Table 1 | Summary of peaks, coverage and genes hit in FANTOM5

	Human							Mouse						
	Peaks	Stranded genome coverage (bp)		Number of aligned reads		Genes hit	Peaks per gene	Peaks	Stranded genome coverage (bp)		Number of aligned reads		Genes hit	Peaks per gene
The whole genome	—	6.2 × 109	100%	4.5×10^9	100%	—	—	—	5.3×10^9	100%	1.9×10^9	100%	—	—
'Permissive' CAGE peaks	1,048,124	1.4×10^7	0.22%	3.6×10^9	80%	20,808	—	652,860	8.4×10^6	0.16%	1.5×10^9	79%	20,480	—
(A) Within 500 bp of annotated 5′	245,514	4.3×10^6	0.07%	3.0×10^9	68%	20,808	11.8	146,185	2.5×10^6	0.05%	1.3×10^9	69%	20,480	7.1
(B) TSS classifier positive	217,572	4.0×10^6	0.06%	2.9×10^9	64%	18,503	—	129,466	2.4×10^6	0.05%	1.0×10^9	52%	17,088	—
(A or B) Likely TSS	308,214	5.3×10^6	0.09%	3.2×10^9	72%	20,808	—	173,564	3.0×10^6	0.06%	1.4×10^9	70%	20,480	—
'Robust' CAGE peaks	184,827	3.9×10^6	0.06%	3.5×10^9	77%	18,961	—	116,277	2.5×10^6	0.05%	1.4×10^9	75%	19,001	—
(A) Within 500bp of annotated 5′	82,150	2.2×10^6	0.04%	3.0×10^9	66%	18,961	4.3	61,134	1.6×10^6	0.03%	1.3×10^9	68%	19,001	3.2
(B) TSS classifier positive	76,445	2.1×10^6	0.03%	2.9×10^9	63%	17,285	—	51,611	1.4×10^6	0.03%	9.9×10^8	51%	16,028	—
(A or B) Likely TSS	92,783	2.4×10^6	0.04%	3.2×10^9	70%	18,961	—	77674	1.7×10^6	0.03%	1.3×10^9	69%	19,001	—
Cross-species projected robust peaks	70,351	1.6×10^6	0.03%	—	—	—	—	105,157	2.4×10^6	0.04%	—	—	—	—
'Homologous' robust peaks	34,041	1.0×10^6	0.02%	—	—	—	—	42,423	1.3×10^6	0.02%	—	—	—	—

confirmed this general observation (Extended Data Fig. 2), however, for the first time the greater depth of sequencing enabled identification of the preferred TSS within broad promoters. Taking each library in turn, using the location of the dominant TSS (that is, the TSS with the highest number of tags), we searched for phased WW dinucleotides (AA/AT/TA/TT) associated with nucleosome location[14] (Extended Data Fig. 2). Remarkably, on a genome-wide scale, there was a periodic spacing of WW motifs with a 10.5 bp repeat downstream of the dominant TSS, exactly as shown previously for well-phased H2A.Z nucleosomes[14] (Extended Data Fig. 2d). The precise phasing was supported further by the pattern of H2A.Z and H3K4me3 chromatin immunoprecipitation sequencing (ChIP-seq) signal seen around TSS in CD14$^+$ monocytes and frontal lobe respectively (Extended Data Fig. 2e, f). This observation indicates that the positioned nucleosome is a key indicator of start site preference in broad promoters.

Expression levels and tissue specificity

The raw tag counts under the DPI peak coordinates were used to generate an expression table across the entire collection. Normalized tags per million (TPM) were then calculated using the relative log expression (RLE) method in edgeR[15]. Almost all peaks (96%) were reproducibly detected above 1 TPM in at least two samples, but most were detected in less than half the samples. Examining the distribution of expression level and breadth across the collection, we classified the 185K robust human peak expression profiles as non-ubiquitous (cell-type-restricted, 80%), ubiquitous-uniform ('housekeeping', 6%) or ubiquitous-non-uniform (14%) (Fig. 2a, b). We define ubiquitous as detected in more than 50% of samples (median >0.2 TPM) and uniform as a less than tenfold difference between maximum and median expression. Estimation using the smaller mouse expression data set or human primary cell, cell line or tissue data subsets resulted in different fractions, yet in all cases ubiquitous-uniform expression profiles were in the minority (Extended Data Fig. 3a–e). Alternative measures such as richness index and Shannon entropy confirm that only a minor fraction of transcripts can be considered as genuine housekeeping genes with broad and uniform expression (Supplementary Note 4 and Supplementary Table 4 for a list of housekeeping genes). In addition many of the 1,225 known genes that were missed in the collection are known to be specifically expressed in cell types that are not easily procured; indicating that even more of the mammalian transcriptome has a cell-type-restricted expression

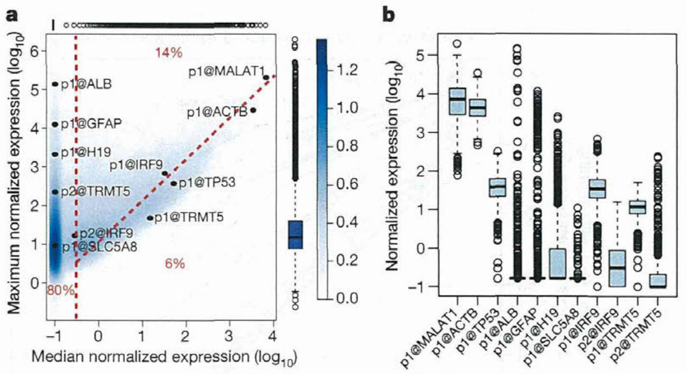

Figure 2 | Cell-type-restricted and housekeeping transcripts encoded in the mammalian genome. **a**, Density plot summarizing the distribution of relative log expression (RLE) normalized maximum and median TPM expression values for the 185K robustly detected human peaks identified by FANTOM5 (colour bar on right indicates relative density). Box and whiskers plots above and to right show distribution of median and maximum values in the data set (box shows the interquartile range). Promoters of named genes are highlighted to show extremes of expression level and expression breadth, note the alternative promoters of *IRF9* and *TRMT5* have different maximums and breadths of expression (see Extended Data Fig. 10). Fraction on left of the red vertical dashed line corresponds to peaks detected in less than 50% of samples with non-ubiquitous (cell-type-restricted) expression patterns (median <0.2 TPM). Fraction below the red diagonal dashed line corresponds to ubiquitous-uniform (housekeeping) expression profiles (maximum < 10× median). Fraction above diagonal and to the right of the vertical dashed lines corresponds to ubiquitous-non-uniform expression profiles (maximum > 10× median). **b**, Box and whisker plots showing the distribution of expression levels for the same peaks as in **a** across the 889 samples (box shows the interquartile range).

pattern (Supplementary Note 3). In overview, the data confirm the argument that most genes are regulated in a tissue-dependent manner[16]. According to Gene Ontology enrichment analysis[17] of genes within each of the three classes (Supplementary Table 5), the non-ubiquitous genes were enriched for proteins involved in cell–cell signalling, plasma membrane receptors, cell adhesion molecules and signal transduction, whereas genes in the housekeeping set were enriched for components of the ribonucleoprotein complex and RNA processing. The ubiquitous-non-uniform set was enriched for cell cycle genes, with 204 of the 268 human genes annotated with the 'mitotic cell cycle' term, a reflection of the fact that the fraction of actively proliferating cells inevitably varies greatly across the collection.

Finally, of the 104,859 peaks expressed at 10 TPM (~3 copies per cell[18]) or greater, an average primary cell sample expressed a median of 8,757 including peaks for 430 transcription factor mRNAs (Extended Data Fig. 3f, g).

Promoter conservation between human and mouse

Regulatory regions such as transcription factor binding sites are often, but not always, located in conserved and orthologous regions[19]. Overall human TSSs were significantly enriched in evolutionarily conserved regions compared to the genome-wide null expectation, with 38% overlapping previously defined mammalian constrained elements (Fisher's exact test, odds ratio 10.2, P value $< 2.2 \times 10^{-16}$; see Supplementary Methods). Despite this general level of conservation, there is evidence of extensive evolutionary remodelling of transcription initiation. For example, 43% (79,670 out of 184,476) of human TSSs could not be aligned to the mouse genome, and 39% (45,926 out of 116,277) of mouse TSSs could not be aligned to the human genome (Supplementary Methods). Alignment between species decayed as a function of neutral sequence divergence (Fig. 3). Housekeeping TSSs showed highest TSS conservation, whereas the TSSs of non-coding RNAs were less conserved than those of protein-coding TSSs. Indeed, the alignment of promoters of broadly expressed non-coding transcripts was not greatly different from randomly selected genomic sites (Fig. 3a). However, it is important to note that the random permutations inevitably overlap constrained elements, so cannot be considered representative of neutral evolution.

TSSs that were highly-restricted or biased in their expression to a single cell type or tissue were more likely to be gained or lost through evolution (Fig. 3a). TSSs preferentially expressed in fibroblasts, chondrocytes and pre-adipocytes were among the most conserved, whereas those enriched in T-cells, macrophages, dendritic cells, whole blood and endothelial cells were the most likely to be gained or lost (Fig. 3b). This suggests a more rapidly evolving immune system. It also suggests contributions of relaxed constraint and positive selection to the remodelling of transcription initiation through the insertion and deletion of promoter sequences.

To enable comparative analysis, we projected the expression patterns from one species to the other (Extended Data Fig. 4) and provide the peak position and orthologous expression profile through a cross-species track in ZENBU[10]. Only 54% and 61% of human and mouse conserved TSSs (of protein coding genes) had an orthologous peak in the other species. This increased to 61% and 63% respectively for TSSs from well matched samples (for example, human and mouse hepatocytes), however, surprisingly, almost 40% of conserved TSS do not appear to be used even in the matched cells (Supplementary Table 6).

