Supplementary text, Table legends and Supplementary figures for “An atlas of active enhancers across human cell types and tissues”

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

[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]School of Clinical and Experimental Medicine, College of Medical and Dental Sciences, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK
[4] Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, NC 27599, USA
[5]RIKEN OMICS Science Centre, RIKEN Yokohama Institute, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama City, Kanagawa, 230-0045, Japan
[6]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
[7]Centre for mRNP Biogenesis and Metabolism, Department of Molecular Biology and Genetics, C.F. Møllers Alle 3, Bldg. 1130, DK-8000 Aarhus, Denmark
[8]Department of Molecular and Cellular Biology, Harvard University, USA
[9]The Finsen Laboratory, Rigshospitalet and Danish Stem Cell Centre (DanStem), University of Copenhagen, Ole Maaloes Vej 5, DK-2200, Denmark
[10] Roslin Institute, Edinburgh University, Easter Bush, Midlothian, EH25 9RG Scotland, UK
[11]Genomics Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Road MS 64-121, Berkeley, CA 94720, USA
[12]EMBL Outstation - Hinxton, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD
[13]RIKEN Preventive Medicine and Diagnosis Innovation Program, RIKEN Yokohama Institute, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama City, Kanagawa, 230-0045, Japan
[14]Center of Biosciences, Karolinska Institutet NOVUM, S-14183 Huddinge, Stockholm, Sweden

[#] These authors contributed equally to this work.
‡ A list of consortium members and their affiliations is included below (page 8)
* Correspondence should be addressed to ARRF (alistair.forrest@gmail.com), PC (carninci@riken.jp), MR (michael.rehli@ukr.de) or AS (albin@binf.ku.dk).

SUPPLEMENTARY TEXT

Overlap between enhancer set and the FANTOM5 CAGE tag clusters

Forrest *et al*[1] defined robust CAGE tag clusters (TCs) to find likely promoter locations (see Methods in ref [1]). This definition is substantially more stringent than the one used for CAGE–defined enhancers for which the number of tags is typically low.

Of the CAGE-defined enhancers, only 1784 (4.1%) could be defined by robust TCs. In Forrest *et al*. [1], a machine learning algorithm that predicted mRNA promoters (T. Lassman, manuscript in preparation) was run on these TCs. Considering only those TCs that were predicted mRNA TSSs, 124 enhancers were overlapped (0.3%).

Supportive external data for CAGE-defined enhancers

We calculated the overlap between bidirectional CAGE peak pairs and external data (Supplementary Fig. 8), and used RefSeq TSSs as a reference for comparison. The CAGE-defined enhancers overlap pooled ENCODE DHSs in 89% of cases, compared to 85% of RefSeq TSSs. Pooled ENCODE P300, RNAPII and transcription factors overlap CAGE pairs as often as RefSeq TSSs, while 5-9% more of CAGE-predicted enhancer sets are overlapped by pooled H3K4me1 and H3K27ac peaks.

Evolutionary conservation of CAGE-defined enhancers

CAGE-defined enhancers are more evolutionarily conserved than randomly selected genomic regions but have on average around 2.5-fold lower PhastCons conservation scores[2] than protein-coding RefSeq TSSs (Supplementary Fig. 7a). The conservation is centered at the derived enhancer midpoint and rapidly drops to background levels at around +/- 250 bp, consistent with the width defined by the CAGE tags. We then repeated this analysis for facet-specific enhancers; the conservation is roughly equal between these, with the noteworthy exception of neural stem-cell specific enhancers, which have low conservation. This is due to an over-representation of Long Terminal Repeat (LTR) elements in these enhancers; repeat regions are generally under-represented in CAGE-defined enhancer regions compared to randomly selected genomic regions - see below (Supplementary Fig. 7b).

Analysis of repeat density in enhancers

Repeat data was obtained from the UCSC database and categorized into different types according to the classes defined in the "repeat mask" track. For both permissive and robust enhancer sets, we calculated repeat counts per enhancer in each position within a +/-500 bp window surrounding the center positions of enhancers. In general, the incidence of repeats is low, and in most cases under-represented compared to randomly selected regions. Simple and tRNA repeats are slightly over-represented, but this is based on very few counts: 0.35% and 0.55% of the midpoints of enhancers overlap tRNA repeats and simple repeats, respectively.

Since the average footprints might be biased by outliers, we computed the percentage of the enhancer regions overlapping different repeat types (Supplementary Fig. 7b) for facet-specific enhancers. There is no major difference between the facets except neural stem cells that have an over-representation of LTRs. This is further explored by Fort *et al.* (Fort *et al.*, submitted).

Comparison of features distinguishing CAGE-defined enhancers and RefSeq TSSs

We extended the overlap analysis above (the fraction of enhancers being covered by ChIP-seq peaks) by additional marks and also split up RefSeq TSSs into CpG-island overlapping/non-overlapping (+/-300 bp from the TSS) (Supplementary Fig. 8).

P300, H3K4me1, H3K4me2, H3K9ac, and H3K27ac overlap TSSs and enhancers with a substantially higher frequency than random genomic regions, but CpG island-overlapping TSSs have the highest overlap, followed by enhancers and then non-CpG TSSs. The lower amount of signal at non-CpG TSSs is consistent with their more cell-constrained usage, as the ChIP-seq data is obtained from only a handful of cell lines, so many of these TSSs might not be expressed in these experiments. This claim is supported by the RNAPII peak overlap.

This is also true for H3K9me1, H3K36me3, H4K20me1 marks, but here the marks have a different positional distribution, with a stronger signal after the TSS for RefSeq TSSs whereas in enhancers the mark is uniformly distributed. This makes immediate sense for H3K36me3, and indicates that the enhancer RNAs are not actively being elongated. The small subset of ubiquitously expressed enhancers shows patterns that are similar, but not identical, to that of CpG-overlapping TSSs (see below).

Other important enhancer features that set them apart from TSSs are the balanced bidirectional transcription, and the much lower RNA abundance: the median TPM of RefSeq CAGE is 19.7-fold higher than that of enhancers (see table below: values are in TPMs). Also see main text for a discussion on RNA fates, DNA sequence signal downstream of TSSs and degradation rates.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
RefSeq TSSs	0	25.02	75.97	623	214.7	1638000
Permissive enhancers	0.3284	2.321	3.838	11.36	8.701	2308

Motif analysis of facet-specific enhancers

We hypothesized that the observed cell type-/tissue-specificity of enhancers is regulated on the level of transcription factor binding. Therefore, we identified enhancers with significantly higher expression in one facet compared to others (Methods); 60% of enhancers are enriched in at least one facet. We analyzed all such facet-enriched sets (with >100 enhancers) for over-represented sequence patterns using HOMER[3]. In line with current knowledge, we find consensus motifs of known key regulators over-represented in corresponding cell types, for instance ETS, C/EBP, and NF-κB in monocyte-specific enhancers, RFX and SOX in neurons, and HNF1 and HNF4a in hepatocytes (Supplementary Fig. 23). Strikingly, AP1 and, to a lesser extent, ETS motifs appear to be enriched across all facets.

Features of ubiquitously expressed enhancers (u-enhancers)

We repeated the ChIP overlap study described above, but also broke up enhancers into ubiquitous (defined by tissue or cell type facets, resulting in 247 and 241 enhancers, respectively, these sets overlap by 106 enhancers) and non-ubiquitous (Supplementary Fig. 26d), and compared the results to RefSeq TSSs, broken up by CpG overlap as above. U-enhancers have several features setting them apart from other enhancers as well as canonical TSSs:

First, u-enhancers are highly enriched for P300 and RNAPII ChIP sites compared to other regions.

Second, for CTCF, H3K4me2, H3K4me3, and H3K9ac the overlap fraction is high and mirrors that of CpG island-overlapping RefSeq TSSs. H3K79me2 peaks also overlap equally often with CpG-island RefSeq TSSs and u-enhancers, but the positional distribution is different: RefSeq TSSs have a higher amount downstream while u-enhancers show a distinct double peak. Since H3K79me2 is a suggested elongation mark, this may indicate that the RNAs from u-enhancers are longer than those of other enhancers, but still bidirectionally transcribed (as opposed to canonical protein-coding TSSs, whose capped transcription is mostly unidirectional). This is corroborated by RNAseq data which shows ~200 nt longer RNAs from u-enhancers (see main text).

Conversely, u-enhancers overlap H3K4me1 and H3K27ac peaks as often as other enhancers (and CpG island-overlapping RefSeq TSSs).

Thus, u-enhancers share features both with CpG island TSSs and canonical enhancers but are even more enriched for P300 and have a unique H3K79me2 profile.

Furthermore, they overlap cohesin peaks (defined as the intersection of STAG1 and RAD21 chip peaks[4]) more than other CAGE-defined enhancers, and are also more associated with interactions defined by ChIA-PET

(Supplementary Fig. 26a, b). A caveat with these results is that since the u-enhancers are defined to be ubiquitous, they implicitly have a higher chance to overlap ChIP-seq peaks or ChIA-PET interactions from a few cell lines than non-ubiquitous enhancers.

Moreover, compared to non-ubiquitous enhancers, the RefSeq TSSs that are closest to u-enhancers are more commonly CpG-overlapping and (consistently) have lower expression specificity. The TSS-enhancer distance also tends to be slightly shorter (Supplementary Fig. 28b). These genes are significantly enriched for zink finger transcription factors, membrane bound proteins including ion channels, and different biosynthesis pathways, using RefSeq TSSs closest to non-ubiquitous enhancers as background (Supplementary Table 14) and the DAVID tool[5].

In vitro validation of blood cell enhancers

To ensure that blood-cell specific enhancers displayed in Figure 3 can act as enhancers, we validated 39 regions, selected solely based on CAGE expression, using enhancer-reporter gene assays in triplicates. We tested the ability of enhancer regions (~1kb) to enhance the activity of a basal E2F promoter in transient transfections of cell line models for 3 of the cell types: THP1 (monocytes), Jurkat (T cells) and Daudi (B cells). B cell, monocyte, and T cell-specific enhancers induced a >4-fold increase in reporter gene signal relative to the enhancer-less control in 92% (12/13), 55% (6/11) and 33% (5/15) of cases in the respective cell line. The validation rate is high considering the artificial nature of reporter assays and the fact that leukemia cell lines are similar, but not identical to the primary cells. In line with previous studies[6,7], we also observe cell type-specific DNA demethylation across the validated enhancer regions, in addition to strong cell type-specific histone signals (Supplementary Fig. 17, and Supplementary Tables 5-8).

For B cell enhancers which gave positive result, we also removed the E2F1 promoter and repeated the analysis, to see to what degree the weak TSS activity of the enhancer influenced the result, since for these experiments the enhancer is placed upstream of the promoter and reporter gene. The median contribution of the enhancer was 13% relative to constructs with both enhancer and promoter.

Co-occurrence analysis of TF motifs and peaks in associated enhancer-promoter pairs

Expression correlation between robust enhancers (n = 38,554) and robust DPI promoters (Forrest *et al.*, same issue) within 500 bp of any known transcript annotation (n = 93,558) was calculated for all enhancer–promoter pairs within 500 kb of each other. Pairs with Pearson correlation >= 0.5 were denoted cEPPs and retained for further analysis (n = 56206). Enhancers were scanned in a 401 bp window (enhancer midpoint +/- 200 bp) for the presence of conserved TFBSs (TRANSFAC motif scans available through the tfbsConsSites track from the UCSC Genome Browser) and ENCODE TF-ChIP signal (wgEncodeRegTfbsClusteredV2 track). Similarly, FANTOM5 CAGE-inferred promoters (regardless of annotation) were searched in a 1 kb window (TC CAGE summit position +/- 500 bp). For each individual motif or TF, the probability of finding it in a randomly picked enhancer–promoter pair was calculated as the product of the overlap frequency of the motif/TF in enhancers and promoters separately. This probability was used along with the number of cEPPs (“trials”) and observed co-occurrences (“successes”) in a binomial test to determine whether the motif/TF was co-occurring in cEPPs more than would be expected by chance, with multiple testing correction using the Benjamini-Hochberg method. Out of 253 motifs, 169 (67%) were co-occurring in cEPPs to a significantly higher degree than expected by chance (FDR 5%). Similarly, a majority (94 out of 112, 84%) of TF binding as represented by TF-ChIPs obtained from the ENCODE project co-occurred in cEPPs significantly more often than expected by chance.

