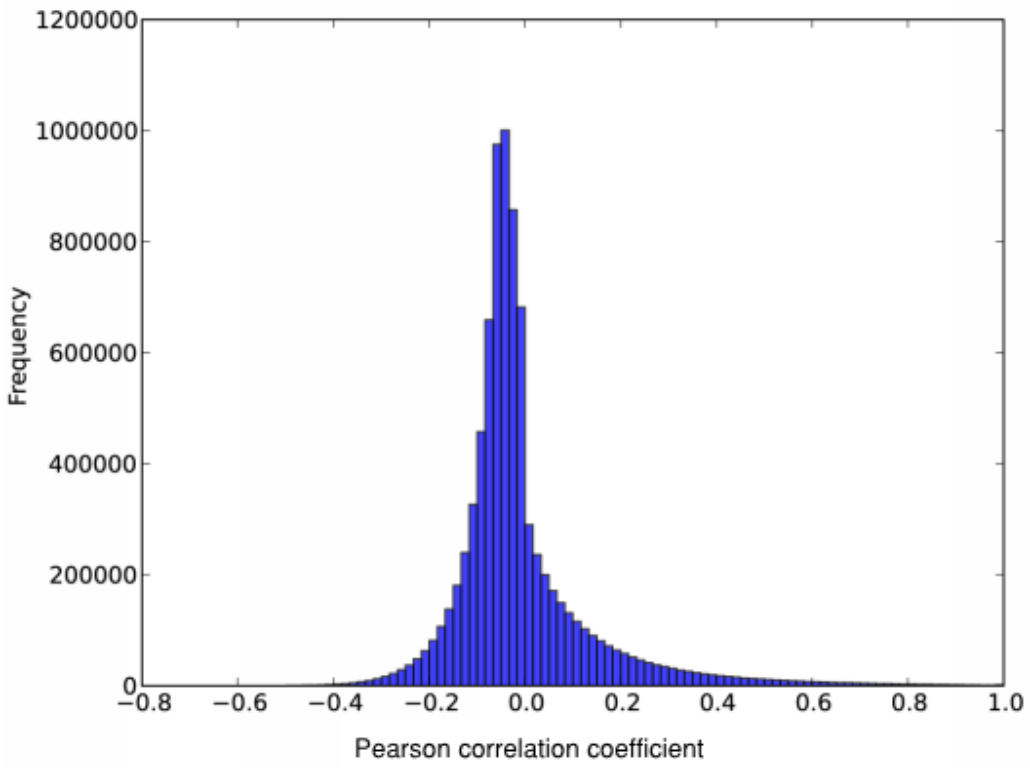


Figure S9: Distribution of Pearson correlation coefficients in the FANTOM5 correlation matrix for Human Robust clusters.

In order to more accurately reflect the biological implications of a strong correlation compared to a weak one, the dataset was then transformed by subtracting 0.75 from each correlation coefficient in the matrix.

Clustering was performed using the MCL algorithm[55], with an inflation value (MCLi) of 2.2 and pre-inflation set at 3.0. The MCL algorithm simulates flow through the network of Pearson correlations, prioritising edges conducting more flow until a stable arrangement of discrete clusters is obtained[55]. It is highly effective for clustering gene expression data[56], and protein interaction networks[57], and is strikingly robust to network perturbations, comparing favourably

to alternative methods[58].
One of the advantages of the MCL algorithm is that clustering is highly data-driven with minimal user input required. One key parameter is set by the user at the outset. The MCL inflation value (MCLi) range from 1 to 30, and determines the granularity of the clustering. A low inflation value results in a small number of large, inclusive clusters. A high value creates a large number of small clusters, with more nodes that do not belong to any cluster. The total number of nodes in the network is almost always lower than the number of entities in the original matrix. This is because many nodes do not form sufficiently strong correlations anywhere in the network, and are discarded since an unconnected node does not add information to the network.
In order to explore the entire network (120,090 promoters), entire clusters were collapsed into a single node and displayed using BioLayout Express3D[59], with node size proportional to the cube root of the number of promoters in each cluster. Edges indicate the Pearson correlation coefficient between the average expressions of each pair of clusters across the entire dataset. Clusters were automatically numbered in consecutive order of size, with the largest cluster designated Cluster 0 (C0), and named with a key word from each of the six samples in which expression of the cluster was greatest (most abundant first).
Sample correlations: Shown in **Extended Data Fig. 12** is a sample-to-sample correlation graph for the human promoterome data. In this instance each node represents an individual sample and edges Pearson correlations (r) > 0.75 between them. A sample-to-sample correlation matrix was calculated using the 'mycor' function within Bioconductor and the resultant graph displayed using BioLayout *Express*[3D]. In **Extended Data Fig. 12a** nodes are coloured according to whether they represent tissues (red), primary cells (cream) or cell lines (grey). It can be seen that all cell line data tends to be in one area of the graph indicating an overall similarity in the data irrespective of the type of cell (cancer) from which the cells are derived. In contrast the tissue and primary cell data is more widely distributed. This is better illustrated in **Extended Data Fig. 12b** where the graph is coloured according to MCLi cluster 2.2. This shows many clusters to be highly enriched in samples derived from related tissues or cell types.
TF Promoter correlations: **Extended Data Fig. 8a** shows a correlation graph constructed from data derived from all promoters associated with human transcription factors expressed in the primary cell samples. A Pearson correlation matrix was calculated so as to compare the profile of expression of each TF promoter and a graph constructed where r > 0.70 and clustered using an MCLi value of 2.2. In this way, groups of promoters showing a similar and often cell-specific expression profile could be identified. In all cases examined the cell-specific clusters recapitulated known the transcription factors associated with a particular cell type, however clusters frequently contained putative new associations between transcription factors and cell lineage specification.

10. Accession numbers

CAGE tag sequences from this study have been deposited at DDBJ DRA under accession number DRA000991.

11. CpG and nonCpG associated CAGE clusters

Expression specificity at CpG island (CGI) versus nonCGI-associated CAGE clusters
The set of 184,827 robust human CAGE clusters was separated into 61,320 CGI and 123,495 nonCGI-associated clusters, uisng the UCSC CpG Island track and then further separated into bins based on expression specificity (log ratio of expression of each cluster in a given sample

versus its pooled expression in the 889 human CAGE libraries for which RLE-normalised expression values were available).

Distribution of general promoter features as a function of expression specificity and CGI-association
We obtained ChIP-Seq data for H3K4me3 and H2A.Z histone modifications as well as DNase-seq data produced by ENCODE[60] as listed in **Supplementary Table 14**. Sequence tags were mapped to the current human reference sequence (GRCh37/hg19) using Bowtie[61]. Average distribution of extended ChIP-Seq reads and DNase I hypersensitivity sites was visualized centered on dominant (most frequently used) TSS for CGI and non-CGI clusters separately across defined expression specificity bins using custom scripts in R.

Distribution of transcription factor binding and motifs as a function of expression specificity and CGI-association
We obtained published ChIP-Seq data for several transcription factors, Pol2, and P300[60,62] as listed in **Supplementary Table 14**. Sequence tags were mapped to the current human reference sequence (GRCh37/hg19) using Bowtie[61] and only uniquely mapped tags were used for downstream analyzes. Histograms showing the average distribution of mapped ChIP-Seq reads around binned CAGE clusters were generated using HOMER[43]. Corresponding transcription factor motifs were also generated and mapped using HOMER.

12. Fantom3, 4, 5 and ENCODE CAGE comparison
We compared CAGE datasets from four project on the basis of their coverage of known genes defined in RefGene. This comparison is challenging due to the different CAGE protocols[12,17,63], processing pipelines and samples used over a period of 7-8 years. We attempted to lift over the coordinates of past CAGE clusters to hg19 but found that many are lost in translation. Hence we re-mapped all old CAGE datasets to hg19 using BWA[64] with default parameters. For each sample we computed the expression (TPM normalized) of all RefGenes based on all reads in the vicinity (+/-500bp) of their annotated 5' start site. Finally we plotted the distribution of maximum expression for all genes separated by project (**Figure S7e,f**). Reassuringly, both the coverage of genes and the distribution of maximum expression in each sample was very comparable between the projects.

13. Mouse and human projections
Human TSS were projected into the mouse reference genome and mouse into human through the Ensembl EPO12 eutherian mammal multiple sequence alignments[65] using the Ensembl API (database and code release 67). The EPO12 alignments incorporate the GRCh37 (hg19) reference human genome assembly and the NCBI37 (mm9) reference mouse assembly. Using a single reference multiple sequence alignment ensured circular consistency of projections (human to mouse to human will find the original human locus) and allowed projection into other sequenced mammalian genomes for comparison and to provide an out-group to assign a branch and direction to evolutionary changes.

Human (or mouse) TSS boundaries were used to define slices of the EPO12 alignments using the alignSlice function to resolve overlapping alignment blocks and to orient and order alignment blocks relative to the human (or mouse) genome. The TSS reference coordinate (modal tag

position) was projected though the alignment slice to obtain a projected reference position. In cases where the projected position falls in an alignment gap, the reference is projected onto the nucleotide at the closest edge of the gap but still within the alignment slice. In cases where the alignment slice is entirely gap but there is flanking aligning sequence on both sides of the alignment slice this is recorded as a "gap" alignment which indicates the deletion or insertion of the TSS sequence during genome evolution. In cases where the alignment slice cannot be projected into a genome at all, i.e. there is no syntenic interval that aligns across the slice, this is recorded as "unaligned". In this case we don't have evidence to discriminate the evolutionary gain or loss of sequence from technical difficulties such as alignment or genome assembly problems or the absence of read coverage in the raw genomic sequence. After projecting reference coordinates, the outer margins of TSS intervals were mapped into the aligned sequences, requiring that they map into the same chromosomal locus as the projected reference position, +/-80 nucleotides. In cases of genomic rearrangements between species the projected interval was trimmed down to the boundary of the rearrangement.

Having projected human TSS into the mouse genome we asked if the projected TSS overlapped an observed TSS in mouse. The reciprocal was also done for mouse TSS projected into human. 119,979 human TSS overlapped with 105,378 mouse TSS when projected through genomic alignments. This discrepancy in numbers is principally due to multiple human TSS projecting into a single broader mouse TSS (**Extended Data Fig. 4**), an expected consequence of the DPI clustering approach being applied to more human than mouse libraries. To directly compare the annotation of human and mouse TSS we defined projected super-clusters where TSS and cross-species projected TSS within 20 nucleotides of each other and on the same strand in either in the human or the mouse genome. The resulting 119,216 projected super-clusters comprise 168,206 human and 156,612 mouse DPI defined TSS. Note that these numbers are greater than the projected overlaps as DPI TSS are clearly clustered and accordingly a 20 nucleotide proximity constraint was included the projected super-clustering. Mouse annotation of the projected super-cluster took the most functional annotation (e.g. protein-coding > known-transcript > Unannotated) of the contributing mouse TSS and the human annotation was similarly the most functional of any contributing human TSS.

14. Pathway enrichment analysis

Canonical pathways

Canonical pathway gene sets were compiled from Reactome[66], Wikipathways[67] and KEGG[68]. For the major signaling pathways, the transcriptionally-regulated genes (downstream targets) were obtained from Netpath[69].

Combined, the canonical pathways and downstream targets totaled 489 human gene sets. The corresponding *M. musculus* gene sets were inferred by homology using the HomoloGene database[70]. The gene membership of these sets is described in **Supplementary Table 15**.

Co-expression cluster pathway enrichment analysis

Enrichment for each of the 489 pathways and gene sets described above was performed for each co-expression cluster. Given that each co-expression cluster has n genes, a pathway or gene set of interest has m genes, and assuming there are a total of N genes on the genome, the probability of having an overlap X of exactly k genes between a co-expression cluster and a pathway of

interest by random chance is given by the hypergeometric probability:

$$P(X=k)=\frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}$$

An enrichment *P* value, i.e. the probability of obtaining the observed overlap result or more extreme, was obtained by summing these probabilities from *k* to *n*. Since enrichment for each co-expression cluster was performed 489 times for each separate pathway, the *P* values were also then adjusted by the Benjamini-Hochberg method for multiple comparisons[33] (ee take N = 19044). All analyses were performed using R[71] version 2.15.0. Entrez gene IDs were used exclusively for enrichment analysis. Gene identifier mapping was performed wherever necessary using the *org.Hs.eg.db*[72] and *org.Mm.eg.db*[72] BioConductor R packages for human and mouse analyses respectively. Significant pathway enrichments found are summarized in **Supplementary Table 16**.

15. Comparison of peaks to H3K4me3, H3K9ac, H3K27ac and RNA-seq from ENCODE

First exons of transcript models based on GENCODE (version 14) annotation and de novo sample specific transcript models were identified for each of the GM12878, HepG2 and K562 cell lines. Both terminal exons were considered for those transcript models where strand assignment was not defined. ENCODE alignments of deep RNA-seq from whole cell poly-A plus and poly-A minus were obtained from the UCSC ENCODE browser (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/ wgEncodeCshlLongRnaSeq/) and from these, expression levels calculated for the identified first exons using coverageBed[73] with the "split" option to give exon specific reads per kilobase exon per million reads (RPKM). Exons were classed as expressed if RPKM > 0.

Histone modification data for H3K4me3, H3K9ac and H3K27ac was obtained from GEO accession GSE26386[74]. Reads were mapped to the hg19 reference genome using Bowtie[61] (version 0.12.8) retaining only unique matches. Peaks were called separately for both replicates of each dataset using MACS[75] (version 14). The union of peaks identified in both replicates was considered the peak interval for each histone modification. Histone peaks were considered to be RNA-seq supported if a transcript model terminal exon supported by RNA-seq reads in the relevant cell type was located within 50 nt of the histone peak interval. In the case of multiple exons being associated with a given peak, the highest associated RPMK was used. The H3K4me3 peaks with transcript model and RNA-seq support were considered candidate active promoters.

To provide a genomic null expectation for annotations and conservation, TSS locations were randomly shuffled over the entire reference genome but excluding the ENCODE DAC Blacklisted Regions. The subset of clusters being analysed were shuffled 100 times and median overlaps/annotations were reported to provide background rates.