Features of cell-type-specific promoters

Carrying out a systematic *de novo* motif discovery analysis in cell-type-specific promoters, recovered motifs similar to the binding motifs of transcription factors known to be relevant to the corresponding cellular states (Extended Data Fig. 5a–c and described in Supplementary Note 5). Examining general promoter features many CpG island (CGI) based promoters (54%) and most non-CGI-non-TATA promoters (92%) had non-ubiquitous expression profiles (Extended Data Fig. 3k–n). Although CGI promoters are generally associated with housekeeping

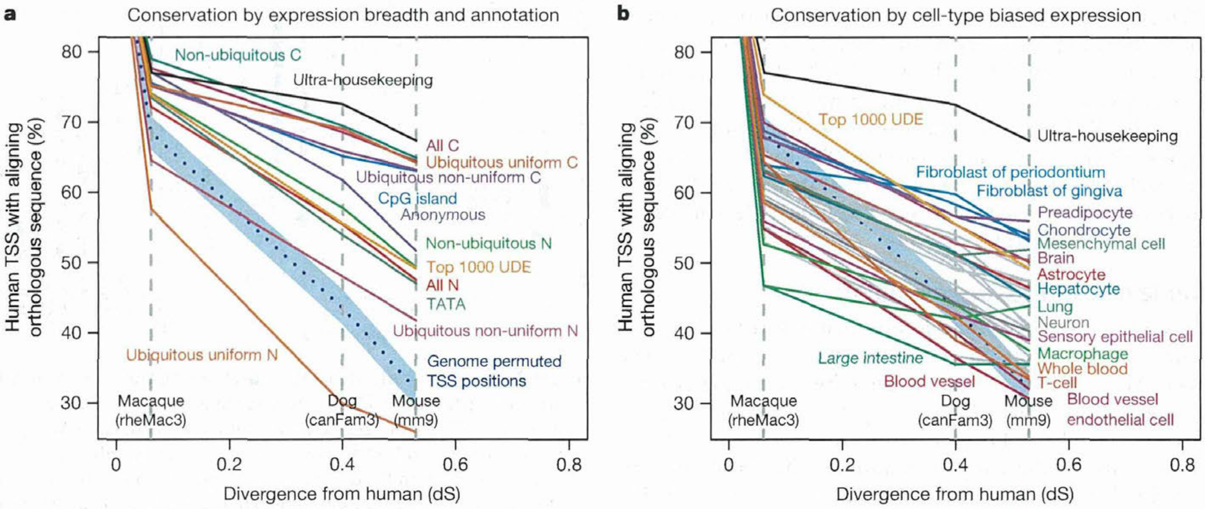

Figure 3 | TSS conservation as a function of expression properties and functional annotation. a, b, Human robust TSS coordinates were projected through EPO12 whole genome multiple sequence alignments (Supplementary Methods). The *y*-axis values show the fraction of human TSSs that align to an orthologous position in the indicated species. The *x* axis shows the relative divergence of macaque, dog and mouse genomes as the substitution rate at fourfold degenerate sites in protein coding sequence. The TSS locations were genome permuted (Supplementary Methods) and then projected through EPO12 alignments to give the null expectation (dashed blue line). The 95% confidence intervals of 1,000 samples of 1,000 TSS are shown (blue shading). **a**, TSS mapped to the 5′ ends of protein coding and non-coding transcripts are labelled (C and N, respectively), those that do not map to a known transcript 5′ end are shown as the 'anonymous' category. With the exception of anonymous, all robust TSSs represented in both panels are associated with the 5′ ends of previously annotated transcripts. Non-ubiquitous (cell-type-restricted), ubiquitous-uniform (housekeeping) and non-uniform-ubiquitous were defined as in Fig. 2. Ultra-housekeeping TSSs were defined as those with less than fivefold difference between maximum and median. The category top 1000 UDE represents the 1,000 ubiquitous TSSs that are most differentially expressed[4]. There are 1,016 ultra-housekeeping TSSs, 276 ubiquitous-uniform non-coding TSSs and all other categories contain over 2,000 TSSs. **b**, Same axes as panel **a** showing TSSs with expression that is biased towards a single expression facet (larger mutually exclusive grouping of the primary cell and tissue samples based on the sample ontologies CO and UBERON, defined in ref. 4). Only expression facets with greater than 250 enriched TSSs are shown. For clarity, only a subset of expression facets are coloured and labelled.

genes, we observed a subset with highly cell-type-restricted expression profiles (right tail of Extended Data Fig. 6a). Examining CGI and non-CGI promoters separately we find that cell-type-specific promoters of both classes were enriched for binding of cell-type-specific transcription factors (evidenced by over-representation of motifs and bound sites in public ChIP-seq data sets). For the human hepatocellular carcinoma cell line HepG2 we observed enrichment of liver-specific transcription factors (HNF4, FOXA2, and TCF7L2) at both CGI and non-CGI HepG2 specific promoters (Extended Data Fig. 6b, c; similar examples are shown in Extended Data Figs 5d and 7). As noted in the accompanying analysis[4], both cell-type-specific CGI and non-CGI promoters tend to have proximal high-specificity enhancers (Extended Data Fig. 6d). This indicates that specific expression at CGI promoters uses the same type of signals as non-CGI promoters: proximal transcription factor motifs and high-specificity enhancers.

Of note, a small number of highly abundant RNAs account for 20% or more of the reads in some libraries: HBB, SMR3B, STATH, PRB4, CLPS, HTN3, SERPINA1, CTRB2, CPB1, CPA1 and MALAT1. Although the abundance of these transcripts is a function of their relatively stability as well as rate of initiation, a modest but significant over representation of ETS and YY1 sites was found in highly expressed promoters compared to weakly expressed ones (Extended Data Fig. 5g). Although the different motif composition may contribute to expression levels, the accompanying manuscript[4] shows that arrays of enhancers with similar usage[20] probably contribute to the higher maximal expression rate.

Key cell-type-specific transcription factors

Among 1,762 human and 1,516 mouse transcription factors compiled from the literature[21–23], promoter level expression profiles for 1,665 human transcription factors (94%) and 1,382 mouse transcription factors (91%) were obtained (Supplementary Tables 7, 8 and 9 and Supplementary Note 6). The distribution of expression levels and cell-type or tissue-specificity of transcription factors (Extended Data Fig. 3f–j) and the number of robust promoter peaks per transcription factor gene was similar to coding genes in general (4.8 compared to 4.6). In any given primary cell type, a median of 430 (306 to 722) transcription factors were expressed at 10 TPM or above (~3 copies per cell based on 300,000 mRNAs per cell[18]) (Extended Data Fig. 3g).

Clustering transcription factors by expression profile revealed sets of transcription factors specifically enriched in each cell type (Extended Data Fig. 8). For each primary cell sample we have made available ranked lists of transcription factors based on their promoter expression in the sample relative to the median across the collection (http://fantom.gsc.riken.jp/5/sstar/Browse_samples). For most cell types we found one transcription factor that was very highly enriched (≥100-fold), 23 highly enriched transcription factors (≥ tenfold) and 82 moderately enriched transcription factors (≥ fivefold) (numbers of transcription factors are based on median number of transcription factors observed at each enrichment threshold across the primary cell samples). To demonstrate their likely relevance we systematically reviewed phenotypes of transcription factor knockout mice at the MGI (see Supplementary Note 7). The clear connection between tissue-specific expression profiles and relevant knockout phenotypes is summarized in Supplementary Table 10. For example, in mouse inner ear hair cells, knockout of six of the top 20 most enriched transcription factor genes in mouse (*Pou3f4* (ref. 24), *Sox2* (ref. 25), *Egr2*, *Six1* (ref. 26), *Fos*[27], *Tbx18* (ref. 28)) as well as patient mutations in a further four top transcription factor genes (*POU4F3* (ref. 29), *ZIC2* (ref. 30), *SOX10* (ref. 31), *FOXF2* (ref. 32)) resulted in hearing-related defects. Similarly, mouse knockouts or patients with mutations in the transcription factors enriched in osteoblasts (*CREB3L1* (ref. 33), *DLX5* (ref. 34), *EBF2* (ref. 35), *HAND2* (ref. 36), *HOXC5* (ref. 37), *NFIX*[38], *PRRX1* (ref. 39), *PRRX2* (ref. 40), *SIX1* (ref. 41), *TWIST1* (ref. 42), *SHOX*[43], *Six2* (ref. 44)) had bone and osteoblast phenotypes. A substantial fraction of top transcription factors (61% of mouse and 40% of human transcription factors) have relevant phenotypes recorded in knockout mice (Supplementary Table 10).

Inferring function from expression profiles

Taking a pair-wise Pearson correlation matrix of the promoter expression profiles we carried out MCL clustering[45] (Supplementary Methods) to group promoters that share similar expression profiles across the atlas. Figure 4 shows a graphical overview of the structure of the data (and the mouse counterpart is shown in Extended Data Fig. 9). We find 6,030 cases of named genes with alternative promoters participating in two or more coexpression clusters (Extended Data Fig. 10). To evaluate and annotate these coexpressed groups, we tested for enrichment in specific Gene Ontology terms and in a curated database of 489 biological pathways. Of these, 356 pathways (174 KEGG, 114 WikiPathways, 46 Reactome, 22 Netpath) were significantly enriched in at least one human coexpression group (FDR <0.05). Using this approach, 38% of the unannotated robust peaks (35,082 out of 91,269) were within a cluster with a significant association to a pathway. The annotated coexpression groups are summarized in the website (http://fantom.gsc.riken.jp/5/sstar/Browse_coexpression_clusters) and a detailed example identifying genes putatively involved in influenza A pathogenesis is shown in Extended Data Fig. 10a.

Introducing sample ontology enrichment analysis (SOEA), we show that expression profiles can also be associated with cell, anatomical and disease ontology terms by testing for overrepresentation of terms in ranked lists of systematically annotated samples expressing each peak (Extended Data Fig. 11 and Supplementary Methods). Novel peaks can be annotated in this way. For example, an un-annotated DPI peak at hg19::chr18:3659943..3659972,+ is linked to the terms classical monocyte (CL:0000860; P value $= 6.35 \times 10^{-124}$, Extended Data Fig. 11h) and bone marrow (UBERON:0002371; P value $= 2.7 \times 10^{-80}$). Manual examination of the profile confirms the transcript is predominantly expressed in myeloid cells with higher levels in CD14^{+} monocytes. Applied to all CAGE peaks, 127,645 human and 44,449 mouse robust peaks were annotated as enriched in at least one CL, DOID or UBERON term (Extended Data Fig. 11i, j). The most commonly-enriched terms at a P value threshold of 10^{-20} were classical monocyte (CL:0000860; 26,634 peaks, 14%), bone marrow (UBERON:0002371; 22,387 peaks, 12%) and neural tube (UBERON:0001049; 20,484 peaks, 11%) (Supplementary Table 13). This is consistent with the coexpression clustering in Fig. 4 (green and purple spheres correspond to leukocyte and central nervous system enriched expression profiles) and indicates that a large fraction of the mammalian genome is dedicated to immune and nervous system specific functions.

Conclusion

The FANTOM5 promoter atlas is a natural extension of earlier maps of active transcripts and promoters complementing the sequencing of mammalian genomes[46,47]. It represents an advance in an order of magnitude in the wide range of cell types and the amount of data produced per sample, and using single-molecule sequencing avoided polymerase chain reaction (PCR), digestion and cloning bias[48]. We have identified and quantified the activity of at least one promoter for more than 95% of annotated protein-coding genes in the human reference genome; only the activity of 1,225 promoters remains uncharacterized. Some of these may not actually be expressed. Some cannot be unambiguously measured with CAGE due to copy number variants or closely related multigene families. The remaining promoters are probably expressed in rare cell types or during windows of development or states of cellular activation that are not readily accessible and remain to be sampled. A continued effort to add profiles from these cells will make it possible to integrate them with the FANTOM5 data, and to extract metadata to identify those regulatory elements that are new and lineage-specific.