GWAS SNPs within enhancer regions

Individual examples of likely regulatory SNPs are shown in Supplementary Figure 32, including SNPs associated to diabetic nephropathy, Crohn's disease, multiple sclerosis and systemic sclerosis all overlapping enhancers close to the TSSs of nearby genes which also are implicated in respective disease[8-10]. Many of the potential interactions are verified by ChIA-PET data (Supplementary Fig. 29a). While the Crohn's disease SNP was recently shown to overlap corresponding hypersensitive sites, this analysis adds the interaction with the PTGER4 gene and detailed expression over the whole body.

While more thorough experiments are needed to infer the exact functional impact of these variations within enhancers, this study highlights the possibility of using the enhancer set to infer the function of non-coding SNPs that are otherwise hard to characterize. We project that similar approaches combined with SNPs from the 1000 genome project data will be fruitful, ideally in combination with high-throughput targeted assays that pinpoint the effect of single nucleotide changes in enhancers[11].

SUPPLEMENTARY TABLE LEGENDS

Table S1: Summary of the FANTOM5 CAGE libraries used in this study.

Table S2: Primers used for reporter plasmid modification for HeLa and HepG2 in *vitro* enhancer validations.

Table S3: Summary of Hela and HepG2 enhancer reporter assay validation results, plasmid construction sequences and amplicons. All sequence IDs refer to the hg19 assembly. “HeLa.rep1.fc” indicates the Firefly/Renilla ratio of construct divided with the corresponding mean ratio of all tested random genomic regions (1st replicate transfection in HeLa). ‘p’ indicates the P-value from one-sided t-test of the three replicates vs. random regions.

Table S4: Summary of ChIP-seq data used in this study (except ENCODE data - see Methods).

Table S5: Sequences of all oligonucleotides used for cloning of enhancer-reporter constructs for *in vitro* validation.

Table S6: Summary of epigenetic and reporter gene data of *in vitro* validated enhancer regions.

Table S7: List of all oligonucleotides that were designed to generate amplicons from bisulfite-treated DNA for EpiTyper (MALDI-TOF MS) analysis. Genomic locations are based on the Build 37 assembly by NCBI (hg19).

Table S8: EpiTyper (MALDI-TOF MS) methylation ratios for individual CpG units of amplicons covering enhancer regions. Mean methylation ratios are given for CD4+CD25- T cells, CD8+ T cells, CD19+ B cells, CD56+ NK cells and human blood monocytes that were measured from two individual healthy donors.

Table S9: A: Selected human Cis-Regulatory Elements (CRE) and control regions used in zebrafish transient reporter assays. Transgene-driven reporter expression at 48 hpf is depicted as a ratio of the number of embryos showing tissue-specific activity versus the total number of embryos injected with the enhancer-containing construct. Merged numbers from at least 3 independent injection experiments are shown.

B: Quantitation of transient expression displayed by 48 hpf zebrafish embryos injected with selected human CREs and control regions. Transgene-driven reporter expression at 48 hpf is depicted as a ratio between the number of embryos showing tissue-specific activity versus the total number of embryos injected with the enhancer-containing construct or “enhancer-less” gata2 promoter containing control vector. Merged numbers from at least 3 independent injection experiments are shown. Ectopic expression includes all expression domains displayed by the control vector. Also See figure S20B.

Table S10: Summary of primary cell facets.

Table S11: Summary of tissue/organ facets.

Table S12: Bed file of locations of ubiquitous enhancers defined by primary cell expression facets. Coordinates refer to the hg19 assembly.

Table S13: Bed file of locations of ubiquitous enhancers defined by tissue/organ expression facets. Coordinates refer to the hg19 assembly.

Table S14: DAVID Gene Ontology results comparing RefSeq genes closest to ubiquitous enhancers (foreground) to RefSeq genes closest to non-ubiquitous enhancers.
Table S15: Locations, lengths and member enhancers of super clusters (see main text).
Table S16: GWAS-associated SNPs overlapping enhancers.

SUPPLEMENTARY REFERENCES

1. The FANTOM Consortium. A promoter level mammalian expression atlas. *Submitted*
2. Siepel, A. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research* **15,** 1034–1050 (2005).
3. Heinz, S. *et al.* Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell* **38,** 576–589 (2010).
4. Schmidt, D. *et al.* A CTCF-independent role for cohesin in tissue-specific transcription. *Genome Res* **20,** 578–588 (2010).
5. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4,** 44–57 (2008).
6. Schmidl, C. *et al.* Lineage-specific DNA methylation in T cells correlates with histone methylation and enhancer activity. *Genome Research* **19,** 1165–1174 (2009).
7. Lister, R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nat Genet* **462,** 315–322 (2009).
8. McDonough, C. W. *et al.* A genome-wide association study for diabetic nephropathy genes in African Americans. *Kidney Int* **79,** 563–572 (2010).
9. Libioulle, C. *et al.* Novel Crohn Disease Locus Identified by Genome-Wide Association Maps to a Gene Desert on 5p13.1 and Modulates Expression of PTGER4. *PLoS Genet* **3,** e58 (2007).
10. Franke, A. *et al.* Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat Genet* **42,** 1118–1125 (2010).
11. Patwardhan, R. P. *et al.* Massively parallel functional dissection of mammalian enhancers in vivo. *Nat Biotechnol* **30,** 265–270 (2012).

MEMBERS OF THE FANTOM CONSORTIUM

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 Schmid[l4{], 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 Dalla40, 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 Ekwall50, 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 Harbers1, 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], MutsumiKanamori-Katayama[1], ArtemS. 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], NaotoKondo[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 Mizuno80, 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-Sato1, Yutaka Nakachi[80,85], Fumio Nakahara[75], Toshiyuki Nakamura[72], Yukio Nakamura[86], Kenichi Nakazato[1], Erik van Nimwegen[24], Noriko Ninomiya1, 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 Rizzu19, Marco Roncador[21], Sugata Roy[1,2], Morten B. Rye[48], Eri Saijyo[1], Antti Sajantila90, Akiko Saka1, 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 Yoneda72, 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), 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa230-0045, 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 andRoyal (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 W120NN,UK.
9 Engelhardt Institute of Molecular Biology,RussianAcademyof Sciences,Vavilov str. 32,Moscow119991,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-17177 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, Charite´ Campus Mitte, Universita¨tsmedizin 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, Kingdomof 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 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. 5
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.
52RIKEN 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 BS81UB, 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 ComputationalBiology 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 Bonomea265, 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 forMolecularMedicine and Genetics,Wayne StateUniversity, 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 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,130Dongdeok-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 BS81UB,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, Charite´ Universita¨tsmedizin 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 andCellularBiology,HarvardUniversity,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 PersonalizedMedicine 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 Universita¨t 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 Academyof Sciences (CeMM),1090Vienna, 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.).

The core members of FANTOM5 phase 1 were:

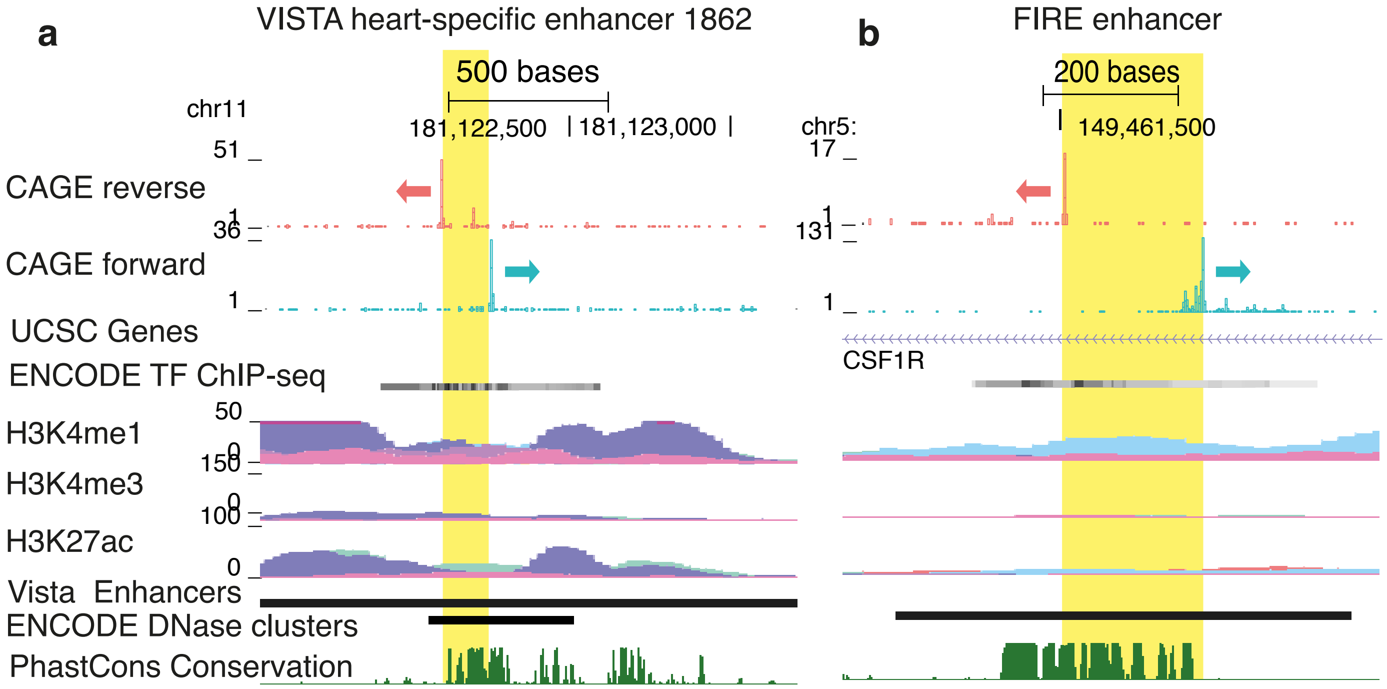
Alistair R. R. Forrest; Hideya Kawajil; J. Kenneth Baillie; Michiel J. L. de Hoon; Timo Lassmann; Masayoshi Itoh; KimM. Summers,Harukazu Suzuki, Carsten O.Daub, Jun Kawai, Peter Heutink, Winston Hide, Tom C. Freeman, Boris Lenhard, Vladimir B. Bajic, Martin S. Taylor, Vsevolod J. Makeev, Albin Sandelin, 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.