Supplementary Notes

Supplementary Note 1: Access to the FANTOM5 results

In order to facilitate access to the FANTOM5 CAGE data set and its analysis results, we set up the following interfaces: (i) ZENBU[76], a genome browser which incorporates an expression summary linked to a genomic view, (ii) SSTAR (Semantic catalogue of, samples, transcription initiations, and regulations, Shimoji *et al.* in prep) which enables us to explore the profiled samples, identified CAGE peaks, and regulatory information (iii) BioMart[77] which enables us to export CAGE peak information via a widely used interface, (iv) a collection of (raw and processed) data files which enable us to download a bunch of data for subsequent analysis, (v) promoter slider to select CAGE peaks showing a specified expression pattern, and (vi) nanopublication for interoperable data exchange.

(i) ZENBU

The ZENBU genome browser system was developed for the FANTOM5 project to enable the interactive exploration of the deep FANTOM5 data sets. ZENBU can perform data processing manipulations to allow the same loaded data to be visualized in many different ways.

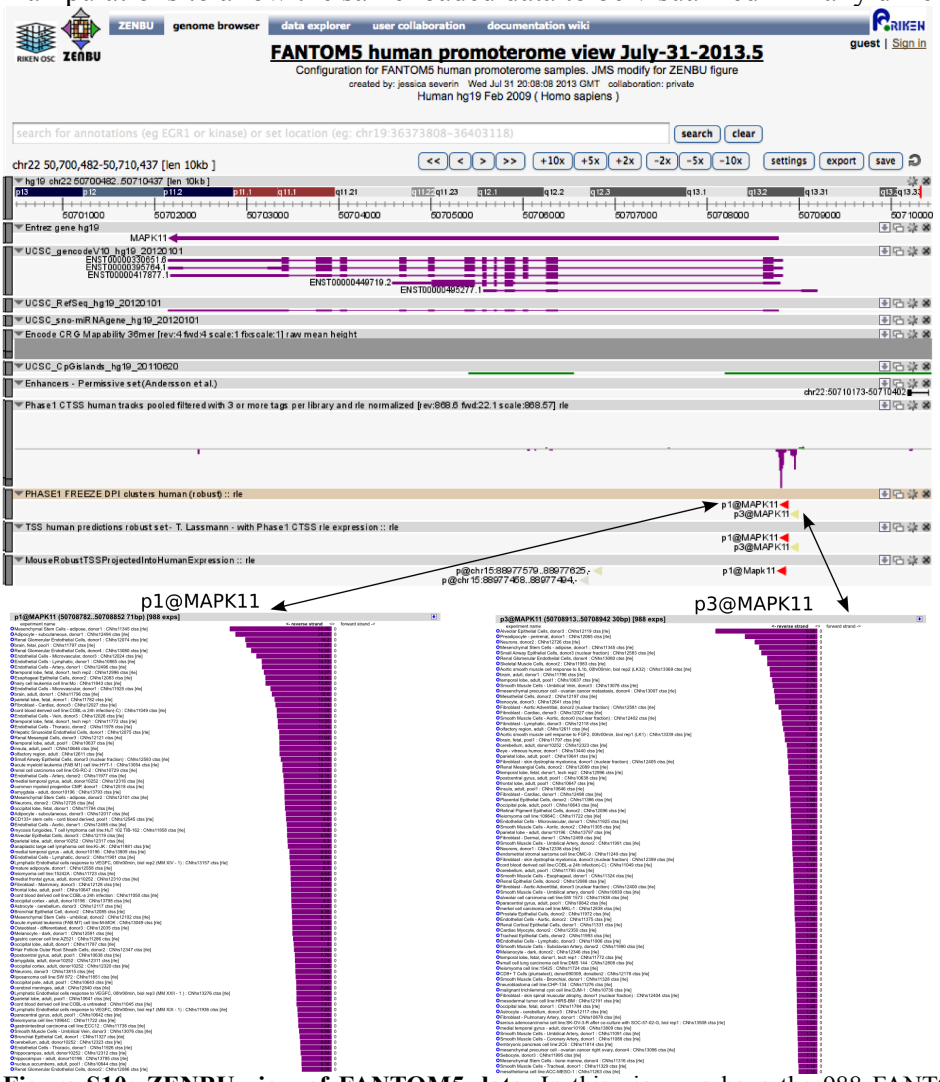


Figure S10: ZENBU view of FANTOM5 data. In this view we have the 988 FANTOM5 CAGE CTSS signal experiments dynamically merged into a single track and displayed along side gene annotation tracks and FANTOM5 promoter annotations for human and a liftover from mouse. The FANTOM5 CTSS expression is dynamically collated by ZENBU into the FANTOM5 promoters both for human (DPI robust, Lassmann classifier robust). On the gene MAPK11 (MAP kinase 11) we see two distinct promoter with tissue and cell specific expression difference. In ZENBU one can click or select objects in tracks and the linked expression facet view will update to show the underlying expression. We see on this gene that promoter p1@MAPK11 shows stronger expression in endothelial derived cells, while p3@MAPK11 shows stronger expression in smooth muscle, neuronal tissues, and epithelial derived cells.

(ii) SSTAR (Semantic catalogue of, samples, transcription initiations, and regulations)
SSTAR allows exploration and searches through the samples, transcriptional initiation regions, motifs and regulation events in the FANTOM5 collection.

Figure S11: Example of a sample entry in SSTAR. The first box contains the ranked list of top Transcription factors identified in FANTOM5.

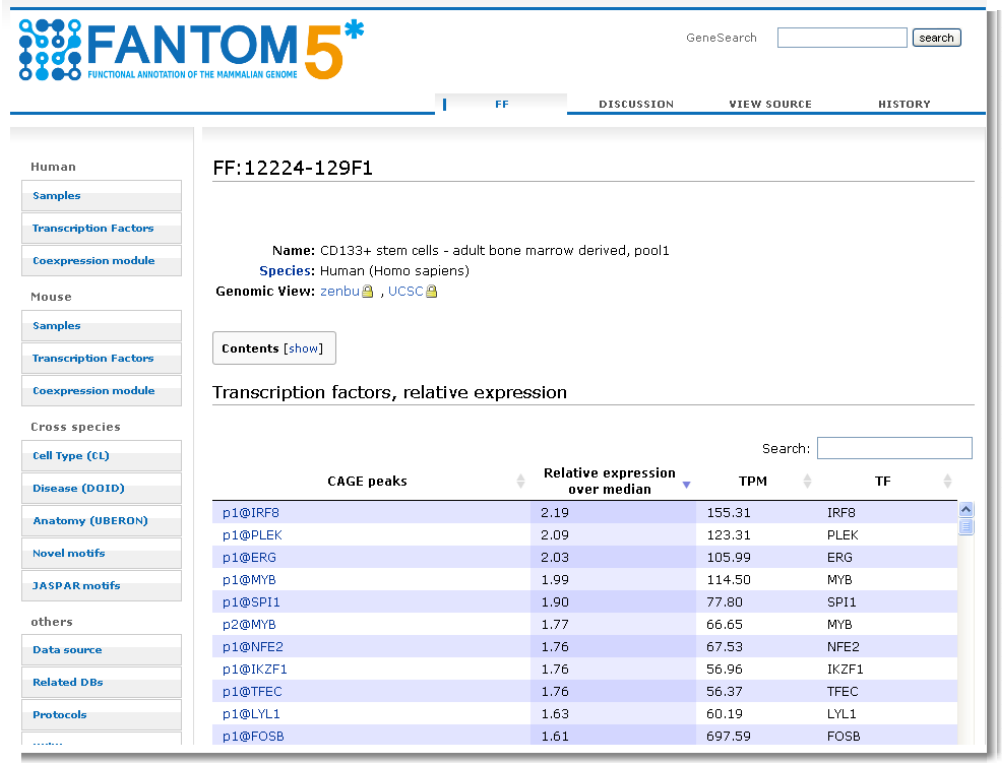


Figure S12: Sample specific motifs in SSTAR

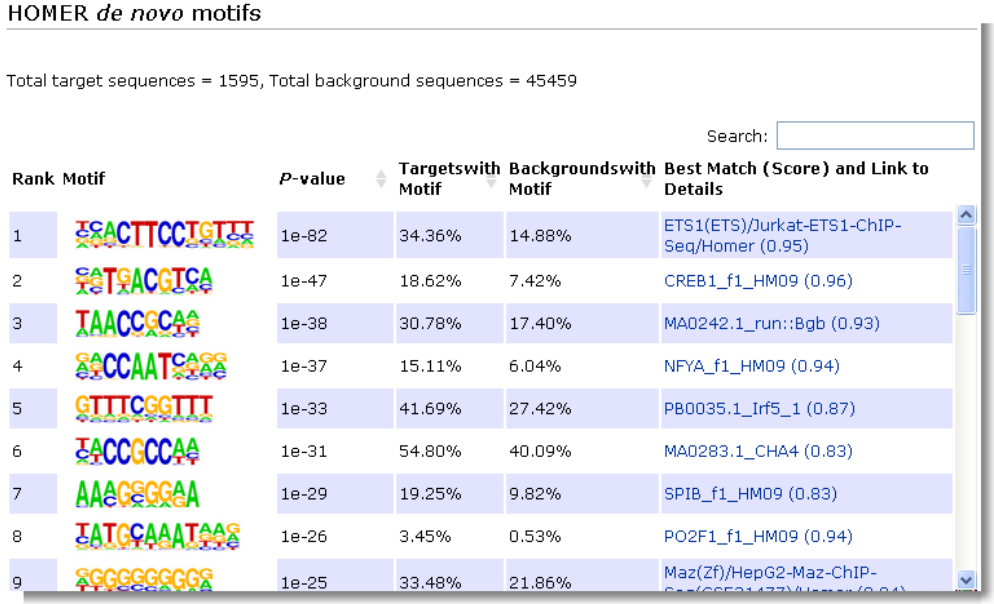


Figure S13: SSTAR summarises the co-expression groups that are specifically active in this sample

Co-expression modules, relative expression

Search:

Co-expression module	Relative expression over median
C4592-acute-myelodysplastic-CD34-CD133-Hodgkin-chronic-CD14CD16	2.38
C1829-acute-CD133-CD34-granulocyte-chronic-biphenotypic-myelodysplastic	1.937
C4563-non-Mast-chronic-acute-granulocyte-myelodysplastic-CD34	1.769
C3297-Monocytederived-CD14CD16-CD14-Macrophage-Dendritic-migratory-splenic	1.552
C3146-acute-CD133-CD34-chronic-biphenotypic-myelodysplastic-granulocyte	1.511
C1598-CD34-CD133-CD4-CD8-Smooth-Cardiac-acute	1.47
C3495-Mast-granulocyte-acute-CD34-immature-CD133-cord	1.433
C4697-Eosinophils-Mast-acute-non-CD14-Neutrophils-myelodysplastic	1.433
C2424-CD14-CD133-Dendritic-Natural-Basophils-chronic-Peripheral	1.391
C3675-Mast-Whole-blood-acute-NK-Neutrophils-CD34	1.384
C3158-acute-CD133-granulocyte-brain-cerebellum-diencephalon-CD34	1.321
C4422-Osteoblast-CD4-CD34-acute-CD133-Natural-CD8	1.321

Showing 1 to 4,883 of 4,883 entries

Figure S14: Detailed information stored in SSTAR on the sample and the sample ontology terms associated with it.

Details

Sample information

strain	
tissue	bone marrow
dev stage	
sex	
age	
cell type	stem cell
cell line	
company	Stem cell technologies
collaboration	FANTOM5 OSC CORE (contact: Al Forrest)

RNA information

lot number	
catalog number	
sample type	
extraction protocol (Details)	OP-RNA-extraction-totalRNA-miRNeasy_Mini-v1.0

Parent terms in the FF ontology

is_a relathionship
EFO:0002091
FF:0000020 CD133-positive progenitor cell- bone marrow derived

Ancestor terms (non development)

CL: Cell type	UBERON: Anatomy	FF: FANTOM5
0000000 (cell)	0000468 (multi-cellular organism)	0000102 (sample by type)
0000003 (native cell)	0002371 (bone marrow)	0000210 (human sample)
0000723 (somatic stem cell)	0001474 (bone)	0000002 (in vivo cell sample)
0000048 (multi fate stem cell;;multipotent stem cell)	0002384 (connective tissue)	0000101 (sample by species)
0000988 (hematopoietic cell)	0000479 (tissue)	0000027 (CD133-positive hematopoeitic stem cell)
0000548 (animal cell)	0000062 (organ)	0000021 (hematopoeitic stem cell)
0002320 (connective tissue cell)	0004120 (mesoderm-derived structure)	0000001 (FANTOM5 Sample Ontology)
0002371 (somatic cell)	0000061 (anatomical structure)	0000020 (CD133-positive
0000255 (eukaryotic cell)		

(iii) BioMart

BioMart allows users to search our CAGE dataset for human and mouse through a widely used interface, which gives the advantage to select parameters and filters in order to retrieve the portion of the dataset that meets the search criteria instead of downloading the whole dataset.

(iv) Data file archive

A main data archive contains the bulk of all primary and contributing analysis data files available for download.

(v) FANTOM5 Table Extraction Tool (TET)

The FANTOM5 expression data is primarily distributed in compressed tab-separated-value (TSV) file format, each file consisting of the full set of CAGE peaks (184,827 rows in human and 116,277 rows in mouse) and expression values over samples (975 columns in human and 399 columns in mouse). In order to assist in the data extraction process we have created the FANTOM5 Table Extract Tool (TET). TET is intended to be a simplified way of extracting relevant sections from a curated set of FANTOM5 data tables.Using TET a user will select one of the FANTOM5 data sets, select the columns they wish to extract (i.e. samples), then specify a set of rows (i.e. cage peaks) using a regular expression search pattern, and finally view or download the resulting subset.