The FANTOM5 data highlights the value in profiling primary cells as opposed to whole tissues. It also highlights the weakness of using cancer cell lines. The cancer cell lines generally fail to cluster in a sample-to-sample correlation graph with their supposed cell type or tissue of origin (Extended Data Fig. 12) and express more transcription factors than primary cells (Extended Data Fig. 3g). The mutations and

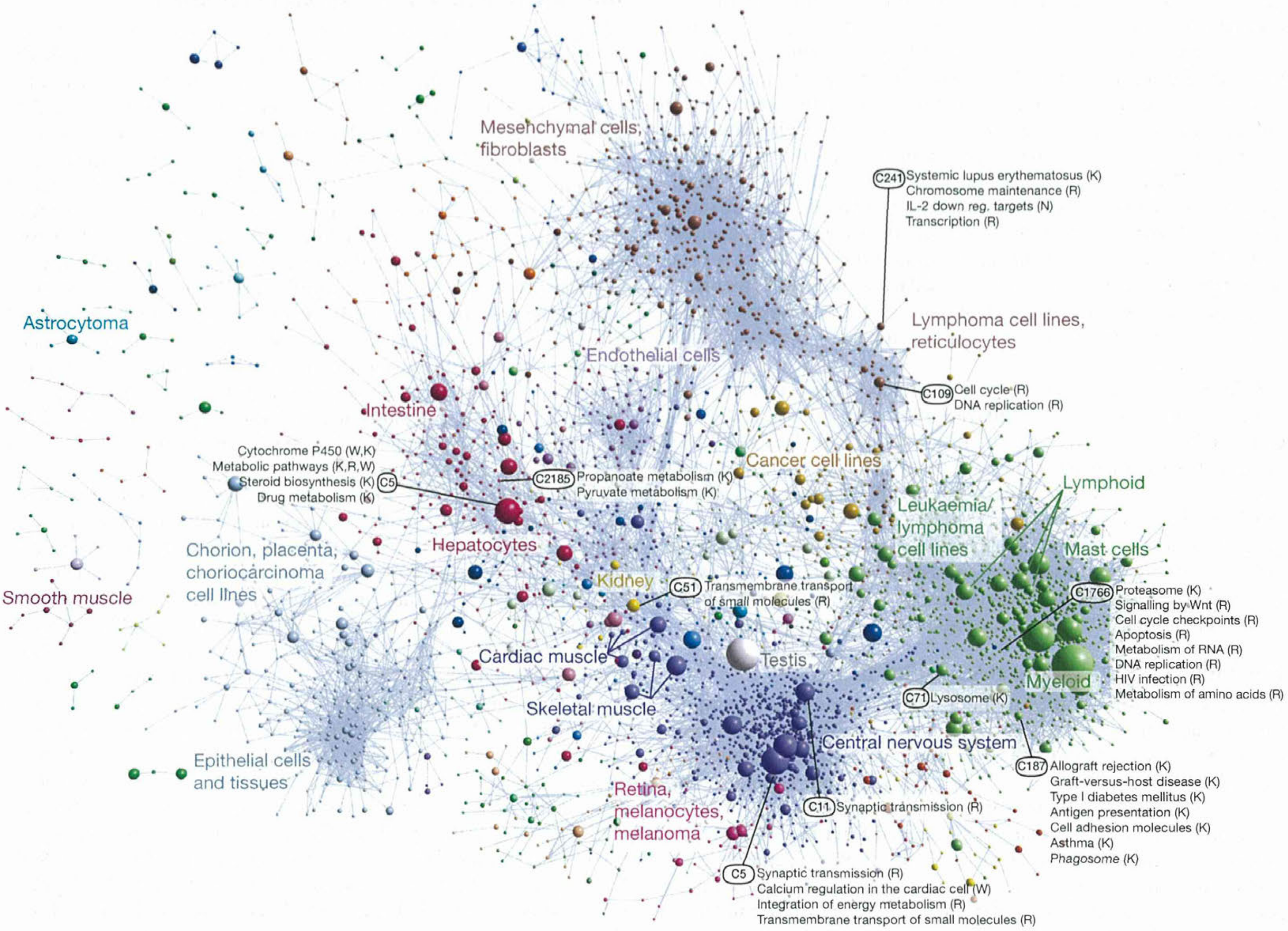
Mesenchymal cells, fibroblasts
C241 Systemic lupus erythematosus (K)
Chromosome maintenance (R)
IL-2 down reg. targets (N)
Transcription (R)
Astrocytoma
Lymphoma cell lines, reticulocytes
Endothelial cells
C109 Cell cycle (R)
DNA replication (R)
Intestine
Cytochrome P450 (W,K)
Metabolic pathways (K,R,W)
Steroid biosynthesis (K) C5
Drug metabolism (K)
C2185 Propanoate metabolism (K)
Pyruvate metabolism (K)
Cancer cell lines
Lymphoid
Leukaemia/ lymphoma cell lines
Chorion, placenta, choriocarcinoma cell lines
Hepatocytes
Mast cells
Smooth muscle
Kidney
C51 Transmembrane transport of small molecules (R)
C1766 Proteasome (K)
Signalling by Wnt (R)
Cell cycle checkpoints (R)
Apoptosis (R)
Metabolism of RNA (R)
DNA replication (R)
HIV infection (R)
Metabolism of amino acids (R)
Cardiac muscle
Testis
Myeloid
C71 Lysosome (K)
Skeletal muscle
Central nervous system
Epithelial cells and tissues
C187 Allograft rejection (K)
Graft-versus-host disease (K)
Type I diabetes mellitus (K)
Antigen presentation (K)
Cell adhesion molecules (K)
Asthma (K)
Phagosome (K)
Retina, melanocytes, melanoma
C11 Synaptic transmission (R)
C5 Synaptic transmission (R)
Calcium regulation in the cardiac cell (W)
Integration of energy metabolism (R)
Transmembrane transport of small molecules (R)

Figure 4 | Coexpression clustering of human promoters in FANTOM5. Collapsed coexpression network derived from 4,882 coexpression groups (one node is one group of promoters; 4,664 groups are shown here) derived from expression profiles of 124,090 promoters across all primary cell types, tissues and cell lines (visualized using Biolayout Express^{3D} (ref. 45), $r > 0.75$, MCLi = 2.2). For display, each group of promoters is collapsed into a sphere, the radius of which is proportional to the cube root of the number of promoters in that group. Edges indicate $r > 0.6$ between the average expression profiles of each cluster. Colours indicate loosely-associated collections of coexpression groups (MCLi = 1.2). Labels show representative descriptions of the dominant cell type in coexpression groups in each region of the network, and a selection of highly-enriched pathways (FDR $<10^{-4}$) from KEGG (K), WikiPathways (W), Netpath (N) and Reactome (R). Promoters and genes in the coexpression groups are available online at (http://fantom.gsc.riken.jp/5/data/).

chromosomal rearrangements that occur in cancer result in unique transcriptional networks that do not exist in the untransformed state and do not necessarily generalize across multiple tumours of the same type. In terms of building mammalian transcriptional regulatory network models that reflect the normal untransformed state, primary cells are the logical choice. They have normal genomes, and express in the order of 430 transcription factors at appreciable levels, ranking of which can be used to reduce the complexity further and identify key known regulators of cellular phenotypes. Focusing on these key regulators and motif searching in the corresponding cell-type-specific promoters provides the data to build cell-type-specific regulatory network models and support a rational approach to identification of drivers required to reprogram cells from one lineage to another. Promoter-based expression data also has direct practical applications in the interpretation (and re-interpretation) of the function of single nucleotide polymorphisms (SNPs) in genome-wide association studies (GWAS), which commonly occur in non-coding sequences. In accompanying manuscripts, reanalysis of several GWAS data sets uncovered new disease associations in FANTOM5 promoters and identification of regulatory SNPs within enhancers that were active in medically relevant samples (ref. 4 and manuscript in preparation). Accordingly, the data will enable the design of genotyping arrays and sequence-capture systems to target regulatory variation, and the design of promoter constructs allowing researchers to specify the cell-type-specificity and absolute expression levels of their constructs (particularly for Cre-conditional knockouts[49] and gene therapy vectors[50]). In all these respects, the FANTOM5 data set greatly extends the data generated by ENCODE[5] to further our knowledge of genome function.

METHODS SUMMARY

All Methods are described in full in the Supplementary Information.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 4 January 2013; accepted 26 February 2014.

1. Vickaryous, M. K. & Hall, B. K. Human cell type diversity, evolution, development, and classification with special reference to cells derived from the neural crest. *Biol. Rev. Camb. Philos. Soc.* **81,** 425–455 (2006).
2. Lenhard, B., Sandelin, A. & Carninci, P. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nature Rev. Genet.* **13,** 233–245 (2012).
3. Kanamori-Katayama, M. *et al.* Unamplified cap analysis of gene expression on a single-molecule sequencer. *Genome Res.* **21,** 1150–1159 (2011).

4. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* http://dx.doi.org/10.1038/nature12787 (this issue).
5. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).
6. Su, A. I. *et al.* A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA* **101,** 6062–6067 (2004).
7. Meehan, T. F. *et al.* Logical development of the cell ontology. *BMC Bioinformatics* **12,** 6 (2011).
8. 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).
9. Osborne, J. D. *et al.* Annotating the human genome with Disease Ontology. *BMC Genomics* **10** (Suppl 1), S6 (2009).
10. Severin, J. *et al.* Interactive visualization and analysis of large-scale NGS data-sets using ZENBU. *Nature Biotechnol.* http://dx.doi.org/10.1038/nbt.2840 (2014).
11. Oja, E., Hyvarinen, A. & Karhunen, J. *Independent Component Analysis* (John Wiley & Sons, 2001).
12. Affymetrix/Cold Spring Harbor Laboratory ENCODE Transcriptome Project. Post-transcriptional processing generates a diversity of 5′-modified long and short RNAs. *Nature* **457,** 1028–1032 (2009).
13. Carninci, P. *et al.* Genome-wide analysis of mammalian promoter architecture and evolution. *Nature Genet.* **38,** 626–635 (2006).
14. Ioshikhes, I., Hosid, S. & Pugh, B. F. Variety of genomic DNA patterns for nucleosome positioning. *Genome Res.* **21,** 1863–1871 (2011).
15. 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 (2010).
16. Schug, J. *et al.* Promoter features related to tissue specificity as measured by Shannon entropy. *Genome Biol.* **6,** R33 (2005).
17. Beissbarth, T. & Speed, T. P. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics* **20,** 1464–1465 (2004).
18. Velculescu, V. E. *et al.* Analysis of human transcriptomes. *Nature Genet.* **23,** 387–388 (1999).
19. Schmidt, D. *et al.* Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* **328,** 1036–1040 (2010).
20. Barolo, S. Shadow enhancers: frequently asked questions about distributed *cis*-regulatory information and enhancer redundancy. *Bioessays* **34,** 135–141 (2012).
21. Roach, J. C. *et al.* Transcription factor expression in lipopolysaccharide-activated peripheral-blood-derived mononuclear cells. *Proc. Natl Acad. Sci. USA* **104,** 16245–16250 (2007).
22. Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. & Luscombe, N. M. A census of human transcription factors: function, expression and evolution. *Nature Rev. Genet.* **10,** 252–263 (2009).
23. Wingender, E., Schoeps, T. & Dönitz, J. TFClass: an expandable hierarchical classification of human transcription factors. *Nucleic Acids Res.* **41,** D165–D170 (2013).
24. de Kok, Y. J. *et al.* Association between X-linked mixed deafness and mutations in the POU domain gene *POU3F4*. *Science* **267,** 685–688 (1995).
25. Kiernan, A. E. *et al.* *Sox2* is required for sensory organ development in the mammalian inner ear. *Nature* **434,** 1031–1035 (2005).
26. Zheng, W. *et al.* The role of *Six1* in mammalian auditory system development. *Development* **130,** 3989–4000 (2003).
27. Paylor, R., Johnson, R. S., Papaioannou, V., Spiegelman, B. M. & Wehner, J. M. Behavioral assessment of *c-fos* mutant mice. *Brain Res.* **651,** 275–282 (1994).
28. Trowe, M. O., Maier, H., Schweizer, M. & Kispert, A. Deafness in mice lacking the T-box transcription factor Tbx18 in otic fibrocytes. *Development* **135,** 1725–1734 (2008).
29. Vahava, O. *et al.* Mutation in transcription factor POU4F3 associated with inherited progressive hearing loss in humans. *Science* **279,** 1950–1954 (1998).
30. Chabchoub, E., Willekens, D., Vermeesch, J. R. & Fryns, J. P. Holoprosencephaly and *ZIC2* microdeletions: novel clinical and epidemiological specificities delineated. *Clin. Genet.* **81,** 584–589 (2012).
31. Pingault, V. *et al.* *SOX10* mutations in patients with Waardenburg-Hirschsprung disease. *Nature Genet.* **18,** 171–173 (1998).
32. Kapoor, S., Mukherjee, S. B., Shroff, D. & Arora, R. Dysmyelination of the cerebral white matter with microdeletion at 6p25. *Indian Pediatr.* **48,** 727–729 (2011).
33. Murakami, T. *et al.* Signalling mediated by the endoplasmic reticulum stress transducer OASIS is involved in bone formation. *Nature Cell Biol.* **11,** 1205–1211 (2009).
34. Acampora, D. *et al.* Craniofacial, vestibular and bone defects in mice lacking the *Distal-less*-related gene *Dlx5*. *Development* **126,** 3795–3809 (1999).
35. Kieslinger, M. *et al.* EBF2 regulates osteoblast-dependent differentiation of osteoclasts. *Dev. Cell* **9,** 757–767 (2005).
36. Funato, N. *et al.* Hand2 controls osteoblast differentiation in the branchial arch by inhibiting DNA binding of Runx2. *Development* **136,** 615–625 (2009).
37. McIntyre, D. C. *et al.* Hox patterning of the vertebrate rib cage. *Development* **134,** 2981–2989 (2007).
38. Driller, K. *et al.* Nuclear factor I X deficiency causes brain malformation and severe skeletal defects. *Mol. Cell. Biol.* **27,** 3855–3867 (2007).
39. Lu, M. F. *et al.* *prx-1* functions cooperatively with another paired-related homeobox gene, *prx-2*, to maintain cell fates within the craniofacial mesenchyme. *Development* **126,** 495–504 (1999).
40. Ten Berge, D., Brouwer, A., Korving, J., Martin, J. F. & Meijlink, F. *Prx1* and *Prx2* in skeletogenesis: roles in the craniofacial region, inner ear and limbs. *Development* **125,** 3831–3842 (1998).
41. Laclef, C. *et al.* Altered myogenesis in *Six1*-deficient mice. *Development* **130,** 2239–2252 (2003).
42. Lee, M. S., Lowe, G. N., Strong, D. D., Wergedal, J. E. & Glackin, C. A. *TWIST*, a basic helix-loop-helix transcription factor, can regulate the human osteogenic lineage. *J. Cell. Biochem.* **75,** 566–577 (1999).
43. Clement-Jones, M. *et al.* The short stature homeobox gene *SHOX* is involved in skeletal abnormalities in Turner syndrome. *Hum. Mol. Genet.* **9,** 695–702 (2000).
44. He, G. *et al.* Inactivation of *Six2* in mouse identifies a novel genetic mechanism controlling development and growth of the cranial base. *Dev. Biol.* **344,** 720–730 (2010).
45. Freeman, T. C. *et al.* Construction, visualisation, and clustering of transcription networks from microarray expression data. *PLoS Comput. Biol.* **3,** e206 (2007).
46. The FANTOM Consortium. The transcriptional landscape of the mammalian genome. *Science* **309,** 1559–1563 (2005).
47. Suzuki, H. *et al.* The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. *Nature Genet.* **41,** 553–562 (2009).
48. Kawaji, H. *et al.* Comparison of CAGE and RNA-seq transcriptome profiling using a clonally amplified and single molecule next generation sequencing. *Genome Res.* http://dx.doi.org/10.1101/gr.156232.113 (2014).
49. Heffner, C. S. *et al.* Supporting conditional mouse mutagenesis with a comprehensive cre characterization resource. *Nature Commun.* **3,** 1218 (2012).
50. Pringle, I. A. *et al.* Rapid identification of novel functional promoters for gene therapy. *J. Mol. Med.* **90,** 1487–1496 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements 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. It was also supported by Research Grants for RIKEN Preventive Medicine and Diagnosis Innovation Program (RIKEN PMI) to Y. Hayashizaki and RIKEN Centre for Life Science Technologies, Division of Genomic Technologies (RIKEN CLST (DGT)) from the MEXT, Japan. Extended acknowledgements are provided in the Supplementary Information.

Author Contributions The core members of FANTOM5 phase 1 were Alistair R. R. Forrest, Hideya Kawaji, Michael Rehli, J. Kenneth Baillie, Michiel J. L. de Hoon, Timo Lassmann, Masayoshi Itoh, Kim M. Summers, Harukazu Suzuki, Carsten O. Daub, Jun Kawai, Peter Heutink, Winston Hide, Tom C. Freeman, Boris Lenhard, Vladimir B. Bajic, Martin S. Taylor, Vsevolod J. Makeev, Albin Sandelin, David A. Hume, Piero Carninci and Yoshihide Hayashizaki. Samples were provided by: A. Blumenthal, A. Bonetti, A. Mackay-sim, A. Sajantila, A. Saxena, A. Schwegmann, A.G.B., A.J.K., A.L., A.R.R.F., A.S.B.E., B.B., C. Schmidl, C. Schneider, C.A.D., C.A.W., C.K., C.L.M., D.A.H., D.A.O., D.G., D.S., D.V., E.W., F.B., F.N., G.G.S., G.J.F., G.S., H. Kawamoto, H. Koseki, H. Morikawa, H. Motohashi, H. Ohno, H. Sato, H. Satoh, H. Tanaka, H. Tatsukawa, H. Toyoda, H.C.C., H.E., J. Kere, J.B., J.F., J.K.B., J.S.K., J.T., J.W.S., K.E., K.J.H., K.M., K.M.S., L.F., L.M.K., L.M.vdB., L.N.W., M. Edinger, M. Endoh, M. Fagiolini, M. Hamaguchi, M. Hara, M. Herlyn, M. Morimoto, M. Rehli, M. Yamamoto, M. Yoneda, M.B., M.C.F.C., M.D., M.E.F., M.O., M.O.H., M.P., M.vdW., N.M., N.O., N.T., P.A., P.G.Z., P.H., P.R., R.F., R.G., R.K.S., R.P., R.V., S. Guhl, S. Gustincich, S. Kojima, S. Koyasu, S. Krampitz, S. Sakaguchi, S. Savvi, S.E.Z., S.O., S.P.B., S.P.K., S. Roy., S.Z., T. Kitamura, T. Nakamura, T. Nozaki, T. Sugiyama, T.B.G., T.D., T.G., T.I., T.J.H., T.J.K., V.O., W.L., Y. Hasegawa, Y. Nakachi, Y. Nakamura, Y. Yamaguchi, Y. Yonekura, Y.I., Y.I.K., Y.M. and Y.O. Analyses were carried out by: A. Mathelier, A. Meynert, A. Sandelin, A.C., A.D.D., A.P.G., A.H., A.J., A.M.B., A.P., A.R.R.F., A.S.K., A.T.K., A.V.F., B. Lenhard, B. Lilje, B.D., B.K., B.M., B.R.J., C. Schmidl, C. Schneider, C.A.S., C.F., C.J.M., C.O.D., C.P., C.V.C., D.A., D.A.M., D.C., E. Dalla, E. Dimont, E.A., E.A.S., E.J.W., E.M., E.V., Ev.N., F.D., G.J., G.J.F., G.M.A., H. Kawaji, H. Ohmiya, H. Shimoji, H.F., H.J., H.P., I.A., I.E.V., I.H., I.V.K., J.A.B., J.A.C.A., J.A.R., J.C.M., J.F.J.L., J.G., J.G.D.P., J.H., J.K.B., J.S., K. Kajiyama, K.I., K.L., L.H., L.L., M. Francescatto, M. Rashid, M. Rehli, M. Roncador, M. Thompson, M.B.R., M.C., M.C.F., M.J., M.J.L.dH., M.L., M.S.T., M.V., N.B., O.J.L.R., O.M.H., P.A.C.tH., P.J.B, R.A., R.S.Y., S. Katayama, S. Kawaguchi, S. Schmeier, S. Rennie, S.F., S.J.H.S., S.P., T. Sengstag, T.C.F., T.F.M., T.H., T.K., T.L., T.R., T.T., U.S., V.B.B., V.H., V.J.M., W.H., W.W.W., X.Z., Y. Chen, Y. Ciani, Y.A.M., Y.S., Z.T. Libraries were generated by: A. Kaiho, A. Kubosaki, A. Saka, C. Simon, E.S., F.H., H.N., J. Kawai, K. Kaida, K.N., M. Furuno, M. Murata, M. Sakai, M. Tagami, M.I., M.K., M.K.K., N.K., N.N., N.S., P.C., R.M., S. Kato, S.N., S.N.-S., S.W., S.Y., T.A., T. Kawashima. The manuscript was written by A.R.R.F. and D.A.H. with help from A. Sandelin, J.K.B., M. Rehli, H.K., M.J.L.dH., V.H., I.V.K., M.T. and K.M.S. with contributions, edits and comments from all authors. The project was managed by Y. Hayashizaki, A.R.R.F., P.C., M.I., M.S., J. Kawai, C.O.D., H. Suzuki, T.L. and N.K. The scientific coordinator was A.R.R.F and the general organizer was Y. Hayashizaki.

Author Information All CAGE data has been deposited at DDBJ DRA under accession number DRA000991. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.R.R.F (alistair.forrest@gmail.com), P.C. (carninci@riken.jp) or Y.H. (yosihide@gsc.riken.jp).

The FANTOM Consortium and the RIKEN PMI and CLST (DGT)

Alistair R. R. Forrest[1,2]*, Hideya Kawaji[1,2,3]*, Michael Rehli[4,5]*, J. Kenneth Baillie[6]*, Michiel J. L. de Hoon[1,2], Vanja Haberle[7,8], Timo Lassmann[1,2], Ivan V. Kulakovskiy[9,10], Marina Lizio[1,2], Masayoshi Itoh[1,2,3], Robin Andersson[11], Christopher J. Mungall[12], Terrence F. Meehan[13], Sebastian Schmeier[14,15], Nicolas Bertin[1,2], Mette Jørgensen[11], Emmanuel Dimont[16], Erik Arner[1,2], Christian Schmidl[4]†, Ulf Schaefer[14], Yulia A. Medvedeva[10,14]†, Charles Plessy[1,2], Morana Vitezic[1,17], Jessica Severin[1,2], Colin A. Semple[18], Yuri Ishizu[1,2], Robert S. Young[18], Margherita Francescatto[19,20], Intikhab Alam[14], Davide Albanese[21], Gabriel M. Altschuler[16], Takahiro Arakawa[1,2], John A. C.