CAGE Libraries were generated by:

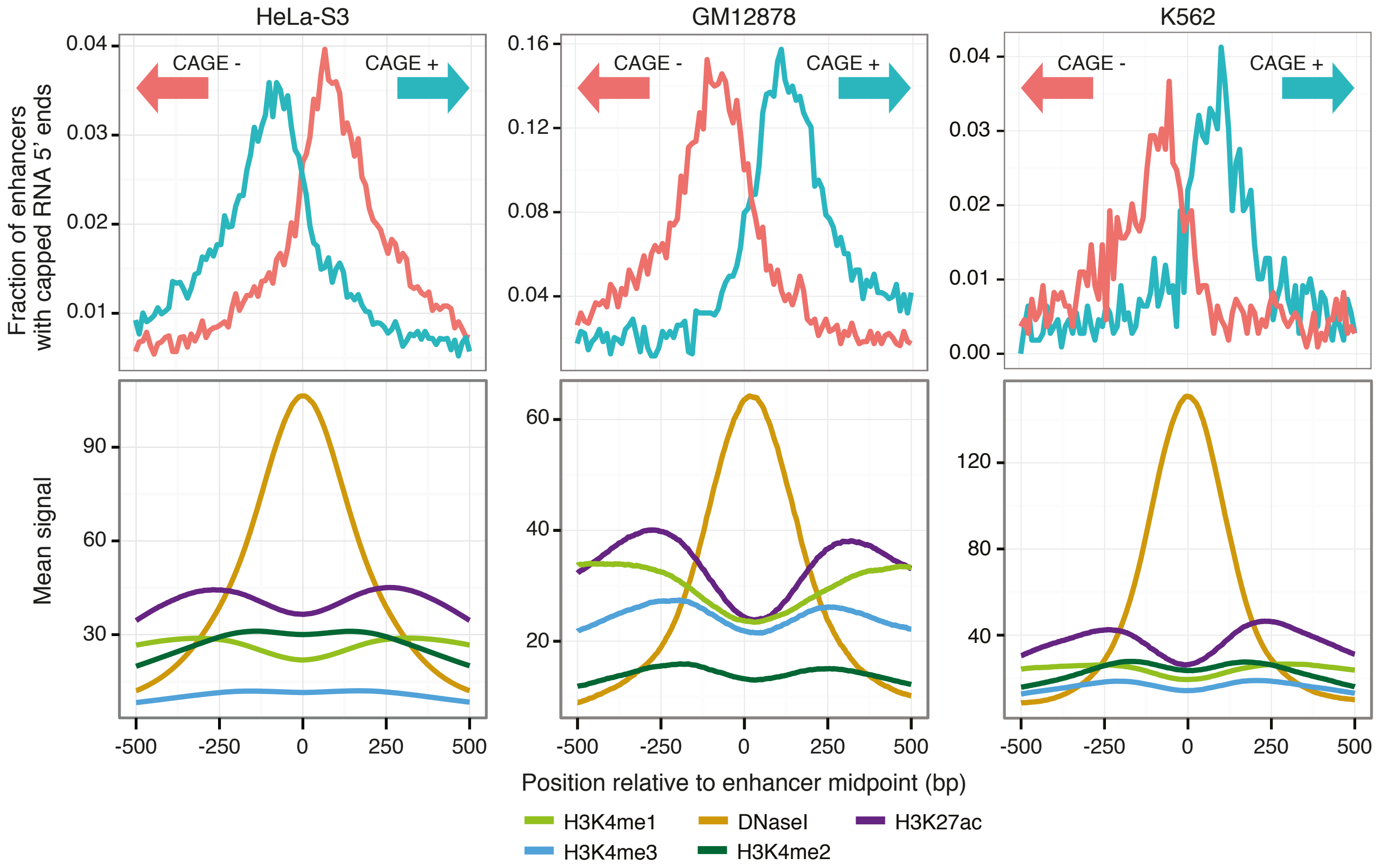
A. Kaiho, A. Kubosaki, A. Saka, C. Simon, E.S., F.H., H.N., J. Kawai, K. Kaida, K.N., M. Furuno, M. Murata, M. Sakai, M. Tagami, M.I., M.K.,M.K.K., N.K.,N.N., N.S., P.C., R.M., S. Kato, S.N., S.N.-S., S.W., S.Y., T.A., T. Kawashima.



UCSC genome browser examples of well-studied enhancers detected by CAGE
a, a VISTA heart enhancer and **b**, the FIRE enhancer, detected by bidirectional CAGE pairs (yellow highlights; arrows show the transcript direction). Also shown are ENCODE transcription factor binding, H3K4me1, H3K4me3 and H3K27ac ChIP-seq, DHS data and PhastCons43 conservation.

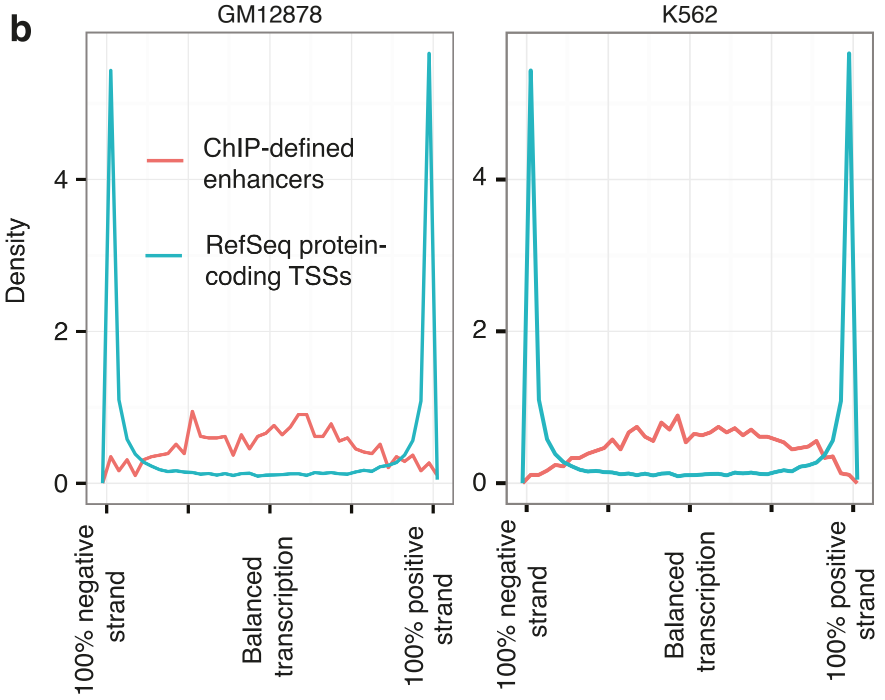
Supplementary Figure 2

a

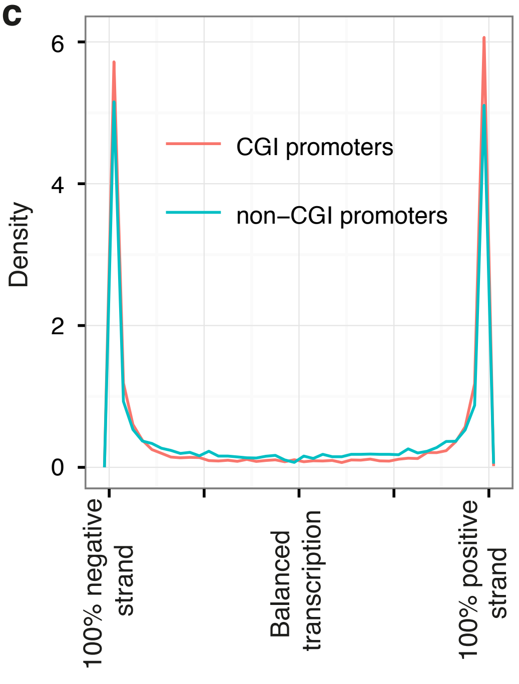


a, CAGE and chromatin profiles over enhancers defined by HeLa-S3 (left), GM12878 (middle) and K562 (right) ENCODE ChIP-seq data
CAGE data are from respective cell lines. Details as in Figure 1a, focusing on P300 sites that are close to H3K4me1 and H3K27ac ChIP-seq peaks. For GM12878 and K562, midpoints are derived from NFKB and GATA1 binding sites, respectively, close to P300, H3K4me1 and H3K27ac ChIP-seq peaks from the same cell line.

b



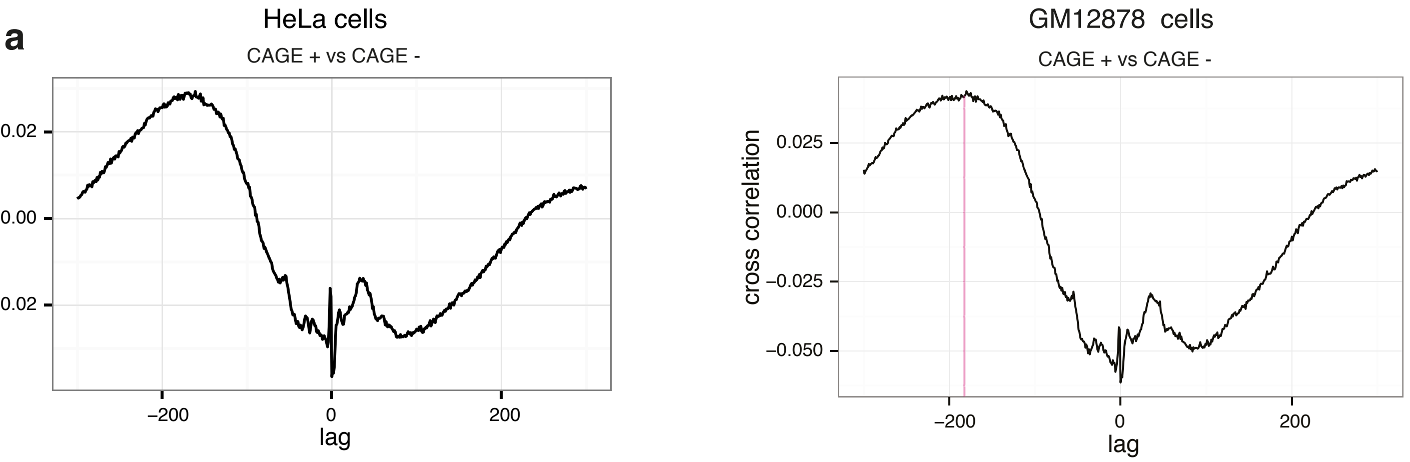
c



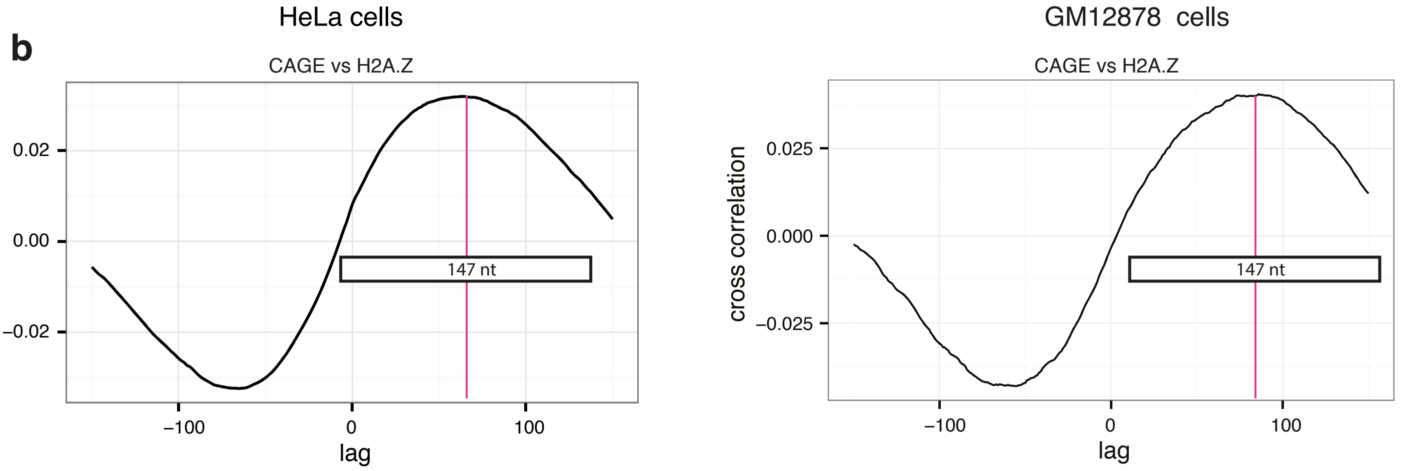
b, Directionality of capped transcription
Densities of transcribed strand bias (directionality) at annotated TSSs and chromatin-defined enhancers in GM12878 (left) and K562 (right) cells as in Figure 1b.

c, Directionality vs. CpG overlap
Directionality (strand bias) of transcription in HeLa cells at RefSeq mRNA TSSs, as in Figure 1b but breaking up the TSSs on their overlap with CpG islands (CGIs). The directionality is unaffected, and the large majority of TSSs are close to unidirectional regardless of CpG content.