Figure S15: Selecting columns and rows in the FANTOM5 phase1 dataset using TET

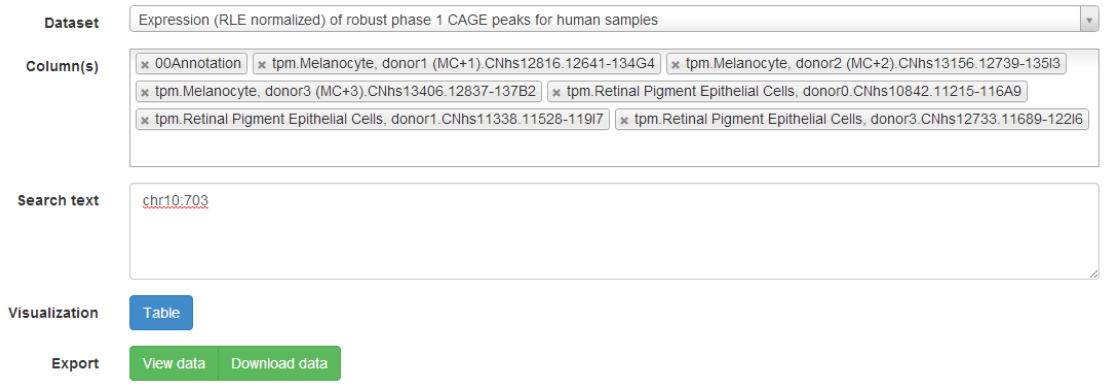
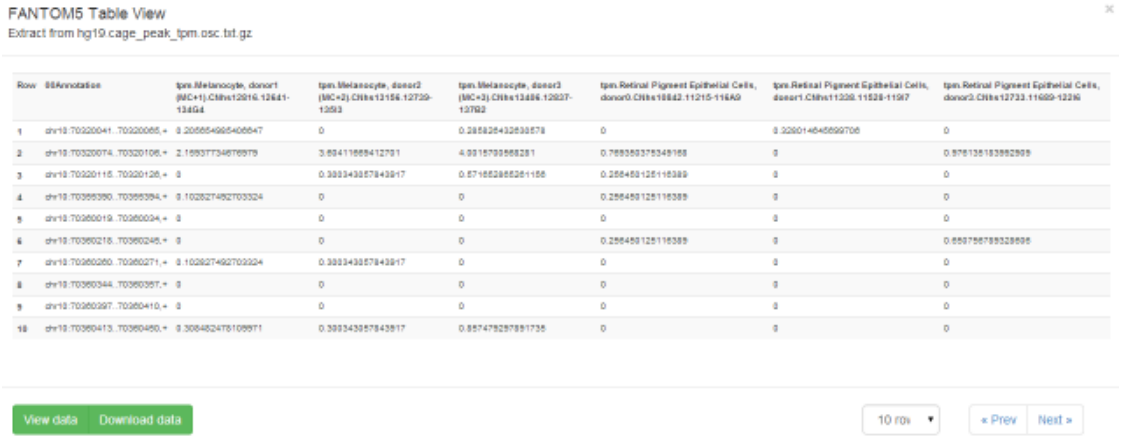


Figure S16: Example results returned by TET



(v) Promoter Slider

This enables us to set expression constraints by moving the sliders below, for primary cells and/or tissues, and export the CAGE peaks fitting to the specification.

(vi) Nanopublications

FANTOM5 CAGE clusters have been exposed using an interoperable exchange publication format called nanopublications[78,79]. A nanopublication is a schema built on top of existing semantic technology that defines (i) a single scientific assertion and (ii) provenance metadata about the assertion such as methods used to create the data and personal and institutional attribution. Nanopublications allow individual assertions and their provenance to be exposed as stand-alone, machine-readable publications.

The statements composing the nanopublication are serialized using the Resource Description Framework (RDF) such that all the entities are identified using resolvable Uniform Resource Identifiers (URIs). URIs ensure machine readability, data interoperability and permit automated search and reasoning. URIs for entities composing the Assertion and Provenance can be obtained from existing ontologies or via stand alone links created by the authors. Detailed specifications and recommendations for nanopublications are given at nanopub.org.

The primary challenges in exposing FANTOM5 data using nanopublications was to create a data model that (i) clarified the actual observations (TPM measurements) from the interpretations (scientific assertions about TSSs), (ii) mark-up this distinction in as unambiguous way as possible, and (iii) allow genomic annotations to be automatically comparable across different genome assemblies. Other considerations include economization of memory usage, time-efficient querying, and avoidance of logical inconsistencies.

In the conversion of raw CAGE datasets to nanopublications we first used the Vocabulary of Interlinked Datasets (VoID) to create a 'nanopublication compliant' RDF description of the data. In this way we make each entry in the dataset (e.g. data row or sample value) referenceable, which in turn makes it possible to specify that a particular assertion was derived from a specific row of the original dataset. To write nanopublications from the VoID dataset we then used various ontologies for representing assertions and provenance as semantic triples.

When attempting to describe genomic regions, we first investigated numerous candidate ontologies but came to the conclusion that this work is lacking for FANTOM5 purposes. Hence, we decided to develop our own ontology - Reference Sequence Ontology (RSO) to fill the gap. We want RSO to accommodate the basic CAGE region description as well as scenarios such as:
-- Allowing a single annotation be mapped onto different reference assemblies, thus providing the mechanism to compare data between FANTOM4 and FANTOM5.
-- Accommodating the most common genomic annotations, thus allowing FANTOM5 data to be cross-queried with other datasets.

As a result, we also developed a series of ontologies that are used alongside RSO to describe FANTOM5 data and the nanopubication provenance. You can access these ontologies at:
http://rdf.biosemantics.org/ontologies/referencesequence
http://rdf.biosemantics.org/ontologies/genomecomponents

http://rdf.biosemantics.org/data/genomes/humangenome
http://rdf.biosemantic.org/data/genomeassemblies/hg19
See figures below for ontology dependencies.

Figure S17: Ontologies for FANTOM5 Assertion data.

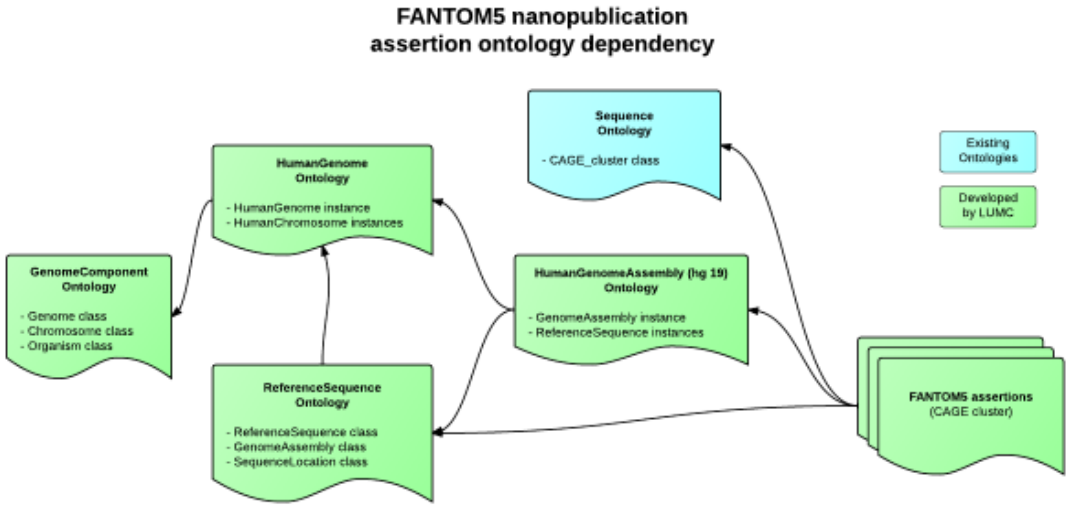
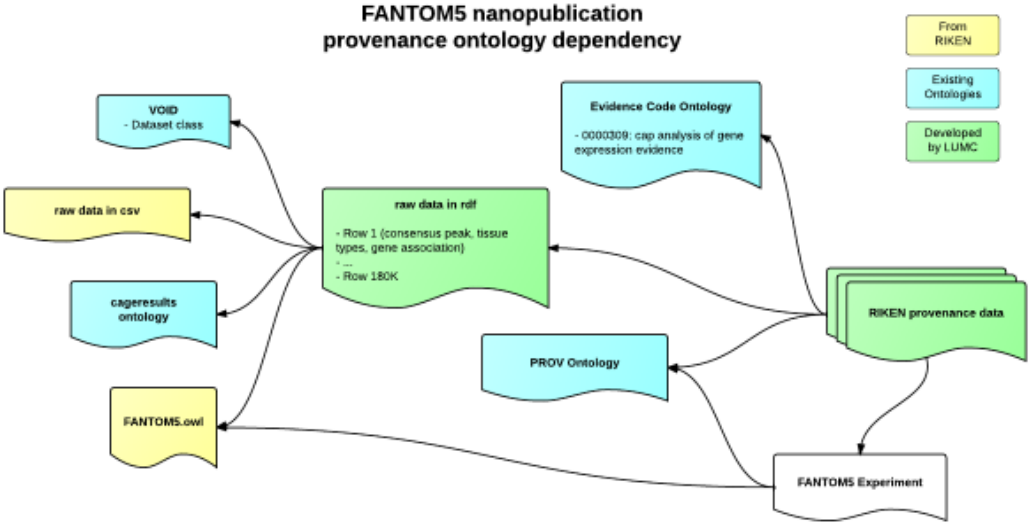


Figure S18: Ontologies for FANTOM5 provenance metadata.



From the VoID descriptions, we can expose three types of nanopublications yielding essential information from the FANTOM5 data. Type I nanopublications associate robust CAGE Clusters with genome locations. We expose 184,827 Type I nanopublications at http://rdf.biosemantics.org/ where the front page has URLs linked to FANTOM5 examples.

On the front page there is also query interface under the "Query" tab. This interface allows the user to query cage clusters by region, and show results to UCSC genome browser. From the

genome browser the user can click on an individual region, and in the description page for that region the item links back to the nanopublication under the "Outside Link".

Forthcoming Type II nanopublications associate CAGE clusters with genes, while Type III nanopublications will associate CAGE clusters with sample type. Detailed descriptions of the data models for these nanopublication assertions, and the provenance metadata can be found at examples pages nanopub.org

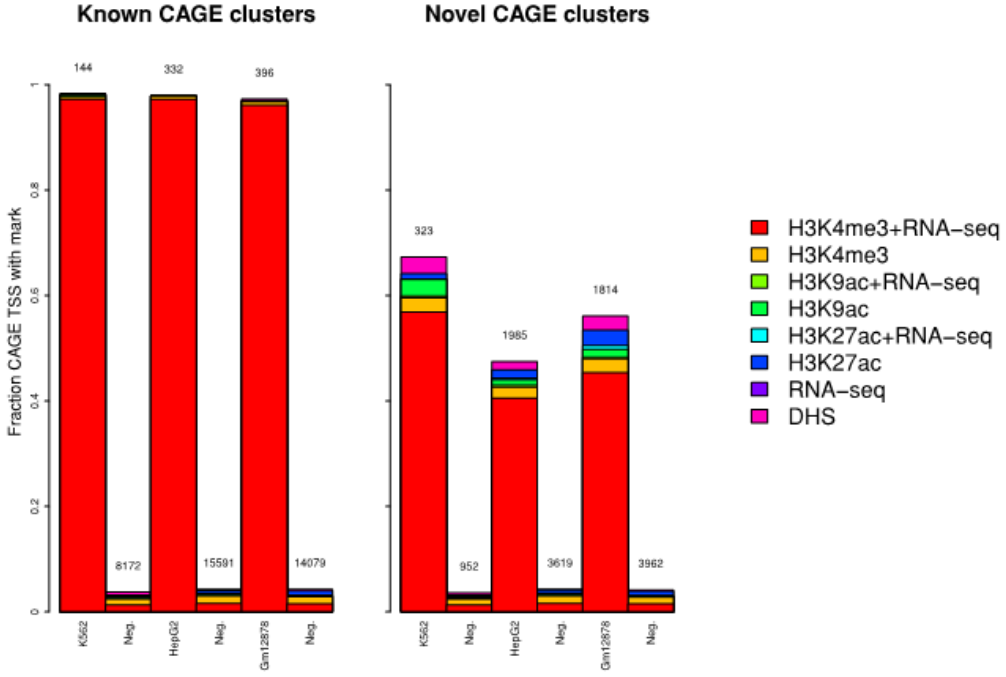
Taken together, these nanopublications expose FANTOM5 observational data (CAGE clusters) and the biological interpretations (transcriptional start sites in biological samples) in a machine-readable and interoperable form, such that these data can also be integrated with other heterogeneous data sources such as those from the ENCODE Consortium (http://genome.ucsc.edu/ENCODE/) or the Leiden Open Variation Database (http://www.lovd.nl/2.0/).

Supplementary Note 2: Support of CAGE peaks as likely TSS by independent datasets

The robust peaks were strongly-supported by 5' ESTs and cDNAs. 70% of human and 79% of mouse robust peaks were within 500 bases of a known 5' end compared to only 6.7% of a randomly-chosen data set of comparable size. The fraction of supported peaks depended on the level of prior analysis by the community. For instance, only 68% of the peaks in CD14+ monocytes were within 500 bases of a known 5' end compared to 97% and 99.9% of the peaks in HeLa and mammary fibroblasts respectively. This reflects the historical focus on sampling ESTs from these easily-accessible cell types. The large majority of libraries (89%), had > 95% of their peaks supported (see **Supplementary table 1**).