Archer[14], Peter Arner[22], Magda Babina[23], Sarah Rennie[18], Piotr J. Balwierz[24], Anthony G. Beckhouse[25,26], Swati Pradhan-Bhatt[27], Judith A. Blake[28], Antje Blumenthal[26,29], Beatrice Bodega[30], Alessandro Bonetti[1,2], James Briggs[25]†, Frank Brombacher[31,32], A. Maxwell Burroughs[1], Andrea Califano[33,34,35,36], Carlo V. Cannistraci[37,38]†, Daniel Carbajo[39], Yun Chen[11], Marco Chierici[21], Yari Ciani[40], Hans C. Clevers[41,42,43], Emiliano Dalla[40], Carrie A. Davis[44], Michael Detmar[45], Alexander D. Diehl[46], Taeko Dohi[47], Finn Drabløs[48], Albert S. B. Edge[49], Matthias Edinger[4,5], Karl Ekwall[50], Mitsuhiro Endoh[51,52], Hideki Enomoto[53], Michela Fagiolini[54], Lynsey Fairbairn[6], Hai Fang[55], Mary C. Farach-Carson[56], Geoffrey J. Faulkner[57], Alexander V. Favorov[10,58,59], Malcolm E. Fisher[6], Martin C. Frith[60], Rie Fujita[61], Shiro Fukuda[1], Cesare Furlanello[21], Masaaki Furuno[1,2], Jun-ichi Furusawa[51,52,62], Teunis B. Geijtenbeek[63], Andrew P. Gibson[64], Thomas Gingeras[44], Daniel Goldowitz[65], Julian Gough[55], Sven Guhl[23], Reto Guler[31,32], Stefano Gustincich[66], Thomas J. Ha[65], Masahide Hamaguchi[67], Mitsuko Hara[68], Matthias Harbers[1], Jayson Harshbarger[1,2], Akira Hasegawa[1,2], Yuki Hasegawa[1,2], Takehiro Hashimoto[1], Meenhard Herlyn[69], Kelly J. Hitchens[25,26], Shannan J. Ho Sui[16], Oliver M. Hofmann[16], Ilka Hoof[11], Fumi Hori[1,2], Lukasz Huminiecki[17], Kei Iida[70], Tomokatsu Ikawa[51,52], Boris R. Jankovic[14], Hui Jia[71], Anagha Joshi[6], Giuseppe Jurman[21], Bogumil Kaczkowski[1,2], Chieko Kai[72], Kaoru Kaida[1,2], Ai Kaiho[1], Kazuhiro Kajiyama[1,2], Mutsumi Kanamori-Katayama[1], Artem S. Kasianov[10], Takeya Kasukawa[2], Shintaro Katayama[1], Sachi Kato[1,2], Shuji Kawaguchi[70], Hiroshi Kawamoto[51], Yuki I. Kawamura[47], Tsugumi Kawashima[1,2], Judith S. Kempfle[49], Tony J. Kenna[29], Juha Kere[50,73], Levon M. Khachigian[74], Toshio Kitamura[75], S. Peter Klinken[76], Alan J. Knox[77], Miki Kojima[1,2], Soichi Kojima[68], Naoto Kondo[1,2], Haruhiko Koseki[51,52], Shigeo Koyasu[51,52,62], Sarah Krampitz[45], Atsutaka Kubosaki[1], Andrew T. Kwon[1,2], Jeroen F. J. Laros[64], Weonju Lee[78], Andreas Lennartsson[50], Kang Li[11], Berit Lilje[11], Leonard Lipovich[71], Alan Mackay-sim[79], Ri-ichiroh Manabe[1,2], Jessica C. Mar[39], Benoit Marchand[14], Anthony Mathelier[65], Niklas Mejhert[22], Alison Meynert[18], Yosuke Mizuno[80], David A. de Lima Morais[81], Hiromasa Morikawa[67], Mitsuru Morimoto[53], Kazuyo Moro[51,52,62,82], Efthymios Motakis[1,2], Hozumi Motohashi[83], Christine L. Mummery[84], Mitsuyoshi Murata[1,2], Sayaka Nagao-Sato[1], Yutaka Nakachi[80,85], Fumio Nakahara[75], Toshiyuki Nakamura[72], Yukio Nakamura[86], Kenichi Nakazato[1], Erik van Nimwegen[24], Noriko Ninomiya[1], Hiromi Nishiyori[1,2], Shohei Noma[1,2], Tadasuke Nozaki[87], Soichi Ogishima[88]†, Naganari Ohkura[67], Hiroko Ohmiya[1,2]†, Hiroshi Ohno[51,52], Mitsuhiro Ohshima[89], Mariko Okada-Hatakeyama[51,52], Yasushi Okazaki[80,85], Valerio Orlando[30,37], Dmitry A. Ovchinnikov[25], Arnab Pain[14,37], Robert Passier[84], Margaret Patrikakis[74], Helena Persson[50], Silvano Piazza[40], James G. D. Prendergast[18], Owen J. L. Rackham[55], Jordan A. Ramilowski[1,2], Mamoon Rashid[14,37], Timothy Ravasi[37,38], Patrizia Rizzu[19], Marco Roncador[21], Sugata Roy[1,2], Morten B. Rye[48], Eri Saijyo[1], Antti Sajantila[90], Akiko Saka[1], Shimon Sakaguchi[67], Mizuho Sakai[1,2], Hiroki Sato[72], Hironori Satoh[61], Suzana Savvi[31,32], Alka Saxena[1]†, Claudio Schneider[40,91], Erik A. Schultes[64], Gundula G. Schulze-Tanzil[92], Anita Schwegmann[31,32], Thierry Sengstag[1], Guojun Sheng[53], Hisashi Shimoji[1], Yishai Shimoni[36], Jay W. Shin[1,2], Christophe Simon[1,2], Daisuke Sugiyama[93], Takaaki Sugiyama[72], Masanori Suzuki[1], Naoko Suzuki[1,2], Rolf K. Swoboda[69], Peter A. C. 't Hoen[64], Michihira Tagami[1,2], Naoko Takahashi[1,2], Jun Takai[61], Hiroshi Tanaka[88], Hideki Tatsukawa[94], Zuotian Tatum[64], Mark Thompson[64], Hiroo Toyoda[87], Tetsuro Toyoda[70], Eivind Valen[95], Marc van de Wetering[41], Linda M. van den Berg[63], Roberto Verardo[40], Dipti Vijayan[25,26], Ilya E. Vorontsov[10], Wyeth W. Wasserman[65], Shoko Watanabe[1], Christine A. Wells[25,26], Louise N. Winteringham[76], Ernst Wolvetang[25], Emily J. Wood[71], Yoko Yamaguchi[96], Masayuki Yamamoto[61], Misako Yoneda[72], Yohei Yonekura[53], Shigehiro Yoshida[1,2], Susan E. Zabierowski[69], Peter G. Zhang[65], Xiaobei Zhao[11], Silvia Zucchelli[66], Kim M. Summers[6], Harukazu Suzuki[1,2], Carsten O. Daub[1], Jun Kawai[1,3], Peter Heutink[19], Winston Hide[16], Tom C. Freeman[6], Boris Lenhard[8,97], Vladimir B. Bajic[14], Martin S. Taylor[18], Vsevolod J. Makeev[9,10,98], Albin Sandelin[11], David A. Hume[6], Piero Carninci[1,2], Yoshihide Hayashizaki[1,3]