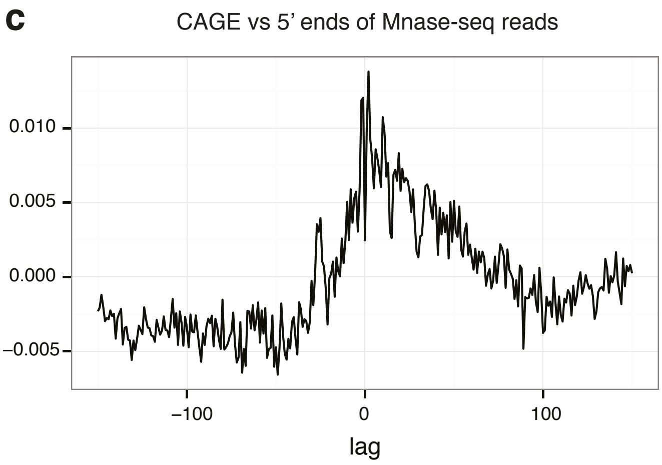
Supplementary Figure 3



a, CAGE cross-correlation of opposite strands
CAGE forward strand vs. CAGE reverse strand cross correlation at HeLa (left) and GM12878 (right) enhancers (ChIP-seq derived) shows that minus strand CAGE tags are most likely 180 bp upstream of plus strand CAGE tags (lag -180: pink line). Unique tags from pooled samples were used. This means that the typical CAGE-defined boundary is 180 bp.



b, CAGE vs H2A.Z cross correlation
CAGE vs H2A.Z cross correlation at HeLa (left) and GM12878 (right) enhancers (ChIP-seq derived) shows that H2A.Z signal from respective cells is most likely 73 bp downstream of CAGE tags (pink line). Assuming that ENCODE signal summits are at the mid point of nucleosomes, that are likely around 147 bp, this means that initiation starts just at the boundary of nucleosomes. Also see panel c as well as DNase I cleavage cross-correlations in Supplementary Figure 4, supporting the same hypothesis.



c, CAGE vs MNase-seq cross-correlation
CAGE vs 5' ends of MNase-seq (nucleosome) reads cross-correlation using ENCODE GM12878 MNase-seq data (9 pooled replicates) at GM12878 enhancers (ChIP-seq derived). The highest correlation is observed close to 0 lag suggesting that transcription initiates at the nucleosome boundary.

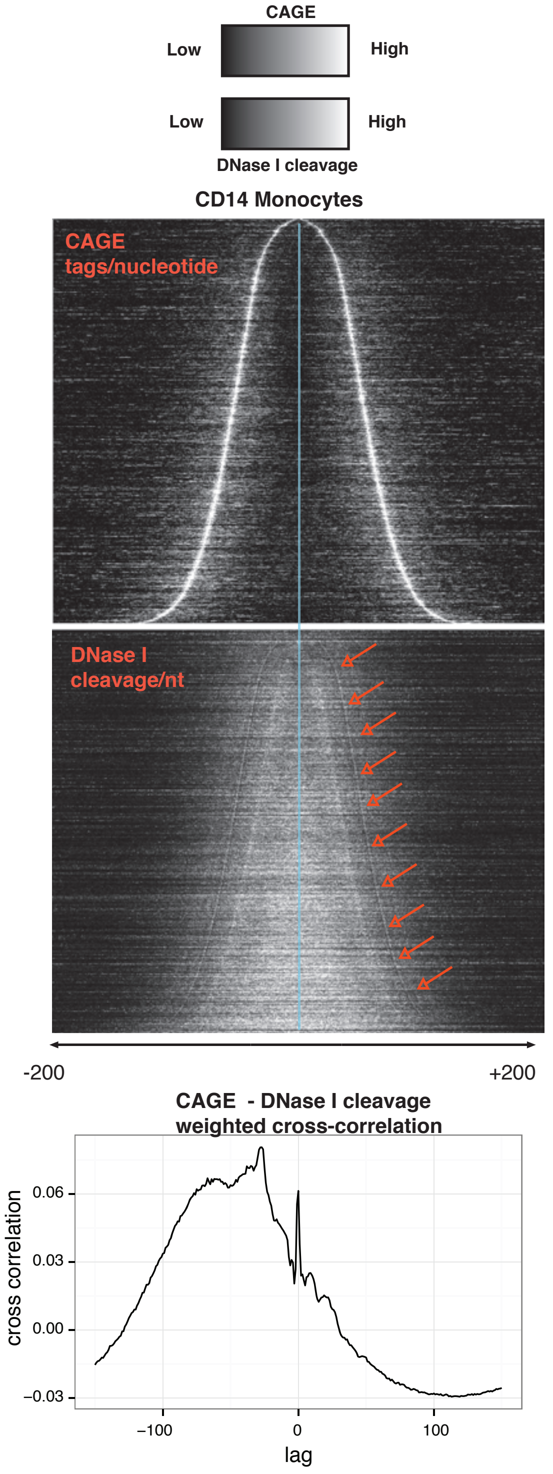
Supplementary Figure 4

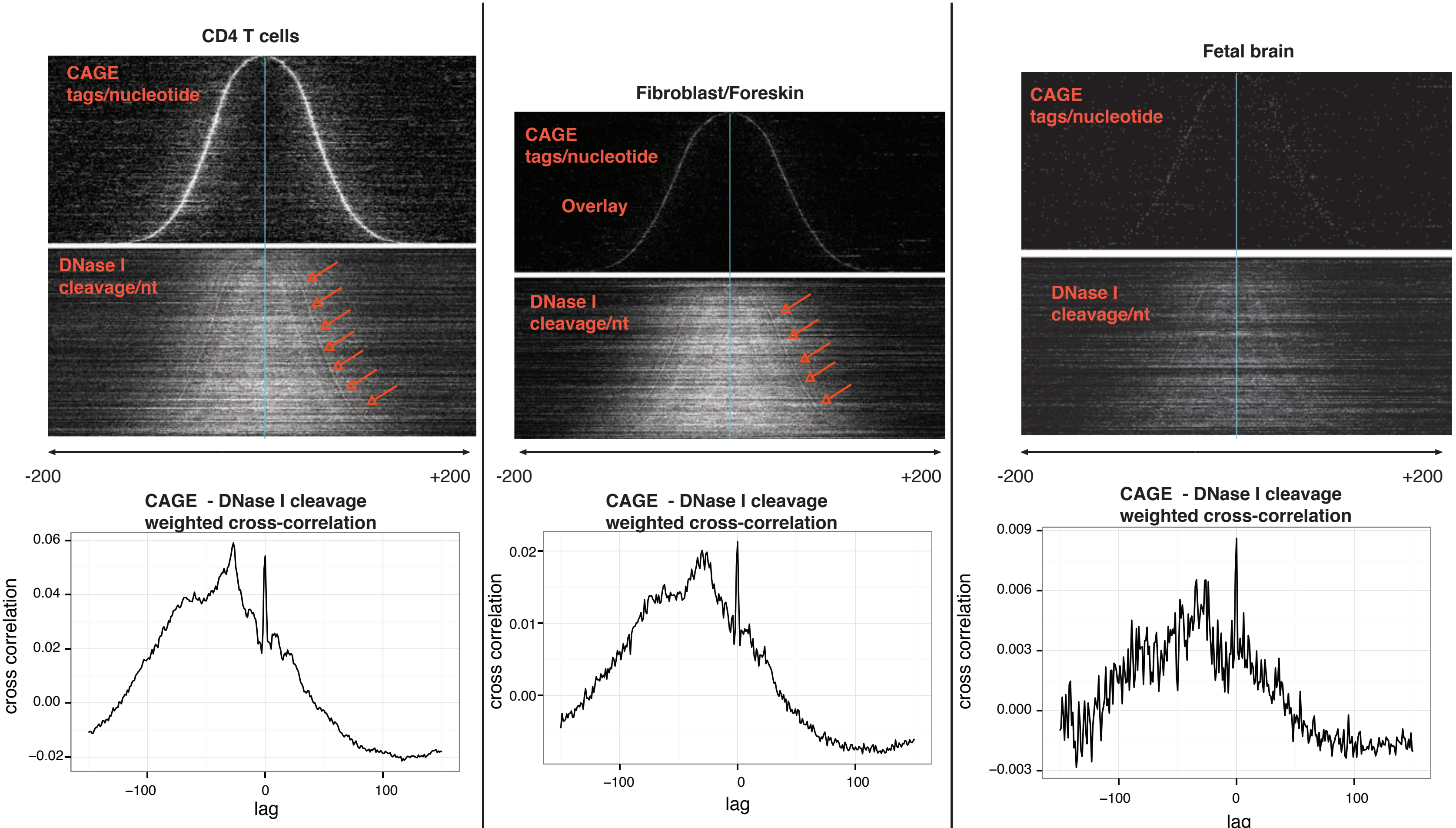
DNase I cleavage and CAGE tag density in enhancer regions in different cell types

The upper two plots in each column show nucleotide resolution DNase I cleavage intensity and the density of CAGE 5' ends (regardless of strand, but in general, CAGE intensity on the left is due to CAGE tags on the negative strand, and vice versa for CAGE tags on the positive strand), where white corresponds to high intensity. Each row corresponds to the +/- 200 bp region centered around the midpoint of an enhancer expressed in a given cell type. The blue line corresponds to the center of the enhancer. The rows are ordered based on the CAGE tag distribution. Note that the cleavage intensity is considerably higher within the sharp boundary defined by the CAGE tags.

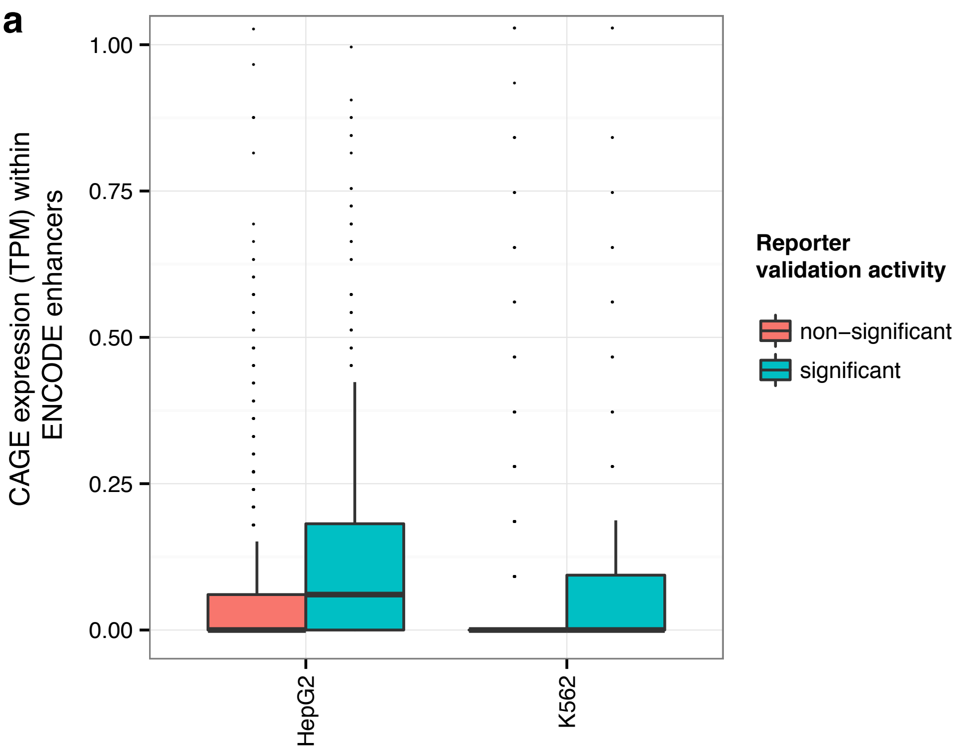
The lower plot shows cross-correlation between the DNase I cleavage intensity and CAGE. The highest correlation is at -28 and 0 corresponding to the two weak DNase cleavage 'bands' evident in the monocyte,T cell and Fibroblast/Foreskin images (red arrows).