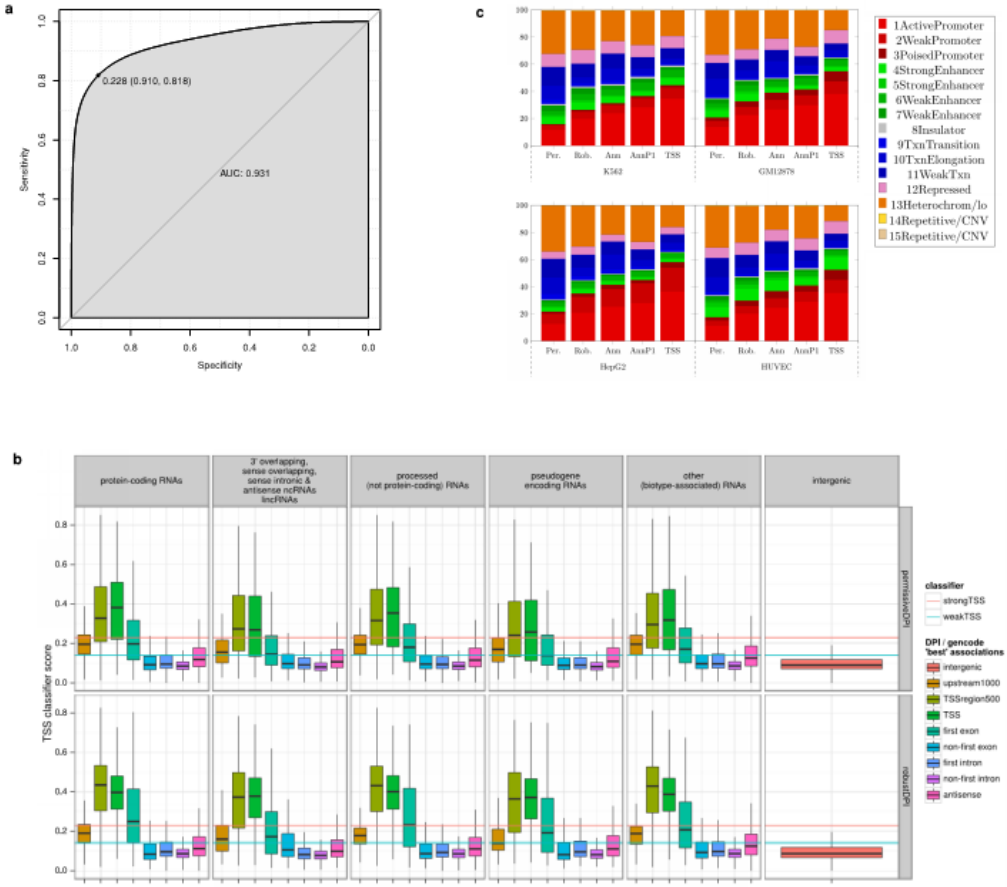
A subset of samples could be compared to alternative genome-wide datasets generated by the ENCODE consortium[19] in matching cells. For HepG2, GM12878 and K562 cell lines we found 86%, 87% and 94% of peaks were within 50 bp of the promoter-specific histone mark H3K4me3[80,81] (**Fig. S16**).

Figure S19: The H3K4me3 promoter associated histone mark is found at the majority of robust peaks. Robust CAGE peaks were annotated by their proximity (overlapping or within 50 nt) to a range of promoter associated measures. For each of the histone marks we also considered whether there was RNA-seq based evidence for transcription initiation in the form of RNA-seq reads supporting Gencode annotated first exons or de novo RNA-seq based transcript models. Annotation was applied progressively: those not annotated by H3K4me3+RNA-seq, were considered for H3K4me3, those remaining unannotated considered for H3K9ac+RNA-seq and so on through the hierachy of annotations listed in the legend. Columns labelled as "Neg." show equivalent annotation for randomly chosen TSS positions. The left panel shows robust peaks at previously annotated TSS and the right panel peaks that have not been previously annotated as TSS. Numbers above the histograms show the count of residual peaks not supported by any of the annotation categories.



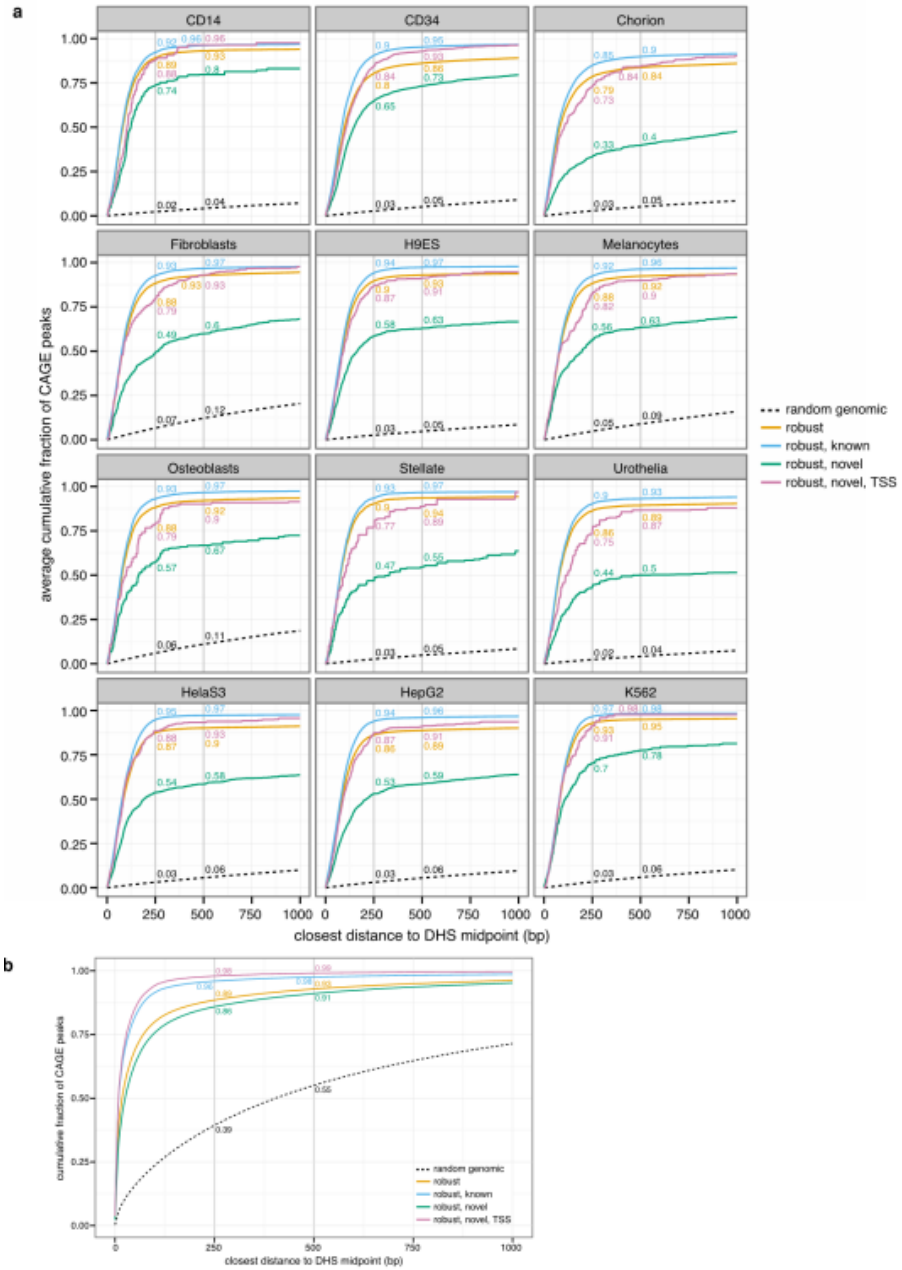
Compared to computationally predicted genome segmentations based upon ENCODE chromatin marks[20,74], robust peaks were enriched in both 'promoter' and 'enhancer'-like segments (**Fig. S17c**). This agrees with previous reports on transcription from enhancer regions[82,83] and permitted mapping of cell type-specific activity of mammalian enhancers in our companion manuscript[84].

Figure S20: TSS classifier performance metrics and overlap with ENCODE genomic segmentation. a, ROC curve showing the agreement of TSS prediction with known human promoter regions. As the standard of truth we used DPI clusters within 100bp of known models. **b**, Distribution of robust and permissiveDPI clusters' TSS classifier score across genomic segments transcriptional categories. Segmentation of the genome into transcriptional features was performed hierarchically using GencodeV16 transcript proximal start sites, exon and intron defined regions (ordered as 1/ transcripts' start site positions; 2/ 500bp proximal promoter regions; 3/ first and 4/ following exons, 5/ first and 6/ following introns; 7/ 1kb upstream regions; and finally 8/ regions within transcripts boundaries but on their opposite strand). Regions not covered are termed "intergenic". Segments were further sub-categorized based on Gencode transcripts' biotype grouped into the following broader categories : "protein-coding"; "non-coding" transcripts; "processed transcript" (kept as an independent non-coding category since it comprises a large fraction of genodeV16 transcriptome); "pseudogenes"; all other biotypes were grouped into "other (biotype-associated) RNAs". The boxplot was computed and drawn using R ggplot2. For each box of the boxplot, the middle line represents the average TSS classifier score of DPI clusters localized within a particular category, the box boundaries represent the first and last quartiles. Outliers were not plotted to preserve the clarity and simplicity of the figure. The red and blue lines emphasize the respective weakTSS (0.14) and strongTSS (0.228) score thresholds selected on the basis of the ROC curves. This figure highlights the robustness of our TSS classifier across a wide range of genomic contexts (pseudogene-encoding, protein-coding and non-coding transcripts, as well as, known TSS, their surrounding, intronic and exonic sequences) and provides an overview of the relative stringency of the weakTSS and strongTSS score thresholds. **c**, Comparison of FANTOM5 peaks with ENCODE Chromatin State Segmentation in 4 cell lines. Datasets used are all permissive (1048124), all robust (184827), all annotated (294765), all P1 annotated (39454) and all TSS classified (217572) DPIs. In each cell line the TSS classified set overlaps 50% or more with promoter annotated segments.



The peaks were also compared to the broader collection of DNase I hypersensitive site (DHS) datasets generated by the ENCODE consortium[48] (**Fig. S18**).

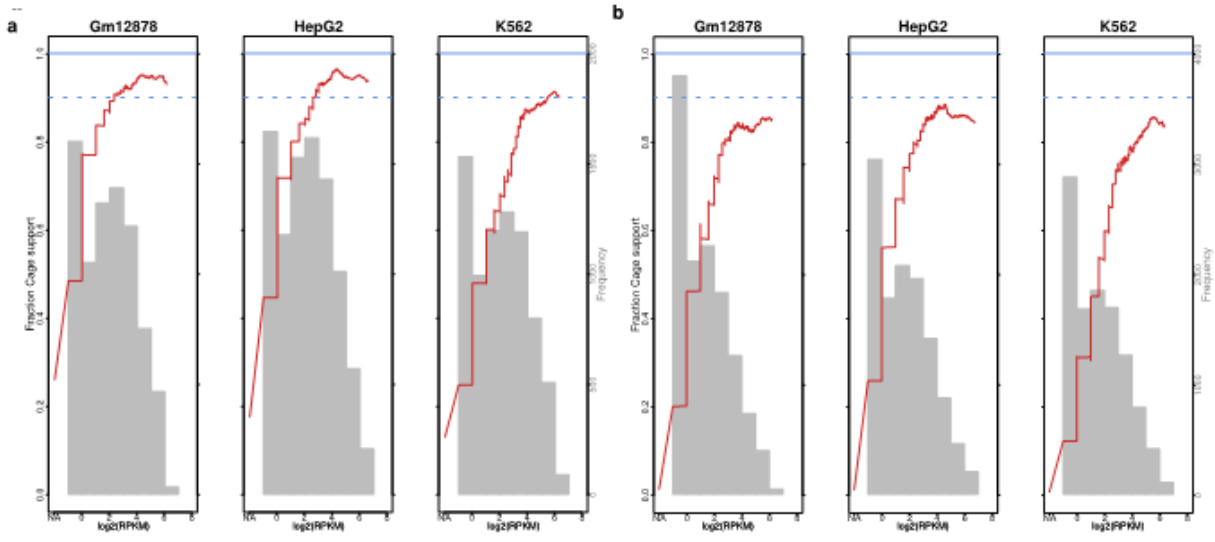
Figure S21. CAGE peak support by proximal DNase I hypersensitive sites. a, The closest distance between the midpoint of any ENCODE DNase I hypersensitive site (DHS) (Jan 2011 integration data, FDR 1% peaks) and the CAGE summit of each expressed robust CAGE peaks in matching samples between ENCODE and FANTOM were calculated. For each CAGE replicate, the cumulative fraction of expressed CAGE peaks supported by DHSs were calculated. These cumulative fractions were then averaged between replicates. This analysis was performed for robust CAGE peaks, robust CAGE peaks within 500bp of known gene 5' ends (robust, known), and robust CAGE peaks distal to 5' ends of known genes (robust, novel). The support for novel peaks predicted as TSSs by the TSS classifier (robust, novel, TSS) was also calculated. The same analysis was repeated and averaged for each sample ten times for each replicate on random genomic regions (same number as the number of expressed robust CAGE peaks in that CAGE replicate), excluding regions with assembly gaps (UCSC gap track) and ENCODE blacklisted regions (wgEncodeDacMapability- ConsensusExcludable). The horizontal axes show the distance between CAGE peak summits and DHS midpoints. The vertical axes give the average cumulative fraction of CAGE peaks with DHS support. Vertical grey lines depict CAGE-DHS distances of 250bp and 500bp. **b**, as in a except showing cumulative fraction of robust CAGE peaks (vertical axis) expressed in any CAGE library supported by proximal DNase I hypersensitive sites (DHSs) from any ENCODE sample (Jan 2011 integration data, FDR 1% peaks).



For matching CAGE and DHS libraries from HeLa-S3, 97% of the CAGE robust peaks corresponding to known 5' ends and 58% of the novel peaks were within 500bases of a DHS peak. Similarly, for matching CD14+ monocyte libraries 96% of the known and 80% of the novel peaks were supported. Less than 7% of a set of random regions sampled from the human genome were near a DHS peak in these cell types. The use of a sequence-based supervised classifier to discriminate likely TSS from post-transcriptionally generated 5' ends (**Supplementary Methods, Fig. S17**), further improved the DHS validation rates of novel peaks in HeLa from 58% to 93% and in CD14+ monocytes from 80% to 96%, respectively. We conclude that the majority of the robust peaks identified (~93%) are supported by other data. Given the high level of independent validation of TSS in cell types used within ENCODE, the large majority of novel TSS discovered in this project are also likely to be genuine TSS, discovered as a consequence of the comprehensive profiling of cells that have not previously been studied.

It is challenging to rigorously assess the limits of detection of CAGE, as it requires a fully comprehensive set of active TSS with perfect accuracy for every sample. We therefore focused on the set of well-supported candidate promoter regions in the extensively studied cell types GM12878, HepG2 and K562. We required that candidate promoters had the H3K4me3 promoter mark and were additionally supported by either RNA-seq support for GENCODE annotated TSS, or TSS defined by *de novo* RNA-seq derived transcript models. Even though RNA-seq and ChIP-seq techniques do not provide definitive identification of genuine 5' ends we found CAGE detection exceeded 90% for the highly-expressed candidate promoters and decayed as expected with decreasing expression levels (**Fig. S19**).