[1]RIKEN Omics Science Center (OSC), 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama 230-0045, Japan. [2]RIKEN Center for Life Science Technologies (Division of Genomic Technologies) (CLST (DGT)), 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan. [3]RIKEN Preventive Medicine and Diagnosis Innovation Program (PMI), 2-1 Hirosawa, Wako-shi, Saitama 351-0198, Japan. [4]Department of Internal Medicine III, University Hospital Regensburg, F.-J.-Strauss Allee 11, D-93042 Regensburg, Germany. [5]Regensburg Centre for Interventional Immunology (RCI), D-93042 Regensburg, Germany. [6]The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush, Edinburgh, Midlothian EH25 9RG, UK. [7]Department of Biology, University of Bergen, Thormøhlensgate 53, NO-5006 Bergen, Norway. [8]Faculty of Medicine, Institute of Clinical Sciences, MRC Clinical Sciences Centre, Imperial College London, Hammersmith Hospital Campus, London W12 0NN, UK. [9]Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Vavilov str. 32, Moscow 119991, Russia. [10]Vavilov Institute of General Genetics, Russian Academy of Sciences, Gubkin str. 3, Moscow 119991, Russia. [11]The Bioinformatics Centre, Department of Biology and BRIC, University of Copenhagen, Ole Maaloes Vej 5, DK 2200 Copenhagen, Denmark. [12]Genomics Division, Lawrence Berkeley National Laboratory, 84R01, 1 Cyclotron Road, Berkeley, California 94720, USA. [13]Mouse Informatics, European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. [14]Computational Bioscience Research Center, King Abdullah University of Science and Technology (KAUST), Ibn Al-Haytham Building -2, Thuwal 23955-6900, Kingdom of Saudi Arabia. [15]Institute of Natural and Mathematical Sciences, Massey University, Private Bag 102-904, North Shore Mail Centre, 0745 Auckland, New Zealand. [16]Department of Biostatistics, Harvard School of Public Health, 655 Huntington Ave, Boston, Massachusetts 02115, USA. [17]Department of Cell and Molecular Biology, Karolinska Institutet, P.O. Box 285, SE-171 77 Stockholm, Sweden. [18]MRC Human Genetics Unit, MRC Institute of Genetics and Molecular Medicine (MRC-IGMM), University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh EH4 2XU, UK. [19]Department of Clinical Genetics, VU University Medical Center Amsterdam, Van der Boechorststraat 7, 1081 BT Amsterdam, The Netherlands. [20]Graduate Program in Areas of Basic and Applied Biology, Abel Salazar Biomedical Sciences Institute, University of Porto, Rua de Jorge Viterbo Ferreira n. 228, 4050-313 Porto, Portugal. [21]Predictive Models for Biomedicine and Environment, Fondazione Bruno Kessler, via Sommarive 18, 38123 Trento, Italy. [22]Department of Medicine, Karolinska Institutet at Karolinska University Hospital, Huddinge, SE-141 86 Huddinge, Sweden. [23]Department of Dermatology and Allergy, Charité Campus Mitte, Universitätsmedizin Berlin, Chariteplatz 1, 10117 Berlin, Germany. [24]Biozentrum, University of Basel, Klingelbergstrasse 50-70, 4056 Basel, Switzerland. [25]Australian Institute for Bioengineering and Nanotechnology (AIBN), University of Queensland, Brisbane St Lucia, Queensland 4072, Australia. [26]Australian Infectious Diseases Research Centre (AID), University of Queensland, Brisbane St Lucia, Queensland 4072, Australia. [27]Department of Biological Sciences, University of Delaware, Newark, Delaware 19713, USA. [28]Bioinformatics and Computational Biology, The Jackson Laboratory, 600 Main Street, Bar Harbor, Maine 04609, USA. [29]Diamantina Institute, University of Queensland, Brisbane St Lucia, Queensland 4072, Australia. [30]IRCCS Fondazione Santa Lucia, via del Fosso di Fiorano 64, 00143 Rome, Italy. [31]Immunology and Infectious Disease, International Centre for Genetic Engineering & Biotechnology (ICGEB) Cape Town component, Anzio Road, Observatory 7925, Cape Town, South Africa. [32]Division of Immunology, Institute of Infectious Diseases and Molecular Medicine (IDM), University of Cape Town, Anzio Road, Observatory 7925, Cape Town, South Africa. [33]Department of Systems Biology, Columbia University Medical Center, 1130 St. Nicholas Avenue, New York, New York 10032, USA. [34]Department of Biochemistry and Molecular Biophysics, Columbia University Medical Center, 701 West 168th Street, New York, New York 10032, USA. [35]Department of Biomedical Informatics, Columbia University Medical Center, 622 West 168th Street, VC5, New York, New York 10032, USA. [36]Institute of Cancer Genetics, Columbia University Medical Center, Herbert Irving Comprehensive Cancer Center, 1130 St. Nicholas Avenue, New York, New York 10032, USA. [37]Biological and Environmental Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Ibn Al-Haytham Building -2, Thuwal 23955-6900, Kingdom of Saudi Arabia. [38]Applied Mathematics and Computational Science Program, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Kingdom of Saudi Arabia. [39]Department of Systems and Computational Biology, Albert Einstein College of Medicine, The Bronx, New York, New York 10461, USA. [40]Laboratorio Nazionale del Consorzio Interuniversitario per le Biotecnologie (LNCIB), Padriciano 99, 34149 Trieste, Italy. [41]Hubrecht Institute, Uppsalalaan 8, 3584 CT Utrecht, The Netherlands. [42]The Royal Netherlands Academy of Arts and Sciences, P.O. Box 19121, NL-1000 GC Amsterdam, The Netherlands. [43]University Medical Centre Utrecht, Postbus 85500, 3508 GA Utrecht, The Netherlands. [44]Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, New York 11797, USA. [45]Institute of Pharmaceutical Sciences, ETH Zurich, Vladimir-Prelog-Weg 3, HCI H 303, 8093 Zurich, Switzerland. [46]Department of Neurology, University at Buffalo School of Medicine and Biomedical Sciences, New York State Center of Excellence in Bioinformatics and Life Sciences, 701 Ellicott Street, Buffalo, New York 14203, USA. [47]Gastroenterology, Research Center for Hepatitis and Immunology Research Institute, National Center for Global Health and Medicine, 1-7-1 Kohnodai, Ichikawa, Chiba 272-8516, Japan. [48]Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology (NTNU), P.O. Box 8905, NO-7491 Trondheim, Norway. [49]Department of Otology and Laryngology, Harvard *Medical School, Massachusetts Eye and Ear Infirmary,* Eaton-Peabody Lab, 243 Charles Street, Boston, Massachusetts 02114, USA. [50]Department of Biosciences and Nutrition, Center for Biosciences, Karolinska Institutet, Hälsovägen 7-9, SE-141 83 Huddinge, Sweden. [51]RIKEN Research Center for Allergy and Immunology (RCAI), 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan. [52]RIKEN Center for Integrative Medical Sciences, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan. [53]RIKEN Center for Developmental Biology (CDB), 2-2-3 Minatojima-minamimachi, Chuo-ku, Kobe, Hyogo 650-0047, Japan. [54]FM Kirby Neurobiology Center, Children's Hospital Boston, Harvard Medical School, 300 Longwood Avenue, Boston, Massachusetts 02115, USA. [55]Department of Computer Science, University of Bristol, Merchant Venturers Building, Woodland Road, Bristol BS8 1UB, UK. [56]Department of Biochemistry and Cell Biology, Rice University, Houston, Texas 77251-1892, USA. [57]Cancer Biology Program, Mater Medical Research Institute, Raymond Terrace, South Brisbane, Queensland 4101, Australia. [58]Department of Oncology, Division of Oncology, Biostatistics and Bioinformatics, Johns Hopkins University School of Medicine, 550 North Broadway, Baltimore, Maryland 21205, USA. [59]State Research Institute of Genetics and Selection of Industrial Microorganisms GosNIIgenetika, 1-st Dorozhniy pr., 1, 117545 Moscow, Russia. [60]Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology (AIST), 2-4-7 Aomi, Koto-ku, Tokyo 135-0064, Japan. [61]Department of Medical Biochemistry, Tohoku University Graduate School of Medicine, 2-1 Seiryo-machi, Aoba-ku, Sendai, Miyagi 980-8575, Japan. [62]Department of Microbiology and Immunology, Keio University School of Medicine, 35 Shinanomachi, Shinjuku, Tokyo 160-8582, Japan. [63]Experimental Immunology, Academic Medical Center, University of Amsterdam, Meibergdreef 9, 1105 AZ Amsterdam, The Netherlands. [64]Department of Human Genetics, Leiden University Medical Center, Einthovenweg 20, 2333 ZC Leiden, The Netherlands. [65]Department of Medical Genetics, Centre for Molecular Medicine and Therapeutics, Child and Family Research Institute, University of British Columbia, 950 West 28th Avenue, Vancouver, British Columbia V5Z 4H4, Canada. [66]Neuroscience, SISSA, via Bonomea 265, 34136 Trieste, Italy. [67]Experimental Immunology, Immunology Frontier Research Center, Osaka University, 3-1 Yamadaoka, Suita, Osaka 565-0871, Japan. [68]RIKEN Advanced Science Institute (ASI), 2-1 Hirosawa, Wako, Saitama 351-0198, Japan. [69]Melanoma Research Center, The Wistar Institute, 3601 Spruce Street, Philadelphia, Pennsylvania 19104, USA. [70]RIKEN Bioinformatics And Systems Engineering Division (BASE), 1-7-22 Suehiro, Tsurumi, Yokohama, Kanagawa 230-0045, Japan. [71]Center for Molecular Medicine and Genetics, Wayne State University, 3228 Scott Hall, 540 East Canfield Street, Detroit, Michigan 48201-1928, USA. [72]Laboratory Animal Research Center, Institute of Medical Science, The University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan. [73]Science for Life Laboratory, Box 1031, SE-171 21 Solna, Sweden. [74]Centre for Vascular Research, University of New South

Wales, Sydney, New South Wales 2052, Australia. [75]Division of Cellular Therapy and Division of Stem Cell Signaling, Institute of Medical Science, University of Tokyo, Tokyo 108-8639, Japan. [76]Harry Perkins Institute of Medical Research, and the Centre for Medical Research, University of Western Australia, QQ Block, QEII Medical Centre, Nedlands, Perth, Western Australia 6009, Australia. [77]Respiratory Medicine, University of Nottingham, Clinical Sciences Building, City Hospital, Hucknall Road, Nottingham NG5 1PB, UK. [78]Department of Dermatology, Kyungpook National University School of Medicine, 130 Dongdeok-ro Jung-gu, Daegu 700-721, South Korea. [79]National Centre for Adult Stem Cell Research, Eskitis Institute for Cell and Molecular Therapies, Griffith University, Brisbane, Queensland 4111, Australia. [80]Division of Functional Genomics and Systems Medicine, Research Center for Genomic Medicine, Saitama Medical University, 1397-1 Yamane, Hidaka, Saitama 350-1241, Japan. [81]Faculty of Engineering, University of Bristol, Merchant Venturers Building, Woodland Road, Clifton BS8 1UB, UK. [82]PRESTO, Japanese Science and Technology Agency (JST), 7 Gobancho, Chiyodaku, Tokyo 102-0076, Japan. [83]Center for Radioisotope Sciences, Tohoku University Graduate School of Medicine, 2-1 Seiryo-machi, Aoba-ku, Sendai, Miyagi 980-8575, Japan. [84]Anatomy and Embryology, Leiden University Medical Center, Einthovenweg 20, P.O. Box 9600, 2300 RC Leiden, The Netherlands. [85]Division of Translational Research, Research Center for Genomic Medicine, Saitama Medical University, 1397-1 Yamane, Hidaka, Saitama 350-1241, Japan. [86]RIKEN BioResource Center (BRC), Koyadai 3-1-1, Tsukuba, Ibaraki 305-0074, Japan. [87]Department of Clinical Molecular Genetics, School of Pharmacy, Tokyo University of Pharmacy and Life Sciences, 1432-1 Horinouchi, Hachioji, Tokyo 192-0392, Japan. [88]Department of Bioinformatics, Medical Research Institute, Tokyo Medical and Dental University, 1-5-45 Yushima, Bunkyo-ku, Tokyo 113-8510, Japan. [89]Department of Biochemistry, Ohu University School of Pharmaceutical Sciences, Misumido 31-1, Tomitamachi, Koriyama, Fukushima 963-8611, Japan. [90]Hjelt Institute, Department of Forensic Medicine, University of Helsinki, Kytosuontie 11, 003000 Helsinki, Finland. [91]DSMB Dipartimento Scienze Mediche e Biologiche University of Udine, P.le Kolbe 3, 33100 Udine, Italy. [92]Department of Orthopedic, Trauma and Reconstructive Surgery, Charité Universitätsmedizin Berlin, Garystrasse 5, 14195 Berlin, Germany. [93]Center for Clinical and Translational Reseach, Kyushu University Hospital, Station for Collaborative Research1 4F, 3-1-1 Maidashi, Higashi-Ku, Fukuoka 812-8582, Japan. [94]Graduate School of Pharmaceutical Sciences, Nagoya University, Furo-cho, Chikusa, Nagoya, Aichi 464-8601, Japan. [95]Department of Molecular and Cellular Biology, Harvard University, 16 Divinity Avenue, Cambridge, Massachusetts 02138, USA. [96]Department of Biochemistry, Nihon University School of Dentistry, 1-8-13, Kanda-Surugadai, Chiyoda-ku, Tokyo 101-8310, Japan. [97]Department of Informatics, University of Bergen, Høgteknologisenteret, Thormøhlensgate 53, NO-5008 Bergen, Norway. [98]Department of Biological and Medical Physics, Moscow Institute of Physics and Technology (MIPT) 9, Institutsky Per., Dolgoprudny, Moscow Region 141700, Russia.

†Present addresses: Institute of Predictive and Personalized Medicine of Cancer, Ctra. de Can Roti, cami de les escoles, s/n, 08916 Badalona (Barcelona), Spain (Y.A.M.); Biomedical Cybernetics Group, Biotechnology Center (BIOTEC), Technische Universität Dresden, Dresden, Germany (C.V.C.); Genomics Core Facility, Biomedical Research Centre, Guy's Hospital, London SE1 9RT, UK (A. Saxena); RIKEN Advanced Center for Computing and Communication (ACCC), 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa, 230-0045 Japan (H. Ohmiya); Research Center for Molecular Medicine of the Austrian Academy of Sciences (CeMM), 1090 Vienna, Austria (C. Schmidl); Department of Biological and Biomedical Sciences, Harvard University, Cambridge, Massachusetts 02138, USA (J.B.); Department of Bioclinical Informatics,Tohoku Medical Megabank Organization,Tohoku University. Sendai 980-8573, Japan (S.O.).