Together with Supplementary Figure 3, this strongly indicates that the CAGE tags define the boundary of the accessible region.

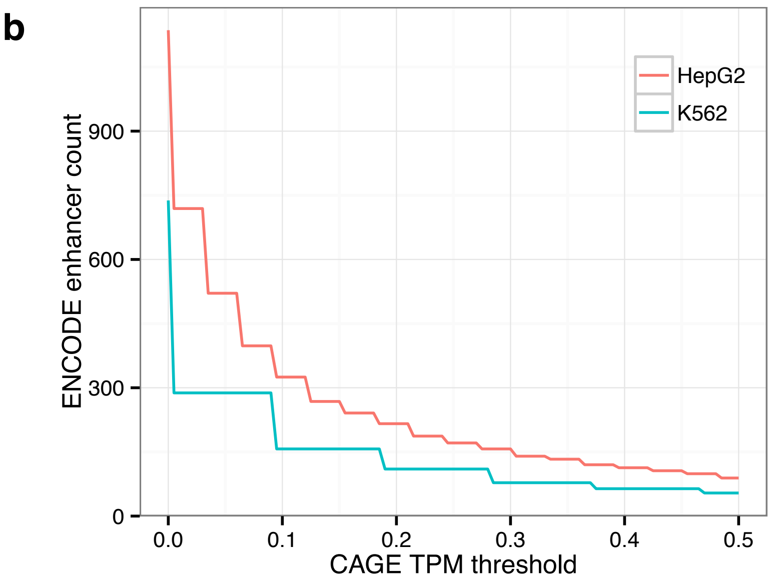




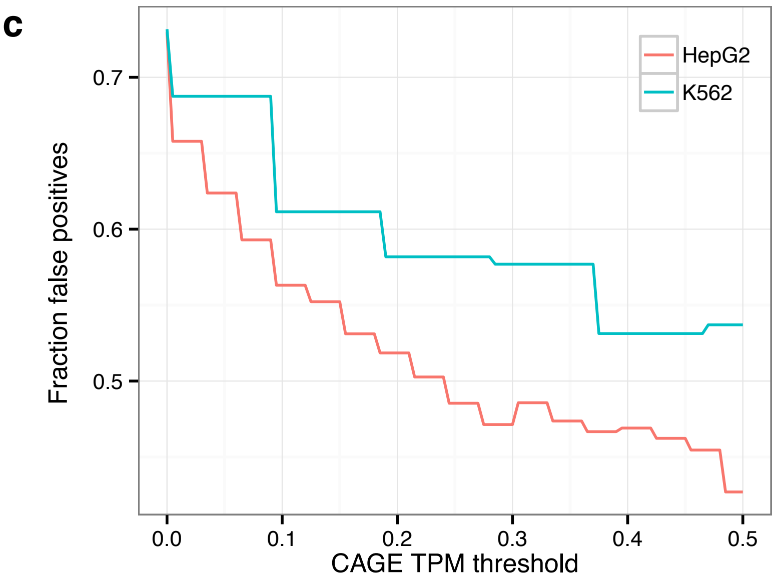
Supplementary Figure 5



a, CAGE expression of predicted ENCODE enhancers with or without enhancer signal
Box-and-whisker plots of TPM-normalized CAGE tag counts (vertical axis) in 401 bp windows centred on mid points of ENCODE-predicted enhancers. The expression of 198 out of 738 K562 enhancers and 307 out of 1136 HepG2 enhancers with significant enhancer reporter activity are plotted separately from the non-significant ones. It is clear that ENCODE enhancers with significant enhancer reporter expression are more transcribed than inactive ones.

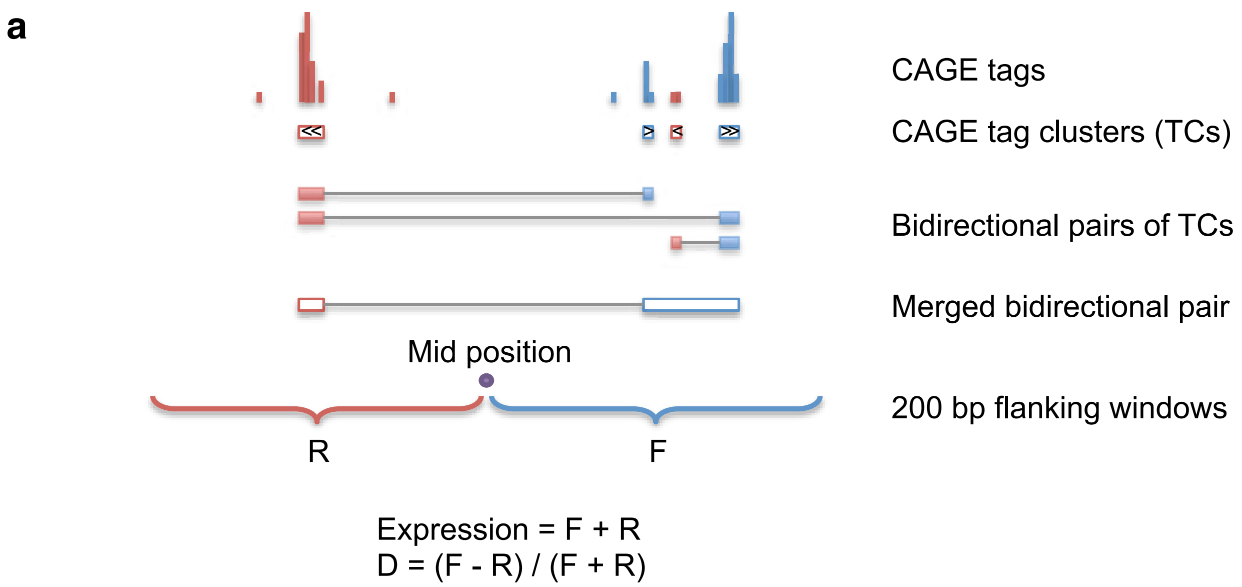


b, ENCODE predicted enhancer counts vs CAGE expression
The number of ENCODE enhancers (vertical axis) as a function of increasing CAGE TPM thresholds (horizontal axis).

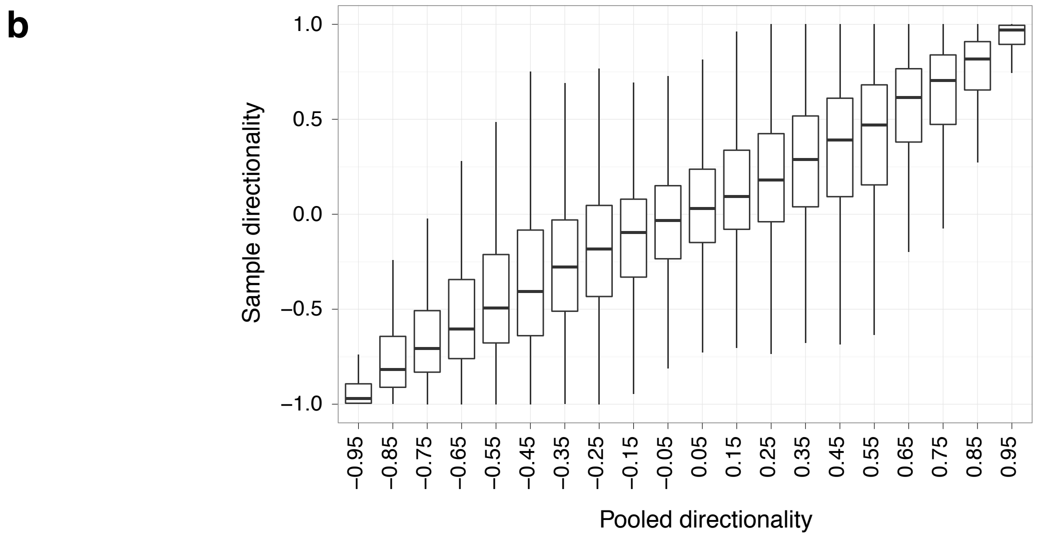


c, ENCODE predicted enhancer FDR as a function of CAGE expression
The fraction of false positives (vertical axis) as a function of increasing CAGE TPM thresholds (horizontal axis), calculated as the fraction of non-significant enhancers among those fulfilling a given expression cutoff. The original sets (with no expression cutoff) have fractions of false positives of 0.730 and 0.731 for HepG2 and K562, respectively.