Figure S22. Independently identified promoter regions identified by robust CAGE tag peaks. For each cell-type shown, a set of candidate active promoter regions was identified that was supported by the H3K4me4 promoter associated histone mark and also by RNA-seq based support for **a**, GENCODE annotated TSS sites only or **b**, a TSS defined by either a de novo RNA-seq based transcript models or GENCODE. Grey histograms show the frequency distribution of the candidate active promoters relative to the RNA-seq based estimate of expression level, given as the log of reads per 1,000 nucleotides of transcript, per million RNA-seq reads (RPKM). Red curves show the fraction of those candidate active promoters supported by a robust CAGE tag cluster within 50 nt. The fraction support is calculated in bins of 1,000 candidate active promoters (step-length=1). NA on the x-axis denotes candidate promoter regions for which there were no supporting RNA-seq reads. Solid blue line is 100% robust CAGE tag support, dashed blue line is 90%.



Supplementary Note 3: Human genes absent from the collection

Known genes were downloaded from HGNC. Withdrawn symbols and the following classes were excluded from analysis (endogenous retrovirus, fragile site, immunoglobulin gene, immunoglobulin pseudogene, phenotype only, pseudogene, region, RNA cluster, RNA, micro, RNA, pseudogene, RNA, ribosomal, RNA, small cytoplasmic, RNA, small misc, RNA, small misc, RNA, small nuclear, RNA, small nucleolar, RNA, transfer, T cell receptor gene, T cell receptor pseudogene, transposable element, virus integration site, protocadherin, , complex locus and unknown).

For HGNC class 'gene with protein product' 97% [17819(17784)] of genes were observed at the permissive peak threshold leaving 3% [1225(1123)] that we did not detect. 91% [17268] were detected at the robust threshold. The numbers in round brackets indicate the subset where an accession supporting an observed transcript has been provided. A check of these lists reveals many cell-type specific genes that would not be expected to be detected in the human collection. This includes 14 interferons known to be induced upon viral challenge, opsins expressed in cones of the eye, 383 olfactory receptors expressed in olfactory epithelium. In addition duplicated genes such as *SMN1* an *SMN2* are filtered out by our requirement of mapping quality (a Q20 alignment). For these duplicated regions we are able to comment on expression patterns (see unfiltered BAM track in ZENBU browser), but are unable to say which copies are contributing to the observed pattern. We estimate for the 1225 coding genes not covered by a permissive DPI 50% correspond to genes that are expressed in rare cell types not-covered in the collection, 40% correspond to duplicated region filtering and 10% correspond to truncated transcript models (i.e. the gene is detected but the promoter is further than 500bases from the transcript model 5').

Protein coding genes not detected by a permissive DPI peak

(*ABCC6, ACCSL, ACPT, ACSM4, ACTRT3, AGAP10, AGAP4, AGAP7, AGAP8, AGAP9, AGBL1, ALG1L, AMY1A, AMY1B, AMY1C, ANHX, ANKRD20A1, ANKRD20A2, ANKRD20A3, ANKRD20A4, ANKRD60, ANXA8L1, AP5B1, APOBEC3A_B, APOL5, ARHGEF28, ARHGEF35, ARL14EPL, ARL17A, ARL17B, ARSH, ASCL4, ASCL5, ASIP, ASTL, ATP5L2, ATXN8, BLACE, BLID, BMP15, BMP2KL, BOLA2B, BPIFB3, BPY2, BPY2B, BPY2C, BTBD18, BTN1A1, BTNL10, C10orf113, C10orf85, C11orf40, C11orf44, C11orf89, C12orf55, C12orf71, C13orf35, C13orf45, C14orf177, C14orf183, C16orf3, C16orf47, C17orf112, C17orf77, C1orf134, C1orf137, C1orf147, C1orf233, C1QTNF9B, C20orf78, C2orf16, C2orf27A, C2orf27B, C2orf78, C2orf91, C3orf27, C3orf35, C3orf36, C3orf79, C4A, C4B, C4orf50, C5orf20, C7orf29, C7orf65, C7orf66, C8orf17, C8orf49, C8orf87, C9orf38, C9orf53, C9orf62, C9orf92, CASP16, CBWD1, CBWD6, CCDC177, CCL4L1, CCL4L2, CCT8L2, CDCP2, CDKL4, CDRT15L2, CDY1, CDY1B, CDY2A, CDY2B, CEACAM16, CELA1, CFC1, CFHR2, CGB2, CHP1, CHRNA10, CIB3, CKMT1A, CKMT1B, CLDN24, CLDN25, CLEC18A, CLLU1, CLLU1OS, CLRN2, CMC4, CNGA2, CNTNAP3, CPQ, CSN2, CSTL1, CT45A1, CT45A2, CT45A3, CT45A4, CT45A6, CT47A1, CT47A10, CT47A11, CT47A12, CT47A2, CT47A3, CT47A4, CT47A5, CT47A6, CT47A7, CT47A8, CT47A9, CT47B1, CTAG1A, CTAG1B, CTAG2, CTAGE4, CTAGE9, CXorf30, CXorf31, CXorf59, CXorf68, CYP11B2, DAOA, DAZ1, DAZ2, DAZ3, DAZ4, DDT, DEFB103A, DEFB104A, DEFB104B, DEFB105A, DEFB105B, DEFB106A, DEFB106B, DEFB107A, DEFB107B, DEFB108B, DEFB112, DEFB113, DEFB115, DEFB130, DEFB133, DEFB136, DNAH10OS, DPRX, DSPP, DUX4, DUX4L2, DUX4L4, DUX4L5, DUX4L6, DUX4L7, DUXA, DYTN, EBLN1, EPPIN, ERAS, ERICH2, FABP12, FAM153A, FAM155B, FAM188B2, FAM197Y1, FAM203A, FAM203B, FAM21B, FAM227A, FAM228B, FAM22A, FAM22D, FAM22F, FAM22G, FAM27A, FAM27E3, FAM27L, FAM48B2, FAM72A, FAM72B, FAM72D, FAM83E, FAM86B1, FAM86B2, FAM87A, FAM90A1, FBXW10, FCGR2B, FCGR2C, FER1L6, FFAR1, FGF16, FOLR4, FOXD4, FOXD4L1, FOXD4L2, FOXD4L3, FOXD4L4, FOXD4L5, FOXD4L6, FRG2, FRG2B, FRMPD3, FSBP, GAGE1, GAGE10, GAGE12B, GAGE12C, GAGE12D, GAGE12E, GAGE12F, GAGE12G, GAGE12H, GAGE12I, GAGE12J, GAGE13, GAGE2A, GAGE2B, GAGE2C, GAGE2D, GAGE2E, GAGE3, GAGE4, GAGE5, GAGE6, GAGE7, GATSL1, GCNT6, GCNT7, GFRA4, GGTLC2, GGTLC3, GIMD1, GJA10, GLT6D1, GOLGA6A, GOLGA6B, GOLGA6C, GOLGA6D, GOLGA6L1, GOLGA6L10, GOLGA6L2, GOLGA6L4, GOLGA6L6, GOLGA6L9, GOLGA8H, GOLGA8J, GOLGA8K, GOLGA8M, GOLGA8N, GOLGA8O, GOLGA8R, GPAT2, GPR142, GPR148, GPR151, GPR152, GPR42, GPR89C, GPX6, GRAP, GRXCR2, GSTT2B, GTF2H2, GTF2IRD2, H2AFB1, H2AFB2, H2AFB3, H2BFWT, HELZ2, HEPN1, HERC2, HHLA1, HIGD1C, HIST1H4G, HLA-DRB3, HNRNPCL1, HSFX2, HSFY1, HSFY2, HSP90AA2, HTR1F, HTR3D, HYAL4, IFNA1, IFNA10, IFNA13, IFNA14, IFNA16, IFNA17, IFNA21, IFNA4, IFNA5, IFNA6, IFNA7, IFNA8, IFNE, IFNW1, IGH@, IGK@,*