*These authors contributed equally to this work.

doi:10.1038/nature12787

An atlas of active enhancers across human cell types and tissues

Robin Andersson[1]*, Claudia Gebhard[2,3]*, Irene Miguel-Escalada[4], Ilka Hoof[1], Jette Bornholdt[1], Mette Boyd[1], Yun Chen[1], Xiaobei Zhao[1,5], Christian Schmidl[2], Takahiro Suzuki[6,7], Evgenia Ntini[8], Erik Arner[6,7], Eivind Valen[1,9], Kang Li[1], Lucia Schwarzfischer[2], Dagmar Glatz[2], Johanna Raithel[2], Berit Lilje[1], Nicolas Rapin[1,10], Frederik Otzen Bagger[1,10], Mette Jørgensen[1], Peter Refsing Andersen[8], Nicolas Bertin[6,7], Owen Rackham[6,7], A. Maxwell Burroughs[6,7], J. Kenneth Baillie[11], Yuri Ishizu[6,7], Yuri Shimizu[6,7], Erina Furuhata[6,7], Shiori Maeda[6,7], Yutaka Negishi[6,7], Christopher J. Mungall[12], Terrence F. Meehan[13], Timo Lassmann[6,7], Masayoshi Itoh[6,7,14], Hideya Kawaji[6,14], Naoto Kondo[6,14], Jun Kawai[6,14], Andreas Lennartsson[15], Carsten O. Daub[6,7,15], Peter Heutink[16], David A. Hume[11], Torben Heick Jensen[8], Harukazu Suzuki[6,7], Yoshihide Hayashizaki[6,14], Ferenc Müller[4], The FANTOM Consortium†, Alistair R. R. Forrest[6,7], Piero Carninci[6,7], Michael Rehli[2,3] & Albin Sandelin[1]

Enhancers control the correct temporal and cell-type-specific activation of gene expression in multicellular eukaryotes. Knowing their properties, regulatory activity and targets is crucial to understand the regulation of differentiation and homeostasis. Here we use the FANTOM5 panel of samples, covering the majority of human tissues and cell types, to produce an atlas of active, *in vivo*-transcribed enhancers. We show that enhancers share properties with CpG-poor messenger RNA promoters but produce bidirectional, exosome-sensitive, relatively short unspliced RNAs, the generation of which is strongly related to enhancer activity. The atlas is used to compare regulatory programs between different cells at unprecedented depth, to identify disease-associated regulatory single nucleotide polymorphisms, and to classify cell-type-specific and ubiquitous enhancers. We further explore the utility of enhancer redundancy, which explains gene expression strength rather than expression patterns. The online FANTOM5 enhancer atlas represents a unique resource for studies on cell-type-specific enhancers and gene regulation.

Precise regulation of gene expression in time and space is required for development, differentiation and homeostasis[1]. Sequence elements within or near core promoter regions contribute to regulation[2], but promoter-distal regulatory regions like enhancers are essential in the control of cell-type specificity[1]. Enhancers were originally defined as remote elements that increase transcription independently of their orientation, position and distance to a promoter[3]. They were only recently found to initiate RNA polymerase II (RNAPII) transcription, producing so-called eRNAs[4]. Genomic locations of enhancers can be detected by mapping of chromatin marks and transcription factor binding sites from chromatin immunoprecipitation (ChIP) assays and DNase I hypersensitive sites (DHSs) (reviewed in ref. 1), but there has been no systematic analysis of enhancer usage in the large variety of cell types and tissues present in the human body.

Using cap analysis of gene expression[5] (CAGE), we show that enhancer activity can be detected through the presence of balanced bidirectional capped transcripts, enabling the identification of enhancers from small primary cell populations. Based upon the FANTOM5 CAGE expression atlas encompassing 432 primary cell, 135 tissue and 241 cell line samples from human[6], we identify 43,011 enhancer candidates and characterize their activity across the majority of human cell types and tissues. The resulting catalogue of transcribed enhancers enables classification of ubiquitous and cell-type-specific enhancers, modelling of physical interactions between multiple enhancers and TSSs, and identification of potential disease-associated regulatory single nucleotide polymorphisms (SNPs).

Bidirectional capped RNAs identify active enhancers

The FANTOM5 project has generated a CAGE-based transcription start site (TSS) atlas across a broad panel of primary cells, tissues and cell lines covering the vast majority of human cell types[6]. Within that data set, well-studied enhancers often have CAGE peaks delineating nucleosome-deficient regions (NDRs) (Supplementary Fig. 1). To determine whether this is a general enhancer feature, FANTOM5 CAGE (Supplementary Table 1) was superimposed on active (H3K27ac-marked) enhancers defined by HeLa-S3 ENCODE ChIP-seq data[7]. CAGE tags showed a bimodal distribution flanking the central P300 peak, with divergent transcription from the enhancer (Fig. 1a). Similar patterns

[1]The Bioinformatics Centre, Department of Biology & Biotech Research and Innovation Centre, University of Copenhagen, Ole Maaloes Vej 5, DK-2200 Copenhagen, Denmark. [2]Department of Internal Medicine III, University Hospital Regensburg, Franz-Josef-Strauss-Allee 11, 93042 Regensburg, Germany. [3]Regensburg Centre for Interventional Immunology (RCI), D-93042 Regensburg, Germany. [4]School of Clinical and Experimental Medicine, College of Medical and Dental Sciences, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK. [5]Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, North Carolina 27599, USA. [6]RIKEN OMICS Science Centre, RIKEN Yokohama Institute, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama City, Kanagawa 230-0045, Japan. [7]RIKEN Center for Life Science Technologies (Division of Genomic Technologies), RIKEN Yokohama Institute, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama City, Kanagawa 230-0045, Japan. [8]Centre for mRNP Biogenesis and Metabolism, Department of Molecular Biology and Genetics, C.F. Møllers Alle 3, Building 1130, DK-8000 Aarhus, Denmark. [9]Department of Molecular and Cellular Biology, Harvard University, Cambridge, Massachusetts 02138, USA. [10]The Finsen Laboratory, Rigshospitalet and Danish Stem Cell Centre (DanStem), University of Copenhagen, Ole Maaloes Vej 5, DK-2200, Denmark. [11]Roslin Institute, Edinburgh University, Easter Bush, Midlothian, Edinburgh EH25 9RG, UK. [12]Genomics Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Road MS 64-121, Berkeley, California 94720, USA. [13]EMBL Outstation - Hinxton, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. [14]RIKEN Preventive Medicine and Diagnosis Innovation Program, RIKEN Yokohama Institute, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama City, Kanagawa 230-0045, Japan. [15]Department of Biosciences and Nutrition, Karolinska Institutet, Hälsovägen 7, SE-4183 Huddinge, Stockholm, Sweden. [16]Department of Clinical Genetics, VU University Medical Center, van der Boechorststraat 7, 1081 BT Amsterdam, Netherlands.
*These authors contributed equally to this work.
†A list of authors and affiliations appears in the Supplementary Information.

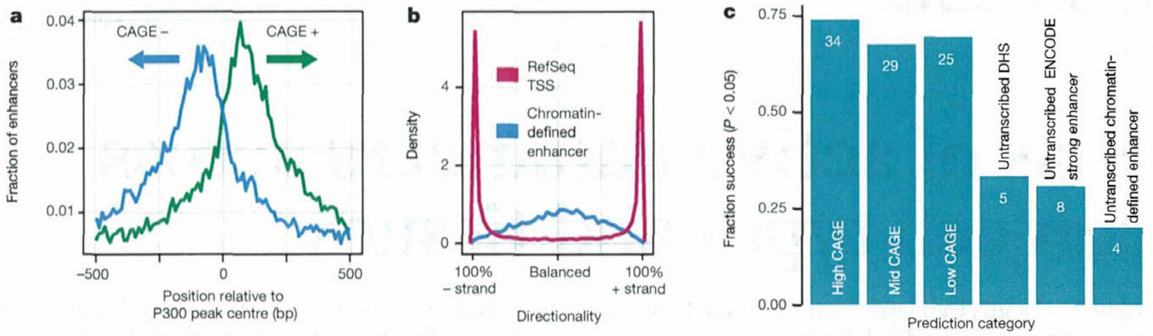

Figure 1 | Bidirectional capped RNAs is a signature feature of active enhancers. a, Enhancers identified by co-occurrence of H3K27ac and H3K4me1 ChIP-seq data[7], centred on P300 binding sites, in HeLa cells were overlaid with HeLa CAGE data (unique positions of CAGE tag 5′ ends, smoothed by a 5-bp window), revealing a bidirectional transcription pattern. Horizontal axis shows the ± 500 bp region around enhancer midpoints. **b**, Density plot illustrating the difference in directionality of transcription according to FANTOM5-pooled CAGE tags mapped within ± 300 bp of 22,486 TSSs of RefSeq protein-coding genes and centre positions of 10,138 HeLa enhancers defined as above. **c**, Success rates of 184 *in vitro* enhancer assays in HeLa cells. Vertical axis shows the fraction of active enhancers (success defined by Student's *t*-test, $P < 0.05$ versus random regions; also see Supplementary Fig. 9). Numbers of successful assays are shown on the respective bar. See main text for details.

were observed in other cell lines (Supplementary Fig. 2a). Enhancer-associated reverse and forward strand transcription initiation events were, on average, separated by 180 base pairs (bp) and corresponded to nucleosome boundaries (Supplementary Figs 3 and 4). As a class, active HeLa-S3 enhancers had 231-fold more CAGE tags than polycomb-repressed enhancers, indicating that transcription is a marker for active usage. Indeed, ENCODE-predicted enhancers[7] with significant reporter activity[8] had greater CAGE expression levels than those lacking reporter activity ($P < 4 \times 10^{-22}$, Mann–Whitney *U* test). A lenient threshold on enhancer expression increased the validation rate of ENCODE enhancers from 27% to 57% (Supplementary Fig. 5).

Although capped RNAs of protein-coding gene promoters were strongly biased towards the sense direction, similar levels of capped RNA in both directions were detected at enhancers (Fig. 1b and Supplementary Fig. 2b, c). Thus, bidirectional capped RNAs is a signature feature of active enhancers. On this basis, we identified 43,011 enhancer candidates across 808 human CAGE libraries (see Supplementary Text and Supplementary Figs 6–8). Interestingly, the candidates were depleted of CpG islands (CGI) and repeats (with the exception of neural stem cells, see ref. 9).

To confirm the activity of newly identified candidate enhancers, we randomly selected 46 strong, 41 moderate and 36 low activity enhancers (as defined by CAGE tag frequency in HeLa cells) and examined their activity using enhancer reporter assays compared to randomly selected untranscribed loci with regulatory potential in HeLa-S3 cells: 15 DHSs[10], 26 ENCODE-predicted 'strong enhancers'[7] and 20 enhancers defined as in Fig. 1a (Supplementary Tables 2 and 3). Whereas 67.4–73.9% of the CAGE-defined enhancers showed significant reporter activity, only 20–33.3% of the untranscribed candidate regulatory regions were active (Fig. 1c and Supplementary Fig. 9a). The same trend was observed in HepG2 cells (Supplementary Fig. 10a, b). Corresponding promoter-less constructs showed that the enhancer transcription read-through is negligible (Supplementary Fig. 9b, c). Many CAGE-defined enhancers overlapped predicted ENCODE 'strong enhancers' or 'TSS' states (25% and 62%, respectively, for HeLa-S3), but there was no substantial difference in validation rates between these classes (Supplementary Fig. 10c, d). In summary, active CAGE-defined enhancers were much more likely to be validated in functional assays than untranscribed candidate enhancers defined by histone modifications or DHSs.