Supplementary Figure 6

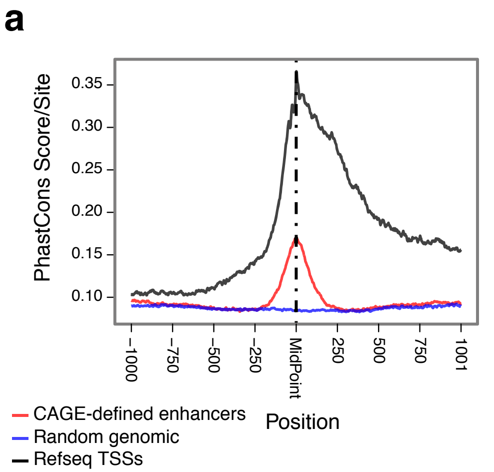


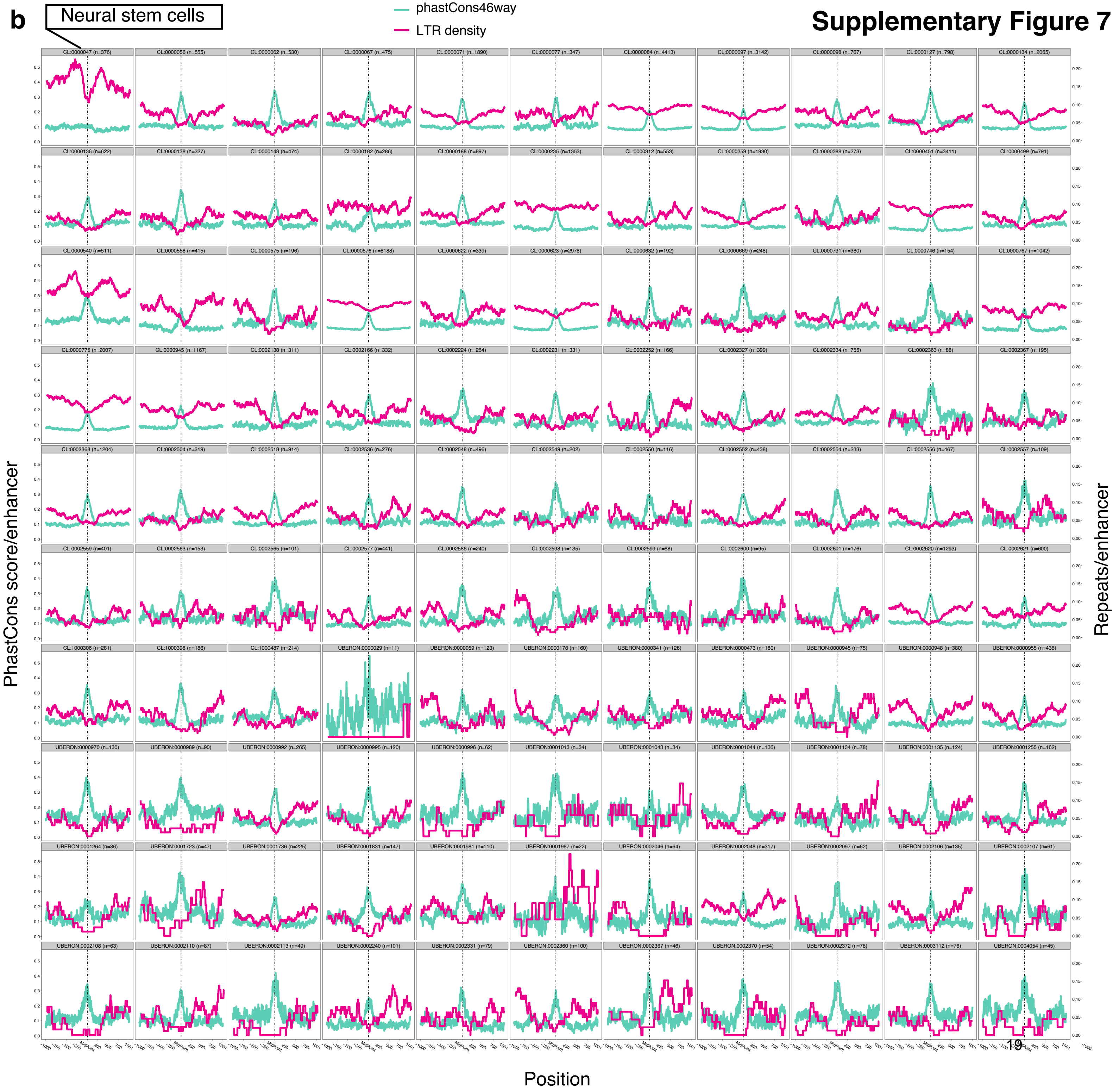
a, Conceptual image illustrating the definition of bidirectionally transcribed loci
Bidirectionally transcribed loci were defined from CAGE tag clusters (TCs) supported by at least two CAGE tags in at least one sample (TCs defined in Forrest *et al.*). Only TCs not overlapping antisense TCs were used. We identified divergent (reverse-forward) TC pairs separated by at most 400 bp and merged all such pairs containing the same TC, while at the same time avoiding overlapping forward and reverse strand transcribed regions (prioritization by expression ranking). A center position was defined for each bidirectional locus from the mid position between the rightmost reverse strand tag cluster (TC) and leftmost forward strand TC included in the merged bidirectional pair. Each bidirectional locus was further associated with two 200 bp regions immediately flanking the center position, one (left) for reverse strand transcription and one (right) for forward strand transcription, in a divergent manner. The merged bidirectional pairs were further required to be bidirectionally transcribed (CAGE tags supporting both windows flanking the center) in at least one individual sample, and to have a greater aggregate of reverse CAGE tags (over all FANTOM5 samples) than forward CAGE tags in the 200 bp region associated with reverse strand transcription, and vice versa.



b, Correlation between sample directionality and estimating directionality over all samples
Box and whisker plots of directionality scores calculated for individual samples binned by the directionality score for all (aggregated) samples shows that pooled directionality is a good estimator of the directionality of individual samples.

Supplementary Figure 7



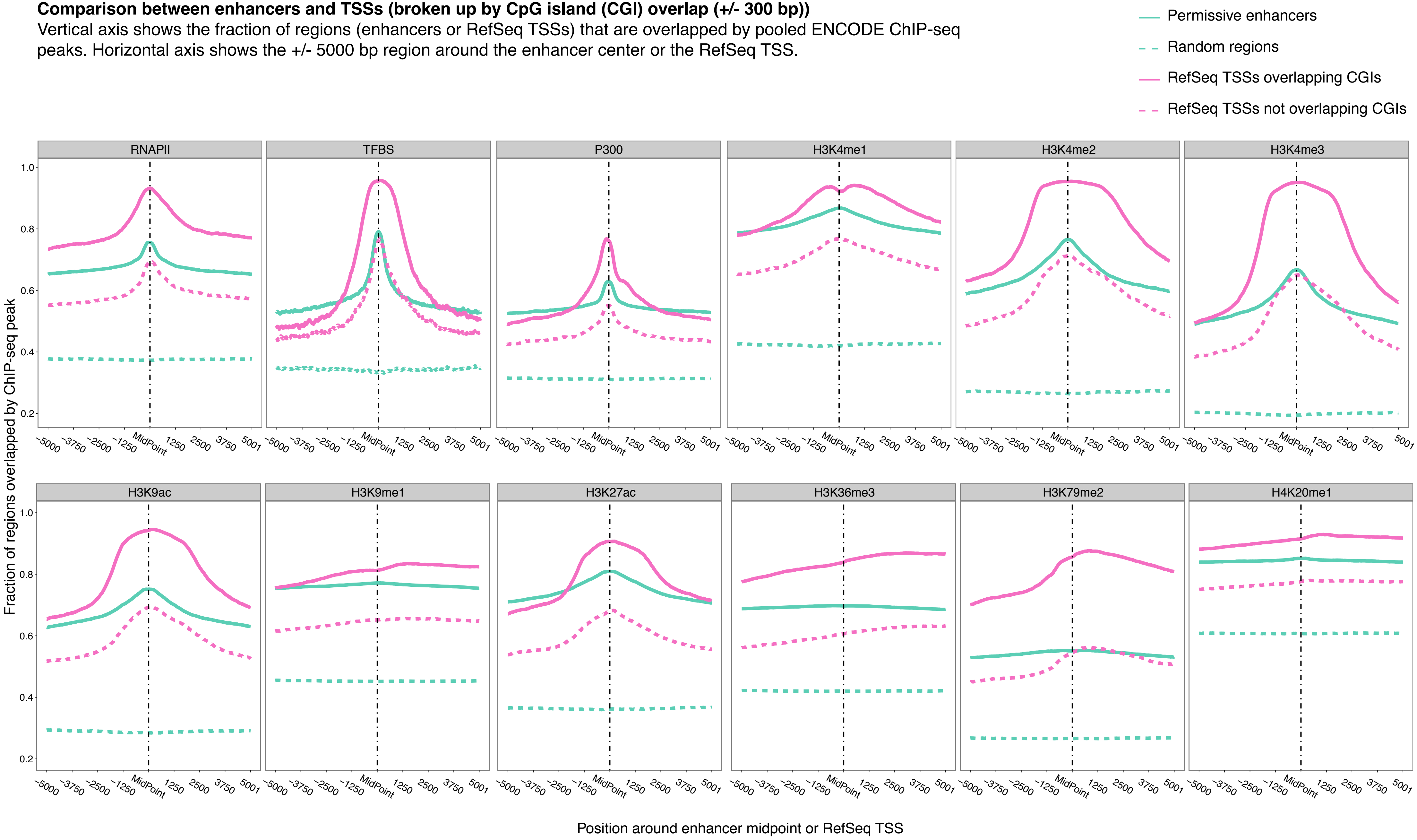


a, Overall conservation of CAGE-defined enhancers
Average PhastCons conservation of enhancers, RefSeq TSSs and randomly selected genomic regions. The midpoint refers to the midpoint of the enhancer or randomly selected region, and the TSS position for RefSeq genes.

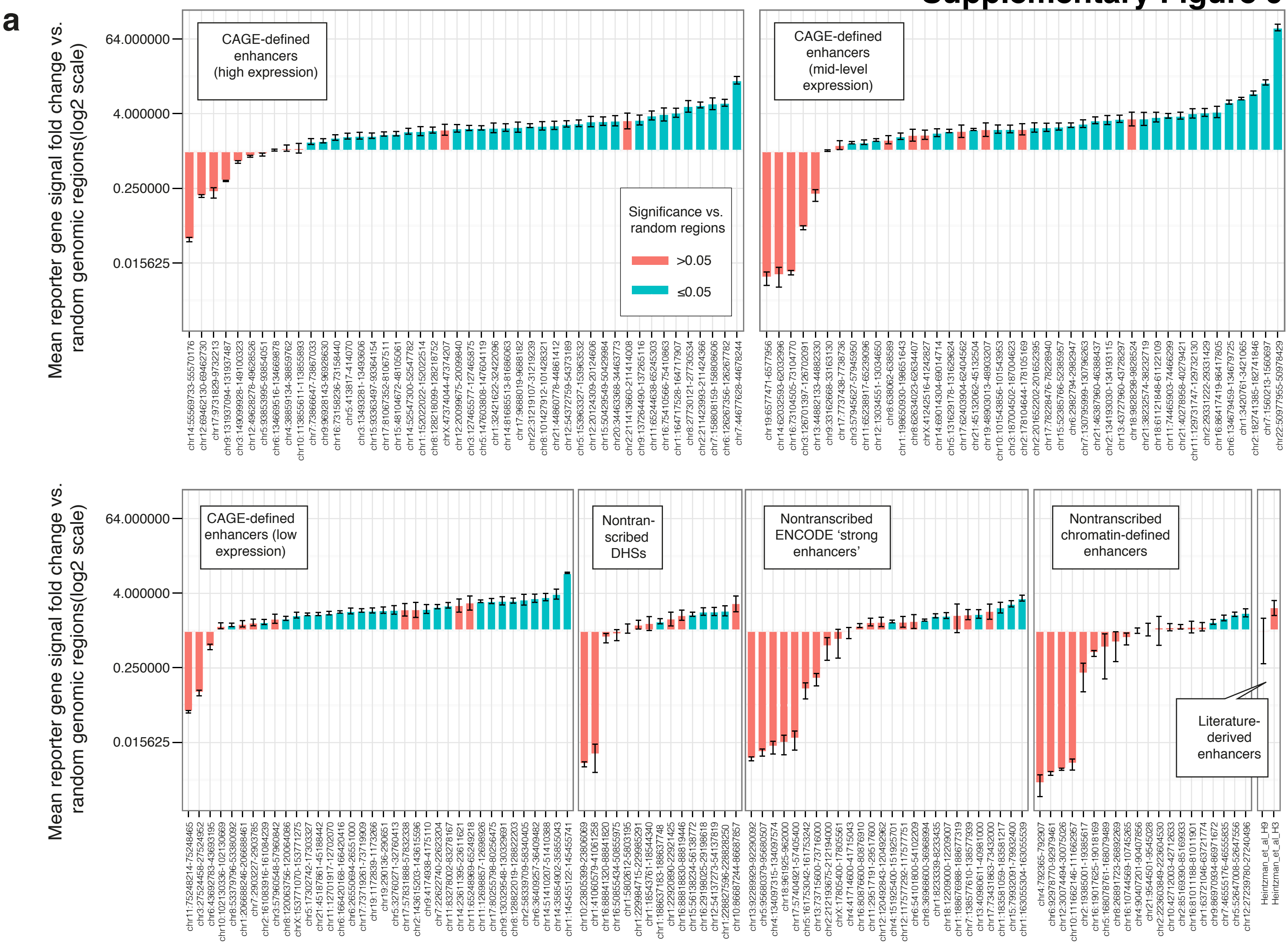
b, Evolutionary conservation and LTR enrichment in differentially expressed enhancer sets
Vertical axis shows the mean PhastCons score (green) and the fraction of enhancers in each set that is overlapped by LTRs (pink). Horizontal axis shows the +/- 1kb genomic region around enhancers, centered on the midpoint. All enhancer sets have similar PhastCons enrichment and LTR depletion, with the exception of neural stem cells, which have low conservation and high LTR overlap. Panel IDs refer to ontology IDs explained in Supplementary Tables 10 and 11.