IGL@, IL22, IL25, IL28A, IL28B, IL37, KCNA10, KCNK18, KDM4E, KIAA1210, KIR2DL2, KIR2DL5A, KIR2DL5B, KIR2DS3, KIR2DS5, KIR3DL1, KIR3DL3, KIR3DS1, KLHL15, KLRC4, KPNA7, KRTAP10-11, KRTAP10-12, KRTAP10-2, KRTAP10-6, KRTAP12-1, KRTAP12-2, KRTAP12-4, KRTAP13-3, KRTAP13-4, KRTAP1-4, KRTAP15-1, KRTAP19-2, KRTAP19-4, KRTAP19-6, KRTAP19-7, KRTAP19-8, KRTAP20-1, KRTAP20-3, KRTAP2-1, KRTAP21-1, KRTAP21-3, KRTAP2-2, KRTAP22-1, KRTAP22-2, KRTAP2-3, KRTAP23-1, KRTAP25-1, KRTAP27-1, KRTAP29-1, KRTAP4-7, KRTAP4-9, KRTAP5-1, KRTAP5-10, KRTAP5-11, KRTAP5-2, KRTAP5-3, KRTAP5-5, KRTAP5-6, KRTAP5-7, KRTAP5-8, KRTAP5-9, KRTAP6-2, KRTAP6-3, KRTAP9-1, KRTAP9-2, KRTAP9-4, KRTAP9-6, LACTBL1, LALBA, LCE1B, LCE2B, LCE2C, LCE3A, LCE3B, LCE4A, LEUTX, LGALS9B, LINC00692, LIPK, LRCOL1, LRIT3, LRRC18, LRRC37A, LRRC37A2, LRRC3C, LRRIQ4, LYG2, LYPD8, MAGEA2B, MAGEA3, MAGEA9B, MAGEB5, MAGEC3, MAS1, MAS1L, MBD3L1, MBD3L2, MBD3L3, MBD3L4, MBD3L5, MC3R, MC5R, MCIN, MICALCL, MILR1, MMP21, MOS, MPC1L, MRC1, MRGPRD, MRGPRG, MRGPRX4, MS4A18, MST1L, MT-ATP6, MT-CO1, MT-CO2, MT-CO3, MT-CYB, MT-ND1, MT-ND2, MT-ND3, MT-ND4, MT-ND4L, MT-ND5, MT-ND6, MTRNR2L1, MYL10, MYO1H, NAIP, NANOGNB, NBPF11, NBPF14, NBPF16, NBPF20, NBPF24, NBPF4, NBPF5, NBPF6, NBPF7, NCBP2L, NCF1, NCR3LG1, NEU2, NKX1-1, NLRP13, NLRP5, NLRP8, NLRP9, NOBOX, NOMO2, NPIP, NPIPL2, NRTN, NTF4, NTN5, NUPR1L, NXF2, NXF2B, NXF5, OBP2A, OBP2B, OCM, OCM2, OFCC1, OMP, OPN1MW, OPN1MW2, OR10A2, OR10A3, OR10A4, OR10A5, OR10A6, OR10A7, OR10AD1, OR10AG1, OR10C1, OR10D3, OR10G2, OR10G3, OR10G4, OR10G7, OR10G8, OR10G9, OR10H1, OR10H2, OR10H3, OR10H4, OR10H5, OR10J1, OR10J3, OR10J4, OR10J5, OR10K1, OR10K2, OR10P1, OR10Q1, OR10R2, OR10T2, OR10V1, OR10W1, OR10X1, OR10Z1, OR11G2, OR11H1, OR11H12, OR11H2, OR11H4, OR11H6, OR11H7, OR11L1, OR12D2, OR12D3, OR13A1, OR13C2, OR13C3, OR13C4, OR13C5, OR13C8, OR13C9, OR13D1, OR13F1, OR13G1, OR13H1, OR13J1, OR14A16, OR14A2, OR14C36, OR14I1, OR14J1, OR14K1, OR1A1, OR1A2, OR1B1, OR1C1, OR1D2, OR1D4, OR1D5, OR1E1, OR1E2, OR1E3, OR1F1, OR1F12, OR1G1, OR1I1, OR1J1, OR1J4, OR1K1, OR1L1, OR1L3, OR1L4, OR1L6, OR1L8, OR1M1, OR1N1, OR1N2, OR1P1, OR1Q1, OR1S1, OR1S2, OR2A1, OR2A12, OR2A14, OR2A2, OR2A25, OR2A4, OR2A42, OR2A5, OR2A7, OR2AE1, OR2AG1, OR2AG2, OR2AK2, OR2AP1, OR2AT4, OR2B11, OR2B2, OR2B3, OR2B6, OR2C1, OR2C3, OR2D2, OR2D3, OR2F1, OR2F2, OR2G2, OR2G3, OR2G6, OR2H2, OR2J1, OR2J2, OR2J3, OR2K2, OR2L2, OR2L3, OR2L5, OR2L8, OR2M2, OR2M4, OR2M5, OR2M7, OR2S2, OR2T1, OR2T10, OR2T11, OR2T12, OR2T2, OR2T27, OR2T29, OR2T3, OR2T33, OR2T34, OR2T35, OR2T4, OR2T5, OR2T6, OR2T7, OR2T8, OR2V1, OR2V2, OR2W1, OR2W3, OR2Y1, OR2Z1, OR3A1, OR3A2, OR3A3, OR4A15, OR4A16, OR4A47, OR4A5, OR4B1, OR4C11, OR4C12, OR4C13, OR4C15, OR4C16, OR4C3, OR4C45, OR4C46, OR4D1, OR4D10, OR4D11, OR4D2, OR4D5, OR4D6, OR4D9, OR4E1, OR4E2, OR4F15, OR4F16, OR4F17, OR4F21, OR4F29, OR4F3, OR4F4, OR4F5, OR4F6, OR4K1, OR4K13, OR4K14, OR4K15, OR4K17, OR4K2, OR4K3, OR4K5, OR4L1, OR4M1, OR4M2, OR4N2, OR4N5, OR4P4, OR4Q2, OR4Q3, OR4S1, OR4S2, OR4X1, OR4X2, OR51A2, OR51A4, OR51A7, OR51B2, OR51B4, OR51B6, OR51D1, OR51F1, OR51F2, OR51G1, OR51G2, OR51I1, OR51I2, OR51J1, OR51L1, OR51M1, OR51Q1, OR51S1, OR51T1, OR51V1, OR52A1, OR52A5, OR52B2, OR52B4, OR52B6, OR52D1, OR52E1, OR52E2, OR52E4, OR52E5, OR52E6, OR52E8, OR52H1, OR52I1, OR52I2, OR52J3, OR52K1, OR52K2, OR52L1, OR52M1, OR52N1, OR52N2, OR52N4, OR52N5, OR52R1, OR52W1, OR52Z1, OR56A1, OR56A3, OR56A4, OR56A5, OR56B4, OR5A1, OR5A2, OR5AC1, OR5AC2, OR5AK2, OR5AL1, OR5AN1, OR5AP2, OR5AR1, OR5AS1, OR5AU1, OR5B12, OR5B17, OR5B2, OR5B21, OR5B3, OR5D13, OR5D14, OR5D16, OR5D18, OR5F1, OR5G3, OR5H1, OR5H14, OR5H15, OR5H2, OR5H6, OR5I1, OR5J2, OR5K1, OR5K2, OR5K3, OR5K4, OR5L1, OR5L2, OR5M1, OR5M10, OR5M11, OR5M3, OR5M8, OR5M9, OR5P2, OR5P3, OR5R1, OR5T1, OR5T2, OR5T3, OR5V1, OR5W2, OR6A2, OR6B1, OR6B2, OR6B3, OR6C1, OR6C2, OR6C3, OR6C4, OR6C6, OR6C65, OR6C68, OR6C70, OR6C74, OR6C75, OR6C76, OR6F1, OR6J1, OR6K2, OR6K3, OR6K6, OR6M1, OR6N1, OR6N2, OR6P1, OR6Q1, OR6S1, OR6T1, OR6V1, OR6X1, OR6Y1, OR7A10, OR7A17, OR7C2, OR7D2, OR7D4, OR7E24, OR7G1, OR7G2, OR7G3, OR8A1, OR8B12, OR8B2, OR8B3, OR8B4, OR8B8, OR8D1, OR8D2, OR8D4, OR8G2, OR8G5, OR8H1, OR8H2, OR8H3, OR8I2, OR8J1, OR8J2, OR8J3, OR8K1, OR8K3, OR8K5, OR8U1, OR8U8, OR8U9, OR9A2, OR9A4, OR9G1, OR9G4, OR9G9, OR9I1, OR9K2, OR9Q1, OR9Q2, ORM2, OTOL1, OVCH1, OVOL3, PABPN1L, PALM3, PBOV1, PCDHA@, PCDHB@, PCDHG@, PET117, PGA4, PGAM4, PGLYRP3, PINLYP, PIWIL3, PLA2G10, PLA2G4E, PLEKHG7, PLGLB1, PLGLB2, PNMA6A, PNMA6B, PNMA6C, PNMA6D, POLR2J3, POMZP3, POTEE, POTEG, POTEH, POTEI, POTEJ, POTEM, PPIAL4A, PPIAL4B, PPIAL4C, PPIAL4D, PPIAL4G, PPIP5K1, PPP5D1, PRAMEF11, PRAMEF12, PRAMEF13, PRAMEF14, PRAMEF15, PRAMEF16, PRAMEF17, PRAMEF18, PRAMEF19, PRAMEF20, PRAMEF21, PRAMEF22, PRAMEF23, PRAMEF3, PRAMEF4, PRAMEF5, PRAMEF6, PRAMEF7, PRAMEF8, PRAMEF9, PRDM7, PRLH, PROX2, PRR20A, PRR20B, PRR20C, PRR20D, PRR20E, PRR21, PRR23C, PRR25, PRSS33, PRSS42, PRSS48, PRY, PRY2, PSPN, PTGES3L, PTPN20A, PTPN20B, PYDC2, PYURF, QRICH2, R3HDML, RAB40AL, RAB44, RAB7B, RBMY1A1, RBMY1B, RBMY1D, RBMY1E, RBMY1F, RBMY1J, RD3L, REXO1L1, RFPL1, RFPL3, RFPL4A, RGPD1, RGPD8, RHOXF2, RHOXF2B, RIMBP3, RIMBP3B, RIMBP3C, RLN3, RNASE10, RNASE8, RSPH10B, RSPH10B2, RTL1, RTP2, S100A7L2, SCGB2B2, SCN10A, SCN11A, SCXA, SCXB, SDIM1, SDR42E2, SEC11B, SERF1A, SERF1B, SERPINB12, SIGLEC16, SKOR2, SLC22A25, SLC35G3, SLC35G4, SLC35G5, SLC35G6, SLC51A, SLC51B, SLCO1B7, SLFN14, SLX1A, SLX1B, SMIM1, SMIM2, SMIM8, SMN1, SMN2, SMTNL1, SPANXA1, SPANXB1, SPANXB2, SPANXE, SPANXF1, SPATA31A1, SPATA31A2, SPATA31A3, SPATA31A4, SPATA31A5, SPATA31A6, SPATA31A7, SPATA31B1, SPATA31C2, SPATA31D3, SPATA31D4, SPDYE1, SPDYE2, SPDYE5, SPDYE6, SPINK14, SPINK8, SPPL2C, SRGAP2C, SSX10, SSX2, SSX2B, SSX4, SSX4B, SSX8, STAG3L3, STARD6, STH, STRA8, STRC, SULT1A3, SULT1A4, SULT1C3, SYNE3, SYT14L, TAAR1, TAAR3, TAAR5, TAAR6, TAAR8, TAAR9, TAL2, TAS1R2, TAS2R1, TAS2R10, TAS2R13, TAS2R16, TAS2R19, TAS2R20, TAS2R3, TAS2R30, TAS2R31, TAS2R38, TAS2R39, TAS2R4, TAS2R40, TAS2R41, TAS2R42, TAS2R43, TAS2R45, TAS2R46, TAS2R5, TAS2R50, TAS2R60, TAS2R7, TAS2R8, TAS2R9, TBC1D26, TBC1D28, TBC1D29, TBC1D3, TBC1D3B, TBC1D3C, TBC1D3F, TBC1D3G, TBC1D3H, TBPL2, TCEB3C, TCEB3CL, TCEB3CL2, TECTA, TGFBR3L, TGIF2LY, THEGL, TMEM114, TMEM133, TMEM14E, TMEM178B, TMEM191B, TMEM211, TMEM236, TMEM247,

TMEM249, TMEM78, TMEM81, TMPRSS9, TP53TG3, TP53TG3B, TP53TG3C, TPBGL, TPRX1, TPTE2, TRA@, TRABD2B, TRB@, TRD@, TRG@, TRIM43, TRIM43B, TRIM49, TRIM49B, TRIM49C, TRIM49L1, TRIM64, TRIM64B, TRIM64C, TRIM73, TRIM74, TRIM75, TRPM5, TSPY1, TSPY10, TSPY3, TSTD3, TTC34, TTLL8, UBTFL1, UGT1A, USP17L10, USP17L11, USP17L12, USP17L13, USP17L15, USP17L17, USP17L18, USP17L19, USP17L20, USP17L21, USP17L22, USP17L23, USP17L24, USP17L25, USP17L26, USP17L27, USP17L28, USP17L29, USP17L30, USP17L5, USP17L8, USP9Y, UTS2R, VCX2, VCX3A, VCY, VCY1B, VHLL, VIMP, VN1R2, VN1R3, VN1R4, VN1R5, WASH1, WEE2, WFDC11, XAGE1A, XAGE1B, XAGE1C, XAGE1D, XAGE1E, XAGE5, XKRY, XKRY2, ZAR1, ZCCHC23, ZDHHC11B, ZIM3, ZNF587B, ZNF658, ZNF705B, ZNF705D, ZNF705G, ZNF728, ZNF735, ZNF80, ZNF806, ZNF888, ZPLD1, ZSCAN4, ZSCAN5B, ZSCAN5C, ZSCAN5D)

For 'RNA, long non-coding' 40% [601(443)] were detected by a permissive peak with 60% missing [909 (605)]. 24% [359] were detected by a robust peak. The numbers in round brackets indicate the subset where an accession supporting an observed transcript has been provided. The low fraction of long non-coding RNAs detected compared to protein coding genes could indicate their general lower expression levels but also the incomplete nature and lower quality of the transcript models available.

Long non-coding genes not detected by a permissive DPI peak

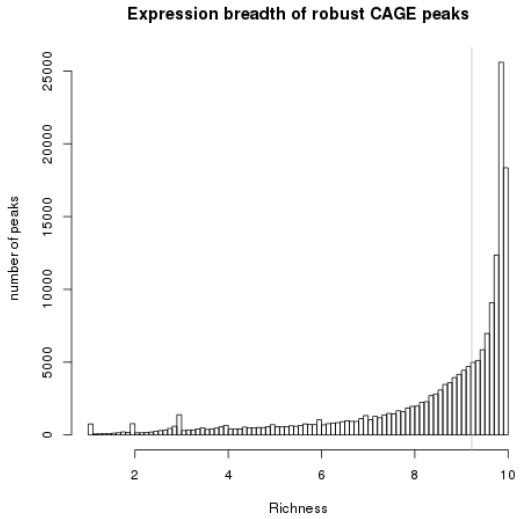
(*A2M-AS1, LRP4-AS1, LRRC2-AS1, LRRC3-AS1, LRRC3DN, LSAMP-AS1, LSAMP-AS2, LZTS1-AS1, MACC1-AS1, MACROD2-IT1, MAGI1-AS1, MAGI2-AS1, MAGI2-IT1, MANEA-AS1, MAP3K14-AS1, MAPT-AS1, MAPT-IT1, MATN1-AS1, MBNL1-AS1, MCCC1-AS1, MDC1-AS1, MED4-AS1, MEG9, MEIS1-AS1, MEIS1-AS2, MEIS1-AS3, MIR381HG, MIR600HG, MIS18A-AS1, MKNK1-AS1, MKRN3-AS1, MLIP-IT1, MME-AS1, MMP24-AS1, MORC1-AS1, MORF4L2-AS1, MPRIP-AS1, MTOR-AS1, MTUS2-AS1, MTUS2-AS2, MYB-AS1, MYCBP2-AS1, MYCBP2-AS2, MYLK-AS2, MYO16-AS2, N4BP2L2-IT2, NAALADL2-AS1, NADKD1-AS1, NAGPA-AS1, NAMA, NAPA-AS1, NARF-OT1, NAV2-AS1, NAV2-IT1, NBEA-AS1, NCBP2-AS1, NCRUPAR, NDUFB2-AS1, NEGR1-IT1, NEXN-AS1, NHS-AS1, NICN1-AS1, NKX2-2-AS1, NLGN1-AS1, NOVA1-AS1, NPSR1-AS1, NR2F2-AS1, NREP-AS1, NRG1-IT1, NRG1-IT2, NRG1-IT3, NRON, NTM-IT1, NTM-IT2, NTM-IT3, NTRK3-AS1, NUCB1-AS1, OCIAD1-AS1, OGFR-AS1, OPA1-AS1, OPCML-IT1, OPCML-IT2, OSGEPL1-AS1, OSTM1-AS1, OSTN-AS1, OTX2-AS1, OXCT1-AS1, P4HA2-AS1, PACRG-AS1, PARD6G-AS1, PCAT1, PCBP3-OT1, PCDH9-AS1, PCDH9-AS2, PCDH9-AS3, PCDH9-AS4, PCED1B-AS1, PCOLCE-AS1, PCYT1B-AS1, PDX1-AS1, PDZK1IP1-AS1, PEG3-AS1, PEX5L-AS1, PGM5-AS1, PHEX-AS1, PHKA1-AS1, PHKA2-AS1, PISRT1, PITPNA-AS1, PLCB1-IT1, PLCB2-AS1, PLCH1-AS1, PLSCR5-AS1, POTEH-AS1, POU4F1-AS1, PPEF1-AS1, PPP2R2B-IT1, PRICKLE2-AS2, PRKAG2-AS1, PRKX-AS1, PRMT5-AS1, PROX1-AS1, PROX1-IT1, PRR7-AS1, PSMD6-AS2, PSMG3-AS1, PSORS1C3, PSPC1-AS1, PSPC1-OT1, PTCSC3, PTOV1-AS1, PWRN2, PXN-AS1, RAB11B-AS1, RABGAP1L-IT1, RAI1-AS1, RAMP2-AS1, RAPGEF4-AS1, RASA2-IT1, RASA3-IT1, RASAL2-AS1, RBM12B-AS2, RBMS3-AS1, RBMS3-AS3, RC3H1-IT1, RERG-AS1, RNA45S1, RNA45S2, RNA45S3, RNA45S4, RNA45S5, RNF144A-AS1, RNF157-AS1, RNF185-AS1, RNF216-IT1, RPL34-AS1, RPS6KA2-AS1, RPS6KA2-IT1, RRM1-AS1, RSBN1L-AS1, RSF1-IT1, RSF1-IT2, SAPCD1-AS1, SATB2-AS1, SBF2-AS1, SCAANT1, SCEL-AS1, SDCBP2-AS1, SEC24B-AS1, SEC62-AS1, SETD5-AS1, SH3BP5-AS1, SH3RF3-AS1, SHANK2-AS2, SHANK2-AS3, SIDT1-AS1, SIK3-IT1, SIX3-AS1, SLC25A30-AS1, SLC26A4-AS1, SLC2A1-AS1, SLC6A1-AS1, SLC7A11-AS1, SLC8A1-AS1, SLC9A9-AS1, SLC9A9-AS2, SLFNL1-AS1, SMAD9-AS1, SMCR2, SMCR5, SMG6-IT1, SMG7-AS1, SNAI3-AS1, SNAP25-AS1, SNAP47-AS1, SNHG16, SNRK-AS1, SOCS2-AS1, SPANXA2-OT1, SPTY2D1-AS1, SRD5A3-AS1, SRGAP2-AS1, SRGAP3-AS1, SRGAP3-AS4, SRRM2-AS1, SSTR5-AS1, ST6GAL2-IT1, STARD13-AS1, STARD13-AS2, STARD13-IT1, STARD4-AS1, STAU2-AS1, STEAP3-AS1, STK4-AS1, STPG2-AS1, STT3A-AS1, STXBP5-AS1, SYNE1-AS1, SYNPR-AS1, SZT2-AS1, TAB3-AS2, TAPT1-AS1, TBC1D4-AS1, TBL1XR1-AS1, TBX5-AS1, TCEAL3-AS1, TDRG1, TET2-AS1, TGFA-IT1, THAP9-AS1, THOC7-AS1, THRB-AS1, THRB-IT1, TLR8-AS1, TMEM106A-AS1, TMEM161B-AS1, TMEM212-IT1, TMEM220-AS1, TMEM254-AS1, TMEM44-AS1, TMEM9B-AS1, TMLHE-AS1, TMPO-AS1, TMPRSS4-AS1, TNR-IT1, TOB1-AS1, TP53COR1, TPRG1-AS2, TPT1-AS1, TRAPPC12-AS1, TRHDE-AS1, TRIM31-AS1, TRMT2B-AS1, TRPC7-AS1, TRPC7-AS2, TSC22D1-AS1, TSIX, TSPAN9-IT1, TSSC1-IT1, TTC3-AS1, TTLL7-IT1, TTN-AS1, TTTY1, TTTY10, TTTY11, TTTY12, TTTY13, TTTY13B, TTTY16, TTTY17A, TTTY17B, TTTY17C, TTTY18, TTTY19, TTTY1B, TTTY2, TTTY20, TTTY21, TTTY21B, TTTY22, TTTY23, TTTY23B, TTTY2B, TTTY3, TTTY3B, TTTY4, TTTY4B, TTTY4C, TTTY5, TTTY6B, TTTY7, TTTY7B, TTTY8, TTTY8B, TTTY9A, TTTY9B, UBE2E2-AS1, UBE2Q1-AS1, UBOX5-AS1, UCKL1-AS1, UFL1-AS1, UGDH-AS1, UGGT2-IT1, ULK4-IT1, UPK1A-AS1, UPP2-IT1, USP12-AS1, USP30-AS1, USP46-AS1, VAV3-AS1, VIPR1-AS1, VPS13A-AS1, VWA8-AS1, VWC2L-IT1, WAC-AS1, WARS2-IT1, WASF3-AS1, WASIR1, WASIR2, WDFY2-AS1, WDFY3-AS1, WDR11-AS1, WDR86-AS1, WWC2-AS1, WWC3-AS1, WWTR1-IT1, XIAP-AS1, XIRP2-AS1, XXYLT1-AS1, XXYLT1-AS2, YEATS2-AS1, ZBED3-AS1, ZBTB20-AS2, ZBTB20-AS3, ZDHHC20-IT1, ZFAT-AS1, ZFHX4-AS1, ZIC4-AS1, ZMYM2-IT1, ZMYM4-AS1, ZNF205-AS1, ZNF32-AS1, ZNF32-AS2, ZNF346-IT1, ZNF582-AS1, ZNF630-AS1, ZNF674-AS1, ZNF790-AS1, ZNRF3-AS1, ZNRF3-IT1, ZRANB2-AS2*)