Initiation and fate of enhancer RNAs versus mRNAs

RNA-seq data from matching primary cells and tissues showed that ~95% of RNAs originating from enhancers were unspliced and typically short (median 346 nucleotides)—a striking difference to mRNAs

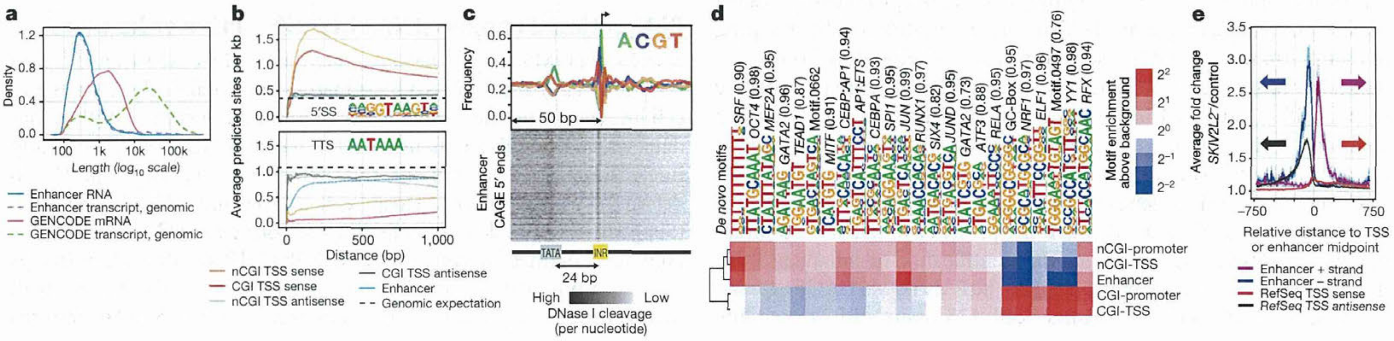

Figure 2 | Features distinguishing enhancer TSSs from mRNA TSSs. a, Densities of the genomic and processed RNA lengths of transcripts starting from enhancer TSSs and mRNA TSSs using assembled RNA-seq reads from 13 pooled FANTOM5 libraries. **b**, Frequencies of RNA processing motifs (5′ splice motif (5′SS, upper panel) and the transcription termination site hexamer (TTS, lower panel)) around enhancer and mRNA TSSs. Vertical axis shows the average number of predicted sites per kb within a certain window size from the TSS (horizontal axis) in which the motif search was done. Dashed lines indicate expected hit density from random genomic background. The window always starts at the gene or enhancer CAGE summits and expands in the sense direction. nCGI, non-CGI. **c**, Average nucleotide frequencies (top panel) and DNase I cleavage patterns (lower panel) of enhancer CAGE peaks (arrow at +1 indicates position of the main enhancer CAGE peaks; direction of transcription goes left to right) reveal distinct cleavage patterns at sequences resembling the INR and TATA elements. **d**, *De novo* motif enrichment analyses around enhancers and non-enhancer FANTOM5 CAGE-defined TSSs (CAGE TSSs matching annotated TSSs are referred to as 'promoters'), contingent on CGI overlap. Top enriched/depleted motifs are shown along with their best-known motif match name. Enrichment versus random background is presented as a heat map. **e**, Vertical axis shows average HeLa CAGE expression fold change versus control at enhancers and RefSeq TSSs after exosome depletion. Horizontal axis shows position relative to the TSS or the centre of the enhancer. Translucent colours indicate the 95% confidence interval of the mean.

(19% unspliced, median 1,256 nucleotides) (Fig. 2a and Supplementary Fig. 11a–c). Unlike TSSs of mRNAs, which are enriched for predicted 5′ splice sites but depleted of downstream polyadenylation signals[11,12], enhancers showed no evidence of associated downstream RNA processing motifs, and thus resemble antisense promoter upstream transcripts (PROMPTs)[11] (Fig. 2b and Supplementary Fig. 11d). Most CAGE-defined enhancers gave rise to nuclear (>80%) and non-polyadenylated (~90%) RNAs[13] (Supplementary Fig. 11e). Based on RNA-seq, few enhancer RNAs overlap exons of known protein-coding genes or large intergenic noncoding RNAs (9 and 1 out of 4,208 enhancers detected, respectively), indicating that they are not a substantial source of alternative promoters for known genes (as in ref. 14).

TSS-associated, uncapped small RNAs (TSSa-RNAs), attributed to RNAPII protection and found immediately downstream of mRNA TSSs[15,16], were detectable in the same positions downstream of enhancer TSSs (Supplementary Fig. 12), indicating that RNAPII initiation at enhancer and mRNA TSSs is similar. Indeed, CAGE-defined enhancer TSSs resembled the proximal position-specific sequence patterns of non-CGI RefSeq TSSs (Fig. 2c and Supplementary Fig. 13a). Furthermore, *de novo* motif analysis revealed sequence signatures in CAGE-defined enhancers closely resembling non-CGI promoters (Fig. 2d and Supplementary Fig. 13b).

Because of the similarity with PROMPTs, we reasoned that capped enhancer RNAs might be rapidly degraded by the exosome. Indeed, small interfering RNA-mediated depletion of the *SKIV2L2* (also known as MTR4) co-factor of the exosome complex resulted in a median 3.14-fold increase of capped enhancer-RNA abundance (Fig. 2e and Supplementary Fig. 14a, b), but only a negligible increase at mRNA TSSs. This increasing trend is similar to that of PROMPT regions upstream of TSSs, although the increase of enhancer RNAs was significantly higher ($P < 4.6 \times 10^{-67}$, Mann–Whitney *U* test; Fig. 2e and Supplementary Fig. 14b, c). Thus, the bidirectional transcriptional activity observed at enhancers is also present at promoters, as suggested previously[17], but in promoters only the antisense RNA is degraded. Furthermore, the CAGE expression of enhancers in control and *SKIV2L2*-depleted cells was proportional (Supplementary Fig. 14d), indicating that virtually all identified enhancers produce exosome-sensitive RNAs. The number of detectable bidirectional CAGE peaks increased 1.7-fold upon *SKIV2L2* depletion and novel enhancer candidates had on average similar, but weaker, chromatin modification signals compared to control HeLa cells (Supplementary Fig. 14e).

CAGE identifies cell-specific enhancer usage

To test whether CAGE expression can identify cell-type-specific enhancer usage *in vivo*, ChIP-seq (H3K27ac and H3K4me1), DNA methylation and triplicate CAGE analyses were performed in five primary blood cell types, and compared to published DHS data (http://www.roadmapepigenomics.org/, Supplementary Table 4). CAGE-defined enhancers were strongly supported by proximal H3K4me1/H3K27ac peaks (71%) and DHSs (87%) from the same cell type. Conversely, H3K4me1 and H3K27ac supported only 24% of DHSs distal to promoters and exons and only 4% of DHSs overlapped CAGE-defined enhancers (Supplementary Fig. 15), indicating that a minority of promoter-distal DHSs identify enhancers. From the opposite perspective, only 11% of H3K4me1/H3K27ac loci overlapped CAGE-defined enhancers and untranscribed loci showed weaker ChIP-seq signals than transcribed ones (Supplementary Fig. 16). Moreover, there was a clear correlation between CAGE, DNase I hypersensitivity, H3K4me1 and H3K27ac for CAGE-defined enhancers expressed in blood cells (Fig. 3a). Accordingly, cell-type-specific enhancer expression corresponds to cell-type-specific histone modifications (Fig. 3b). The majority of selected cell-type-specific enhancers could be validated in corresponding cell lines and were associated with cell-type-specific DNA demethylation (Supplementary Text, Supplementary Fig. 17 and Supplementary Tables 5–8, see also ref. 18). Thus, bidirectional CAGE pairs are robust predictors for cell-type-specific enhancer activity.

An atlas of transcribed enhancers across human cells

The FANTOM5 CAGE library collection[6] enables the dissection of enhancer usage across cell types and tissues comprehensively sampled across the human body. Clustering based on enhancer expression clearly grouped functionally related samples together (Fig. 3c and Supplementary Figs 18 and 19). Although fetal and adult tissue often grouped together, two large fetal-specific clusters were identified: one brain-specific (pink) and one with diverse tissues (green). The fetal-brain cluster is associated with enhancers that are located close to known neural developmental genes, including *NEUROG2*, *SCRT2*, *POU3F2* and *MEF2C* (Supplementary Fig. 18b), for which gene expression patterns correlate with enhancer RNA abundance across libraries, suggesting regulatory interaction (see below). The results corroborate the functional relevance of these enhancers for tissue-specific gene expression and indicate that they are an important part of the regulatory programs of cellular differentiation and organogenesis.

To confirm that candidate enhancers can drive tissue-specific gene expression *in vivo*, five evolutionarily conserved CAGE-defined human enhancers (including the *POU3F2* and *MEF2C*-proximal enhancers

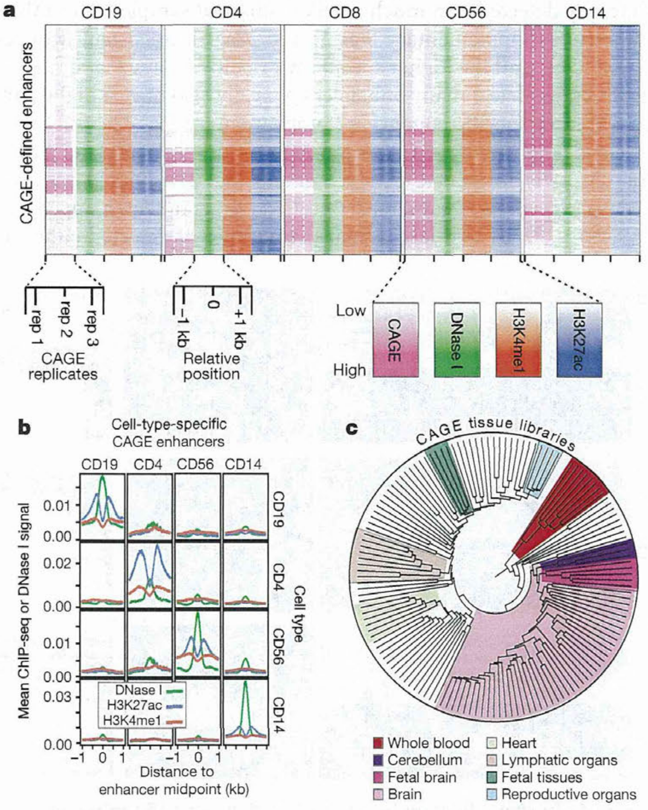

Figure 3 | CAGE expression identifies cell-type-specific enhancer usage. **a**, Relationship between CAGE and histone modifications in blood cells. Rows represent CAGE-defined enhancers that are ordered based on hierarchical clustering of CAGE expression. Columns for the CAGE tags (pink) represent the expression intensity for three biological replicates. DNase I hypersensitivity and H3K27ac and H3K4me1 ChIP-seq signals ± 1kb around the enhancer midpoints are shown in green, blue and orange, respectively. **b**, Mean signal of DNase-seq as well as ChIP-seq for H3K27ac and H3K4me1 (vertical axes) per cell type (rows) in ± 1kb regions (horizontal axes) around enhancer midpoints, for enhancers with blood-cell type-specific CAGE expression (columns). **c**, Dendrogram resulting from agglomerative hierarchical clustering of tissue samples based on their enhancer expression: each leaf of the tree represents one CAGE tissue sample (for a labelled tree and the corresponding results on primary cell samples, see Supplementary Figs 18 and 19). Sub-trees dominated by one tissue/organ type or morphology are highlighted. Some of the enhancers responsible for the fetal-specific subgroup in the larger brain sub-tree are validated *in vivo* (Fig. 4).