Supplementary Figure 8

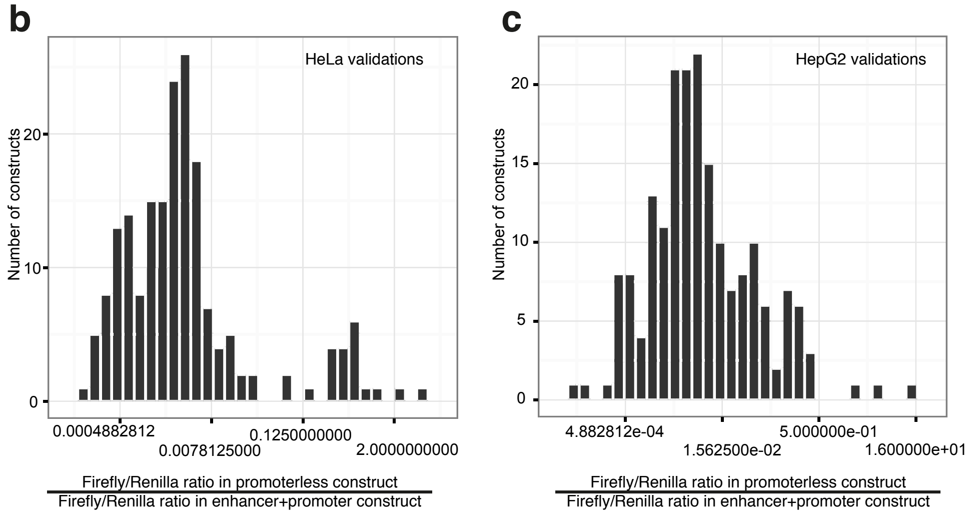
Comparison between enhancers and TSSs (broken up by CpG island (CGI) overlap (+/- 300 bp))
Vertical axis shows the fraction of regions (enhancers or RefSeq TSSs) that are overlapped by pooled ENCODE ChIP-seq peaks. Horizontal axis shows the +/- 5000 bp region around the enhancer center or the RefSeq TSS.

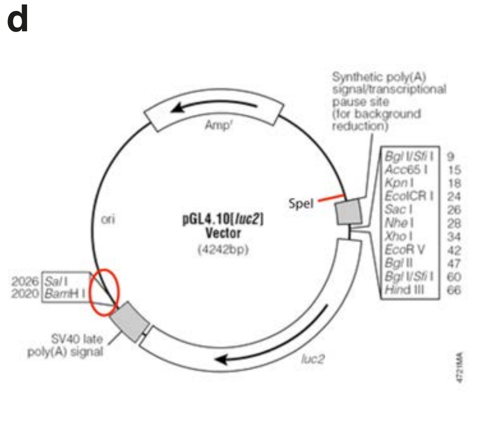


Supplementary Figure 9



a, HeLa enhancer reporter assays for enhancer candidate regions defined by CAGE, chromatin or DHS signatures. The vertical axis shows the average Firefly/Renilla signal of respective regions (~ +/- 250 bp from their center) divided by the corresponding average signal from 8 randomly selected genomic regions. Error bars indicate the standard error of the mean over three independent transfections. Each bar corresponds to one predicted enhancer, split up by prediction method and, in the case of CAGE, also expression strength. Bars are sorted by vertical axis values within each group. Bar color indicates the results of a t-test vs. random genomic regions.

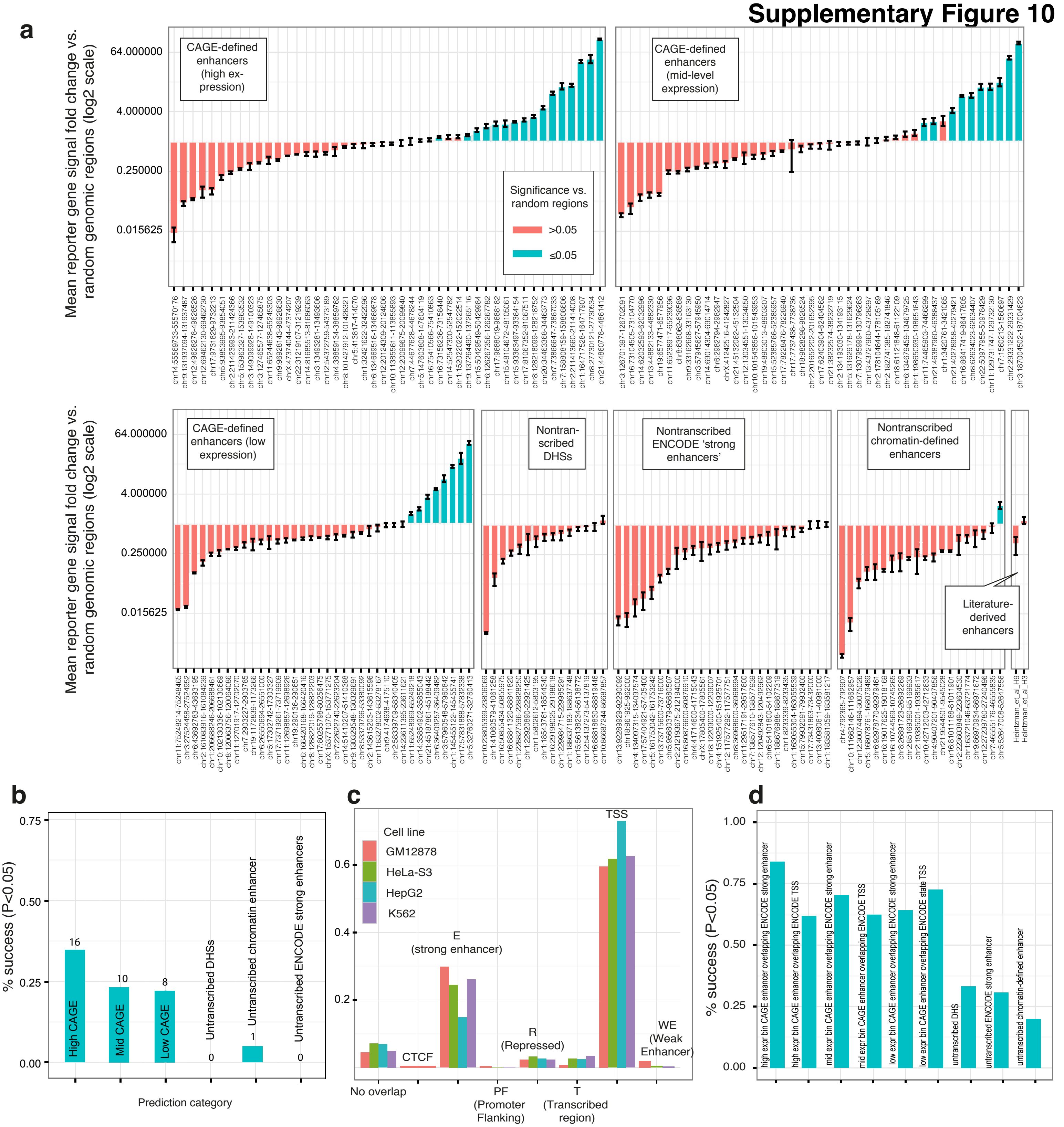




b-c, Influence of enhancer read-through. Histograms showing influence of read-through from enhancer transcription vs. the full construct, measured by calculating the ratio between pairs of promoter less and full (enhancer + basal promoter) constructs. All values are normalized by Firefly-/Renilla ratios. **b** shows read-through for HeLa cell validations, **c** for HepG2. In both cases, the median is below 3%.

d, pGL4.10[*luc2*] Vector. Large-scale in vitro validations on randomly selected enhancers were performed using Firefly/Renilla luciferase reporter plasmids with enhancer sequences cloned upstream of an EF1α basal promoter separated by a synthetic polyA signal/transcriptional pause site in a modified pGL4.10 (Promega) vector. Full details are provided in the Supplementary Methods online.

Supplementary Figure 10



a, HepG2 enhancer reporter assays for enhancer candidate regions defined by CAGE, chromatin or DHS signatures. The vertical axis shows the average Firefly/Renilla signal of respective region (~ +/- 250 bp from its center) divided by the corresponding average signal from 8 randomly selected genomic regions. Error bars indicate the standard error of the mean over three independent transfections. Each bar corresponds to one predicted enhancer, split up by prediction method and, in the case of CAGE, also expression strength. Bars are sorted by vertical axis values within each group. Bar color indicates the results of a t-test vs. random genomic regions.

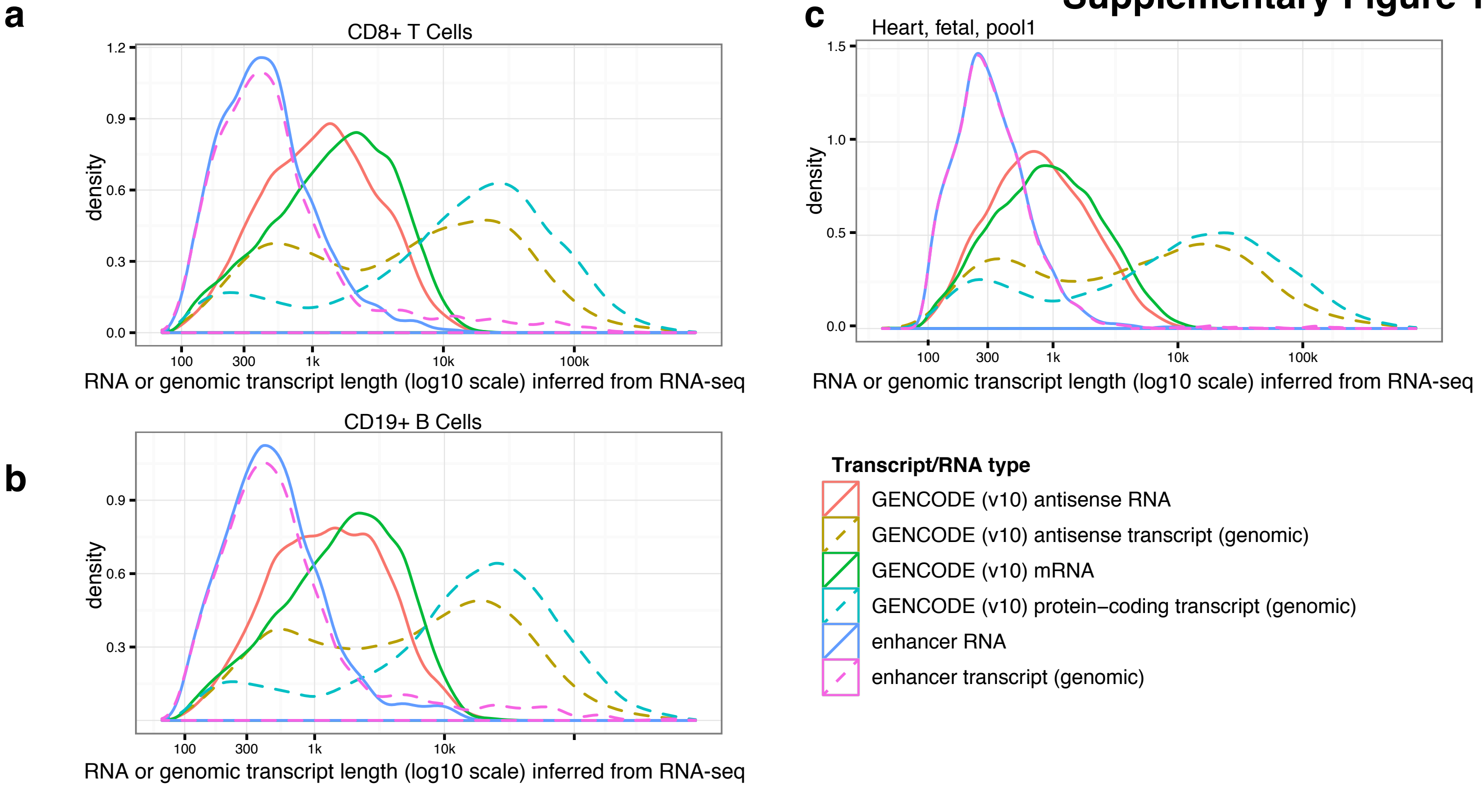
b, Summary of success rate of the HepG2 validations of selected groups in panel a. Vertical axis shows the percentage successes (t-test, P<0.05 vs. random regions). Numbers within bars indicate the number of successes. The success rates are much lower than in HeLa overall since the enhancers were selected based on HeLa expression; however, the general trend that transcribed enhancers have higher validation rates than untranscribed ones is still evident.

c, Overlap between CAGE-defined enhancers and ENCODE state segmentations. Fraction of CAGE-defined enhancers expressed in ENCOCE cell lines by at least 2 out of 3 replicates (in matching cells) that overlap ENCODE segmentation states (combined results of Segway and ChromHMM).