Supplementary Note 4: Estimates on tissue specific transcripts

Richness of expression

To quantify the breadth of expression of the robust CAGE peaks, we calculated a *richness* index[85]. This index represents the number of different libraries where the CAGE peaks would expected to be detected, if all the peaks would contain an arbitrary number of tags, here chosen to be 10. To avoid to under-estimate the expression breadth of peaks found in libraries that yielded less tags than average, and to obtain normalized expression values that are still round tag counts, we down-sampled without replacement each library to a total of one million mapped tags, using the `rrarefy` function of the R `vegan` package[86], discarding the 57 libraries where the total count was lower than one million. We also discarded 10,168 peaks that had less than 10 tags in total across all the libraries after down-sampling, as it is not possible to estimate their richness on scales smaller than 10. We then calculated the richness of each CAGE peak with a sample size of 10.

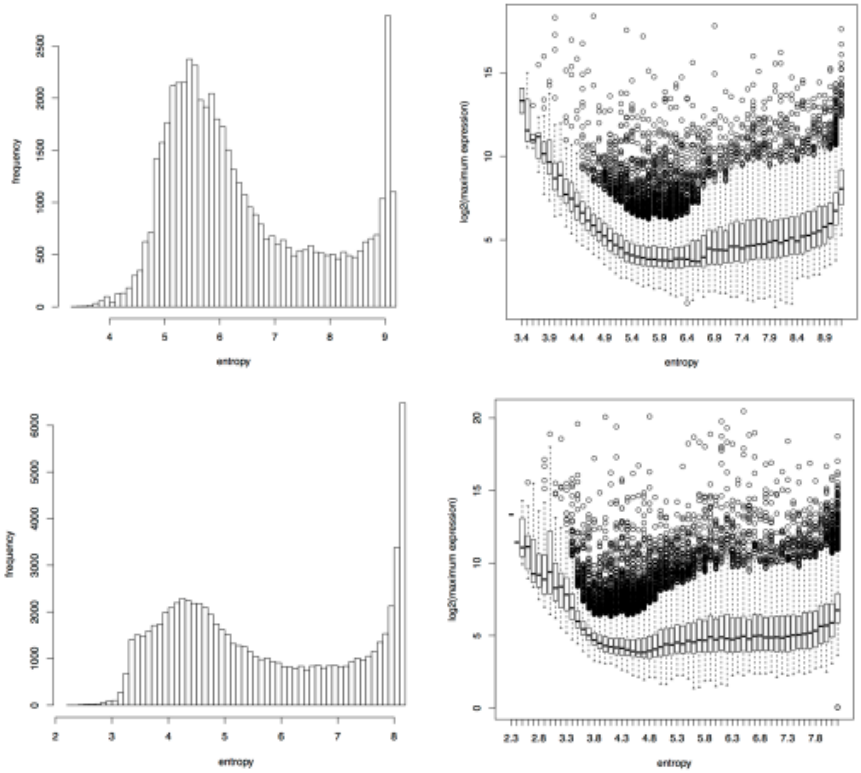
Figure S23. The distribution of richness indexes. Shows a peak for high values representing ubiquitously or very broadly expressed clusters. The median is 9.2 (vertical grey line), which is outside of the peak, showing that roughly half of the clusters are not ubiquitously expressed.



Entropy of expression

As an alternative metric of sample specificity, we calculated Shannon entropy[68], where smaller entropy means a more sample restricted (specific) pattern of expression. Averaged TPM (tags per million) values over replicates are used as expression intensities, as follows: -Σ p(r) log2(p(r)), where p(r) represent log2(TPM) value in a replicate group R.

Figure S24: Frequencies and distribution of maximum expression according to entropy bins are plotted for the human robust peaks with maximum expression > 10TPM. Human Top, Mouse Bottom.



These plots clearly show two populations of promoter activities: very restricted expression and less specific ones (left panels). The difference in mouse and human distributions is likely to reflect that the human collection contains many more states and the mouse collection predominantly contains tissue libraries (mixtures of cells).

Supplementary Note 5: Inferring key regulatory motifs in cell-type-specific promoters

The accurate definition of tissue-specific promoters enables the definition of sequence motifs that may bind tissue-specific transcription factors. We performed *de novo* motif discovery in genomic sequences in the vicinity of promoters with sample-specific expression to identify putative transcription factor binding patterns. Four complementary algorithms were used: DMF[42], HOMER[43], ChIPMunk[38], and ScanAll (Dalla *et al.*, manuscript in preparation) (**Supplementary Methods & Extended Data Figure 5**). 8,699 overrepresented motifs were identified *de novo*, including motifs resembling the consensus binding sequences of most known regulators. Importantly motifs corresponding to key regulators of cellular state such as HNF factors in hepatocytes, PU.1 and AP1 in monocytes, SOX and RFX in testis, and CRX in retina were identified in the corresponding cell-type enriched promoters (**Extended Data Figure 5b**). TomTom[51] comparison of our combined set of *de novo* motifs with the JASPAR[45], HOMER[43], SwissRegulon[46], UniPROBE[47], HOCOMOCO[39] and TRANSFAC[49] motif collections revealed that approximately 90% of all known motifs were re-identified *de novo* (**Extended Data Figure 5c**). We also rediscovered motifs similar to 234 of the 289 novel 'LEXICON' motifs recently identified in ENCODE DNase I footprints[48], providing independent confirmation of their biological relevance (**Supplementary Table 11**). 1,221 of the 8,699 *de novo* motifs did not resemble known motifs; upon clustering[40] this reduced to a set of 169 novel non-redundant motifs (see **Supplementary Methods**). While these motifs were not further functionally characterized, the significant correlation between the expression of the CAGE peaks and their associated TFBSs in 37 cases and the significant enrichment of gene ontology terms for almost all predicted TFBSs for the 169 novel motifs (see Supplemental text and table 12) suggest that some of these are recognized by novel transcription factors. Summaries on each of the novel motifs are provided online in the SSTAR resource [http://fantom.gsc.riken.jp/5/sstar/Browse_Novel_motifs].

Supplementary Note 6: Transcription factors absent from the collection

For the 50 missing human TFs, 15 were from duplicated regions of the genome and 35 were not detected. Orthologs for 24 of these were detected in the mouse embryonic development samples while the remaining TFs are expressed in early developmental samples missing from both species collections.

The 24 TFs that were not detected in the human collection that were detected in the mouse robust cluster set: (*ARID3C* (neonatal eyeball), *ASCL3* (submandibular gland), *BHLHA9* (neonatal skin), *CPXCR1* (testis), *DMRT3* (hematopoietic progenitors, hippocampal astrocytes), *EVX2* (Developing forelimb[87]), *FIGLA* (embryonic ovary and eyeball, neonatal ovary, eyeball, testis), *HELT* (hippocampal neurons, granule cells), *HSFY2* (testis), *KLF14* (E14-E18 various organs), *MBD3L1* (testis), *OTP* (neurons), *PRDM12* (embryo and neurons), *PRDM9* (various), *PROP1* (embryonic pituitary, down regulated during development), *PROX2* (testis), *SKOR2* (spinal neurons, granule cells), *TAL2* (granule cells, macrophages, striatal neurons), *TFAP2D* (granule cells, hippocampal and striatal neurons), *TFAP2E* (neonatal skin and eyeball), *YY2* (neonatal testis), *ZC3H12B* (adult brain regions), *ZFP28* (neonatal hippocampus, corpus striatum, eyeball, neurons), *ZFP92* (corpora quadrigemina, pituitary gland, neurons), *ZNF286A* (various tissues)). These tended to be detected in samples where we did not have a matching human counterpart sample. In particular this corresponded to transcription factors expressed in neuronal subsets and embryonic or neonatal development. Interestingly four of these were from testis, which is represented in the human collection. Due to the cellular complexity of testis it may be that the ratio of cell types is different in human and mouse as testis volume scales and hence these weakly expressed testis specific genes are not detected in human.

E.g. *MBD3L1* is only expressed in round spermatids[88] and *YY2* in spermatocytes but not sperm cells[76]. Perhaps isolated cell types in testis would recover this.

A set of 16 transcriptional regulators with homologs in both human and mouse and with uniquely mappable promoter regions were not detected in the robust clusters for either species (*ASCL4, CDX4, GSC2, MSGN1, NKX1-1, NOBOX, NOTO, OVOL3, SRY, TBPL2, TBX10, UBTFL1, ZNF648, ZIM3, ZSCAN4, ZSCAN5B*). Checking the literature these factors are expressed in more rare samples:

Early embryogenesis: E.g. *CDX4*[89], *MSGN1* (paraxial mesoderm[90]) *NKX1-1*[91], (*SRY* testis formation[92]), *TBX10* (rhombomere 4 and rhombomere 6, Hindbrain development[93]), *UBTFL1* (preimplantation-specific[94]), *ZIM3* (meiotic cells in testis[95]), *ZSCAN4* (2-cell stage embryos[96]),

Rare cell types in specialized regions: *GSC2* (interpeduncular nucleus[97]), *NOBOX* (primordial and growing oocytes[98]), *NOTO* (organizer node and in the nascent notochord[99]), *TBPL2* (oocyte[100]).

For three of these factors a review of the literature could not find any information on their expression patterns (*OVOL3, ZNF648, ZSCAN5B*), perhaps indicating they are not expressed, and while *ASCL4* was reported to be expressed in fetal skin[101], we find no such expression in either mouse or human fetal skin.

Supplementary Note 7: Comparison of top TFs with mouse phenotypes

For each sample we generated ranked lists of top TFs based on the expression of a TF promoter in a sample compared to the median expression across the entire collection [log10(expression + 0.1) –log10(median expression + 0.1)]. The results of which are available online using the SSTAR tool [http://fantom.gsc.riken.jp/5/sstar/Main_Page]. To demonstrate the likely relevance of the top transcription factors identified in each sample we examined the fraction of top 20 TF promoters with relevant knockout mouse phenotypes at the MGI[102,103](retrieved July 2013). We specifically focused on a subset of 316 human primary cell samples and 89 mouse primary cell samples. The remaining primary cell samples were excluded from this analysis either because they were treated samples (pathogens, expanded in culture etc. e.g. CD14+ monocytes - treated with Candida) or there was no clear way of assessing them for relevant phenotypes (e.g. mesenchymal precursor cell - ovarian cancer left ovary). For the set of top TFs identified in the triaged set of primary cells we then downloaded all associated mammalian phenotypes and manually scored associations. Examples of manual associations include

Atoh1+ Inner ear hair cells [MP:0001967: deafness, MP:0006325: impaired hearing, MP:0004699: unilateral deafness, MP:0002855: abnormal cochlear ganglion morphology]

Reticulocytes [MP:0000245: abnormal erythropoiesis, MP:0002026: leukemia, MP:0002874: decreased hemoglobin content, MP:0002875: decreased erythrocyte cell number, MP:0010763: abnormal hematopoietic stem cell physiology]

Renal Proximal Tubular Epithelial Cell [MP:0000527: abnormal kidney development, MP:0002135: abnormal kidney morphology, MP:0002703: abnormal renal tubule morphology, MP:0002989: small kidney, MP:0003918: decreased kidney weight, MP:0000520: absent kidney]

Adipocyte [MP:0002644: decreased circulating triglyceride level, MP:0010025: decreased total body fat amount, MP:0000013: abnormal adipose tissue distribution, MP:0001783: decreased white adipose tissue amount, MP:0002118: abnormal lipid homeostasis, MP:0008844: decreased subcutaneous adipose tissue amount, MP:0001547: abnormal lipid level, MP:0009115: abnormal fat cell morphology, MP:0000187: abnormal triglyceride level, MP:0001261: obese]

Considering the top 20 cell type enriched TF promoters for the triaged set 61% of mouse and 40% of human TFs for which there were knockouts available had relevant phenotypes.

References for supplementary information

1 Pradhan, S. *et al.* Perlecan domain IV peptide stimulates salivary gland cell assembly in vitro. *Tissue Eng Part A* **15**, 3309-3320 (2009).

2 Lee, W. J., Cha, H. W., Sohn, M. Y., Lee, S. J. & Kim do, W. Vitamin D increases expression of cathelicidin in cultured sebocytes. *Arch Dermatol Res* **304**, 627-632 (2012).

3 Ohshima, M., Yamaguchi, Y., Micke, P., Abiko, Y. & Otsuka, K. In vitro characterization of the cytokine profile of the epithelial cell rests of Malassez. *J Periodontol* **79**, 912-919 (2008).

4 You, Y., Richer, E. J., Huang, T. & Brody, S. L. Growth and differentiation of mouse tracheal epithelial cells: selection of a proliferative population. *Am J Physiol Lung Cell Mol Physiol* **283**, L1315-1321 (2002).

5 Kajiya, K., Hirakawa, S., Ma, B., Drinnenberg, I. & Detmar, M. Hepatocyte growth factor promotes lymphatic vessel formation and function. *EMBO J* **24**, 2885-2895 (2005).

6 Hori, S., Nomura, T. & Sakaguchi, S. Control of regulatory T cell development by the transcription factor Foxp3. *Science* **299**, 1057-1061 (2003).

7 Kobune, F. *et al.* A novel monolayer cell line derived from human umbilical cord blood cells shows high sensitivity to measles virus. *J Gen Virol* **88**, 1565-1567 (2007).

8 Watanabe, A. *et al.* Peroxiredoxin 1 is required for efficient transcription and replication of measles virus. *J Virol* **85**, 2247-2253 (2010).

9 Ikawa, T. *et al.* An essential developmental checkpoint for production of the T cell lineage. *Science* **329**, 93-96 (2010).

10 Klinken, S. P., Nicola, N. A. & Johnson, G. R. In vitro-derived leukemic erythroid cell lines induced by a raf- and myc-containing retrovirus differentiate in response to erythropoietin. *Proc Natl Acad Sci U S A* **85**, 8506-8510 (1988).

11 Itoh, M. *et al.* Automated workflow for preparation of cDNA for cap analysis of gene expression on a single molecule sequencer. *PLoS One* **7**, e30809 (2012).

12 Kanamori-Katayama, M. *et al.* Unamplified cap analysis of gene expression on a single-molecule sequencer. *Genome Res* **21**, 1150-1159 (2011).

13 Myers, G. A fast bit-vector algorithm for approximate string matching based on dynamic programming. *Journal of the ACM (JACM)* **46**, 395-415 (1999).

14 Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489**, 101-108 (2012).

15 Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* **18**, 1851-1858 (2008).

16 Oja, E., Hyvarinen, A. & Karhunen, J. (John Wiley & Sons, 2001).

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

18 Wootton, J. C. & Federhen, S. Analysis of compositionally biased regions in sequence databases. *Methods Enzymol* **266**, 554-571 (1996).

19 Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).

20 Ernst, J. & Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol* **28**, 817-825 (2010).

21 Hoffman, M. M. *et al.* Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat Methods* **9**, 473-476 (2012).

22 Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140 (2009).

23 Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol* **11**, R106 (2010).

24 Pruitt, K. D., Tatusova, T., Brown, G. R. & Maglott, D. R. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res* **40**, D130-135 (2011).

25 Hsu, F. *et al.* The UCSC Known Genes. *Bioinformatics* **22**, 1036-1046 (2006).

26 Harrow, J. *et al.* GENCODE: producing a reference annotation for ENCODE. *Genome Biol* **7 Suppl 1**, S4 1-9 (2006).

27 Flicek, P. *et al.* Ensembl 2011. *Nucleic Acids Res* **39**, D800-806 (2010).

28 Meehan, T. F. *et al.* Logical development of the cell ontology. *BMC Bioinformatics* **12**, 6 (2011).

29 Mungall, C. J., Torniai, C., Gkoutos, G. V., Lewis, S. E. & Haendel, M. A. Uberon, an integrative multi-species anatomy ontology. *Genome Biol* **13**, R5 (2012).

30 Osborne, J. D. *et al.* Annotating the human genome with Disease Ontology. *BMC Genomics* **10 Suppl 1**, S6 (2009).

31 Malone, J. *et al.* Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics* **26**, 1112-1118 (2010).

32 Day-Richter, J., Harris, M. A., Haendel, M. & Lewis, S. OBO-Edit--an ontology editor for biologists. *Bioinformatics* **23**, 2198-2200 (2007).

33 Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 289-300 (1995).

34 Fan, W., McCloskey, J. & Yu, P. S. in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining.* 136-146 (ACM).

35 Fan, W., Wang, H., Yu, P. S. & Ma, S. in *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on.* 51-58 (IEEE).

36 Robin, X. *et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 77 (2011).

37 Yu, X., Lin, J., Zack, D. J. & Qian, J. Computational analysis of tissue-specific combinatorial gene regulation: predicting interaction between transcription factors in human tissues. *Nucleic Acids Res* **34**, 4925-4936 (2006).

38 Kulakovskiy, I. V., Boeva, V. A., Favorov, A. V. & Makeev, V. J. Deep and wide digging for binding motifs in ChIP-Seq data. *Bioinformatics* **26**, 2622-2623 (2010).

39 Kulakovskiy, I. V. *et al.* HOCOMOCO: a comprehensive collection of human transcription factor binding sites models. *Nucleic Acids Res* **41**, D195-202 (2012).

40 Vorontsov, I. E., Kulakovskiy, I. V. & Makeev, V. J. Jaccard index based similarity measure to compare transcription factor binding site models. *Algorithms Mol Biol* **8**, 23 (2013).

41 Huang, E., Yang, L., Chowdhary, R., Kassim, A. & Bajic, V. An algorithm for ab initio DNA motif detection. *Information Processing and Living Systems*, 611-614 (2005).

42 Marchand, B., Bajic, V. B. & Kaushik, D. K. in *Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis.* 56 (ACM).

43 Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**, 576-589 (2010).

44 Ho Sui, S. J. *et al.* oPOSSUM: identification of over-represented transcription factor

binding sites in co-expressed genes. *Nucleic Acids Res* **33**, 3154-3164 (2005).

45 Portales-Casamar, E. *et al.* JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res* **38**, D105-110 (2009).

46 Pachkov, M., Erb, I., Molina, N. & van Nimwegen, E. SwissRegulon: a database of genome-wide annotations of regulatory sites. *Nucleic Acids Res* **35**, D127-131 (2007).

47 Berger, M. F. *et al.* Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell* **133**, 1266-1276 (2008).

48 Neph, S. *et al.* An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489**, 83-90 (2012).

49 Matys, V. *et al.* TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res* **34**, D108-110 (2006).

50 Sokal, R. R. A statistical method for evaluating systematic relationships. *Univ Kans Sci Bull* **38**, 1409-1438 (1958).

51 Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L. & Noble, W. S. Quantifying similarity between motifs. *Genome Biol* **8**, R24 (2007).

52 Cock, P. J. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422-1423 (2009).

53 Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res* **14**, 1188-1190 (2004).

54 McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol* **28**, 495-501 (2010).

55 van Dongen, S. M. Graph clustering by flow simulation. (2000).

56 Theocharidis, A., van Dongen, S., Enright, A. J. & Freeman, T. C. Network visualization and analysis of gene expression data using BioLayout Express(3D). *Nat Protoc* **4**, 1535-1550 (2009).

57 Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* **30**, 1575-1584 (2002).

58 Brohee, S. & van Helden, J. Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics* **7**, 488 (2006).

59 Freeman, T. C. *et al.* Construction, visualisation, and clustering of transcription networks from microarray expression data. *PLoS Comput Biol* **3**, 2032-2042 (2007).

60 A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol* **9**, e1001046 (2011).

61 Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25 (2009).

62 Pham, T. H. *et al.* Dynamic epigenetic enhancer signatures reveal key transcription factors associated with monocytic differentiation states. *Blood* **119**, e161-171 (2012).

63 Suzuki, H. *et al.* The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. *Nat Genet* **41**, 553-562 (2009).

64 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).

65 Hubbard, T. J. *et al.* Ensembl 2009. *Nucleic Acids Res* **37**, D690-697 (2009).

66 Croft, D. *et al.* Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res* **39**, D691-697 (2011).

67 Pico, A. R. *et al.* WikiPathways: pathway editing for the people. *PLoS Biol* **6**, e184 (2008).

68 Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27-30 (2000).

69 Kandasamy, K. *et al.* NetPath: a public resource of curated signal transduction pathways. *Genome Biol* **11**, R3 (2010).

70 Sayers, E. W. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **40**, D13-25 (2012).

71 Team, R. D. C. (ISBN 3-900051-07â€“0. Available at: http://www. R-project. org, 2008).

72 Pages, H. *et al.* Annotation Database Interface. *R package version* **1** (2008).

73 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).

74 Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43-49 (2011).

75 Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137 (2008).

76 Severin, J. *et al.* Interactive visualization and analysis of large-scale NGS data-sets

using ZENBU. *Nature Biotechnology* (2014).

77 Kasprzyk, A. BioMart: driving a paradigm change in biological data management. *Database (Oxford)* **2011**, bar049 (2011).

78 Groth, P., Gibson, A. & Velterop, J. The anatomy of a nanopublication. *Information Services and Use* **30**, 51-56 (2010).

79 Mons, B. *et al.* The value of data. *Nat Genet* **43**, 281-283 (2011).

80 Bernstein, B. E. *et al.* Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell* **120**, 169-181 (2005).

81 Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39**, 311-318 (2007).

82 Kim, T. K. *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**, 182-187 (2010).

83 Tuan, D., Kong, S. & Hu, K. Transcription of the hypersensitive site HS2 enhancer in erythroid cells. *Proc Natl Acad Sci U S A* **89**, 11219-11223 (1992).

84 Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* (2014).

85 Hurlbert, S. H. The nonconcept of species diversity: a critique and alternative parameters. *Ecology* **52**, 577-586 (1971).

86 Oksanen, J. *et al.* vegan: Community Ecology Package, 2007. *R package version* **1** (2009).

87 Herault, Y., Hraba-Renevey, S., van der Hoeven, F. & Duboule, D. Function of the Evx-2 gene in the morphogenesis of vertebrate limbs. *EMBO J* **15**, 6727-6738 (1996).

88 Jiang, C. L. *et al.* MBD3L1 and MBD3L2, two new proteins homologous to the methyl-CpG-binding proteins MBD2 and MBD3: characterization of MBD3L1 as a testis-specific transcriptional repressor. *Genomics* **80**, 621-629 (2002).

89 Kinkel, M. D., Eames, S. C., Alonzo, M. R. & Prince, V. E. Cdx4 is required in the endoderm to localize the pancreas and limit beta-cell number. *Development* **135**, 919-929 (2008).

90 Yoon, J. K., Moon, R. T. & Wold, B. The bHLH class protein pMesogenin1 can specify paraxial mesoderm phenotypes. *Dev Biol* **222**, 376-391 (2000).

91 Simon, R. & Lufkin, T. Postnatal lethality in mice lacking the Sax2 homeobox gene

homologous to Drosophila S59/slouch: evidence for positive and negative autoregulation. *Mol Cell Biol* **23**, 9046-9060 (2003).

92 Koopman, P., Munsterberg, A., Capel, B., Vivian, N. & Lovell-Badge, R. Expression of a candidate sex-determining gene during mouse testis differentiation. *Nature* **348**, 450-452 (1990).

93 Xue, X. D. *et al.* A unique expression pattern of Tbx10 in the hindbrain as revealed by Tbx10(LacZ) allele. *Genesis* **48**, 295-302 (2010).

94 Yamada, M. *et al.* Involvement of a novel preimplantation-specific gene encoding the high mobility group box protein Hmgpi in early embryonic development. *Hum Mol Genet* **19**, 480-493 (2009).

95 Kim, J., Bergmann, A., Wehri, E., Lu, X. & Stubbs, L. Imprinting and evolution of two Kruppel-type zinc-finger genes, ZIM3 and ZNF264, located in the PEG3/USP29 imprinted domain. *Genomics* **77**, 91-98 (2001).

96 Falco, G. *et al.* Zscan4: a novel gene expressed exclusively in late 2-cell embryos and embryonic stem cells. *Dev Biol* **307**, 539-550 (2007).

97 Gong, S. *et al.* A gene expression atlas of the central nervous system based on bacterial artificial chromosomes. *Nature* **425**, 917-925 (2003).

98 Suzumori, N., Yan, C., Matzuk, M. M. & Rajkovic, A. Nobox is a homeobox-encoding gene preferentially expressed in primordial and growing oocytes. *Mech Dev* **111**, 137-141 (2002).

99 Beckers, A., Alten, L., Viebahn, C., Andre, P. & Gossler, A. The mouse homeobox gene Noto regulates node morphogenesis, notochordal ciliogenesis, and left right patterning. *Proc Natl Acad Sci U S A* **104**, 15765-15770 (2007).

100 Muller, F. & Tora, L. TBP2 is a general transcription factor specialized for female germ cells. *J Biol* **8**, 97 (2009).

101 Jonsson, M. *et al.* Hash4, a novel human achaete-scute homologue found in fetal skin. *Genomics* **84**, 859-866 (2004).

102 Eppig, J. T., Blake, J. A., Bult, C. J., Kadin, J. A. & Richardson, J. E. The Mouse Genome Database (MGD): comprehensive resource for genetics and genomics of the laboratory mouse. *Nucleic Acids Res* **40**, D881-886 (2012).

103 Smith, C. L. & Eppig, J. T. The Mammalian Phenotype Ontology as a unifying standard for experimental and high-throughput phenotyping data. *Mamm Genome* **23**, 653-668 (2012).