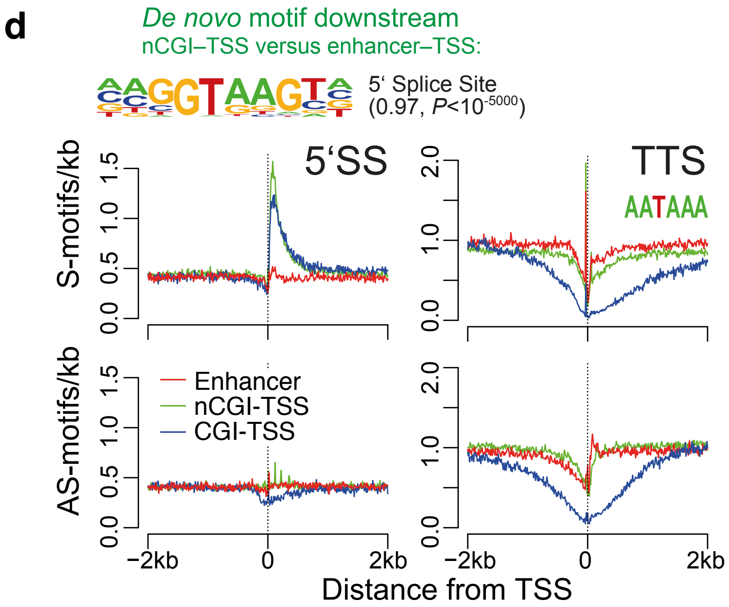
d, Summary of success rate of the HeLa validations as a function of ENCODE state overlap

Succes rates are shown as in Figure 1c, but with CAGE-predicted enhancers broken up according to overlap with ENCODE 'TSS' and 'strong enhancer' predictions. Vertical axis shows the percentage successes (t-test, P<0.05 vs. random regions).

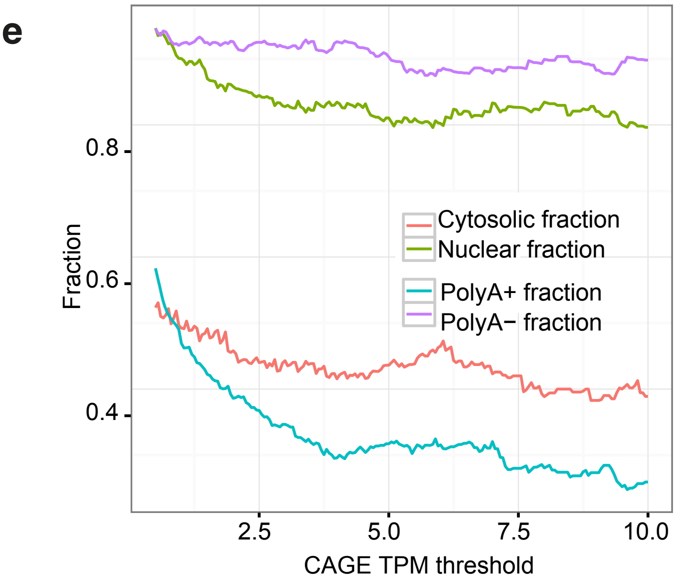
Supplementary Figure 11



a-c, Lengths of transcripts from enhancer and mRNA TSSs
Distributions of genomic lengths of Cufflinks transcripts (dotted lines) and also derived nucleotide lengths of (intron-less) RNAs (solid lines), inferred from RNA-seq (total RNA) in CD8+ T cells (a), CD19+ B cells (b) and fetal heart tissue (c) whose 5' ends originate from CAGE-defined enhancers or promoters of GENCODE (v10) protein-coding gene transcripts. RNA-seq was run on the same samples analyzed with CAGE within FANTOM5. Enhancer RNAs are clearly shorter than mRNAs. Note also the clear separation between RNA length and genomic transcript length for protein-coding genes, which is in strong contrast with enhancer RNAs that are most often unspliced.

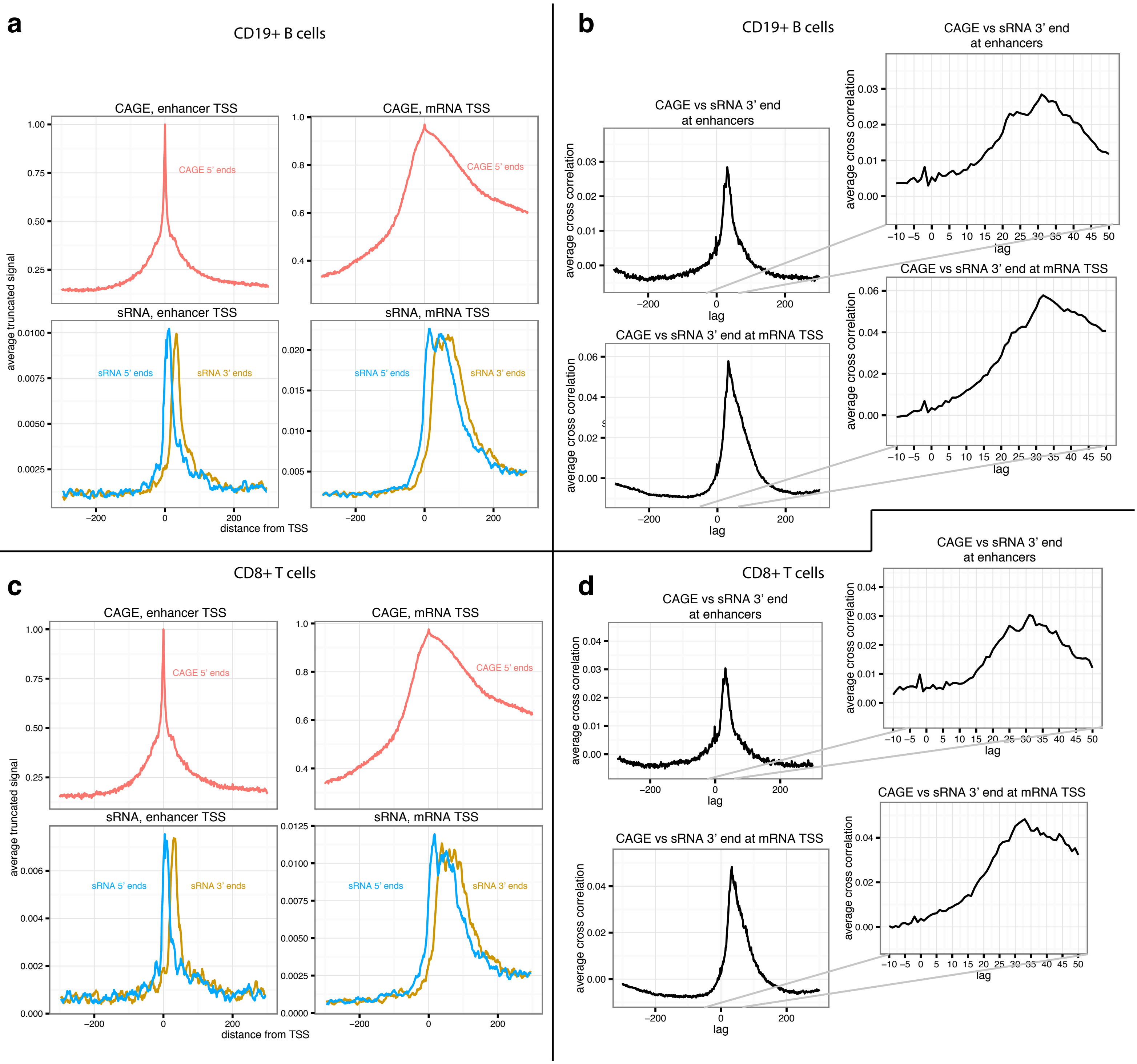


d, Over-representation of RNA processing motifs around enhancers and mRNA TSSs.
De novo motif finding identifies the 5' splice site motif (5'SS) as over-represented around RefSeq TSSs vs enhancer TSSs (top logo). Counting the number of 5'SS motif occurrences around respective TSSs shows that this over-representation is located in the first 100 bp of mRNA transcripts but is not present around enhancers (left panels). Conversely, the transcription termination motif is depleted downstream of RefSeq TSSs but not enhancer TSSs (right panels).



e, RNA fractionation and polyA selection show that most enhancer RNAs are nuclear and non-polyadenylated
Using ENCODE HeLa-S3 CAGE data, we calculated the number (vertical axis) of HeLa-S3 enhancers (expressed in at least 2 out of 3 FANTOM5 HeLa-S3 replicates) whose RNAs were nuclear, cytosolic, poly-adenylated and non-polyadenylated at increasing TPM threshold (horizontal axis). Fractions were then calculated. The nuclear fraction was calculated as the fraction of enhancers with nuclear or cytoslic RNAs that were nuclear and similarly for cytoslic fraction, PolyA+ fraction and PolyA- fraction.
Nuclear fraction = #nuclar / #(cytosolic or nuclear)
Cytosolic fraction = #cytosolic / #(cytosolic or nuclear)
PolyA+ fraction = #polyA+ / #(polyA+ or polyA-)23
PolyA- fraction = #polyA- / #(polyA+ or polyA-)

Supplementary Figure 12



Locations of small RNAs (sRNAs) at enhancer and mRNA TSSs

a, Distributions of CD19+ B cell CAGE 5' ends (corresponding to capped >100 nt long RNAs) centered on the CAGE sense strand summits of RefSeq mRNA TSSs (upper right) and forward strand summits of enhancer TSSs (upper left), and CD19+ B cell RNA (uncapped 18-30 nt RNAs) 3' and 5' ends in the same locations (lower panels). The vertical axis shows CAGE or sRNA tags/bp, but only 1 tag from any unique nucleotide is counted to avoid undue influence of outliers (truncated signal).

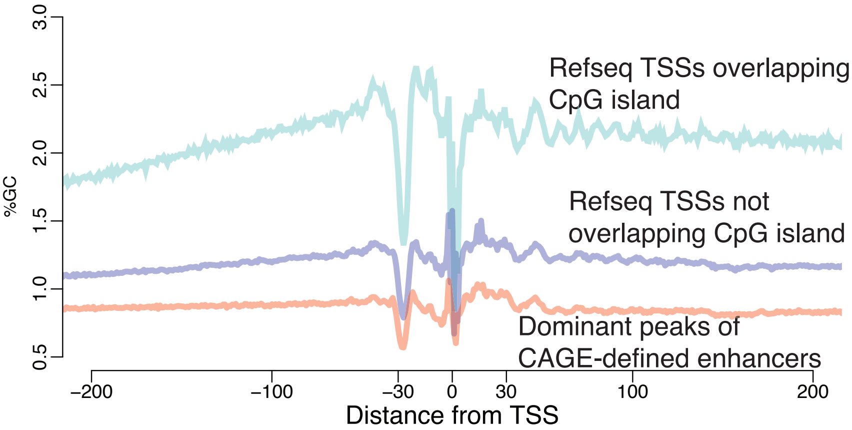
b, Average cross-correlation between CAGE 5' ends and sRNA 3' signal as a function of the shift (lag) between the two datasets in bp, for enhancer CAGE summits (top panels) and RefSeq TSS CAGE summits (lower panels) The right panels show zoom-ins of lags between -10 and +50. For both enhancer and RefSeq TSSs, the most common lag is between 30 and 35, consistent with previously described TSS-associated small RNAs (e.g. Valen *et al.*, Taft *et al.*). That means that both enhancer TSSs and RefSeq TSSs have the same properties in terms of emitting small RNAs.

c, As in panel a, but for CD8+ T cells

d, As in panel b, but for CD8+ T cells

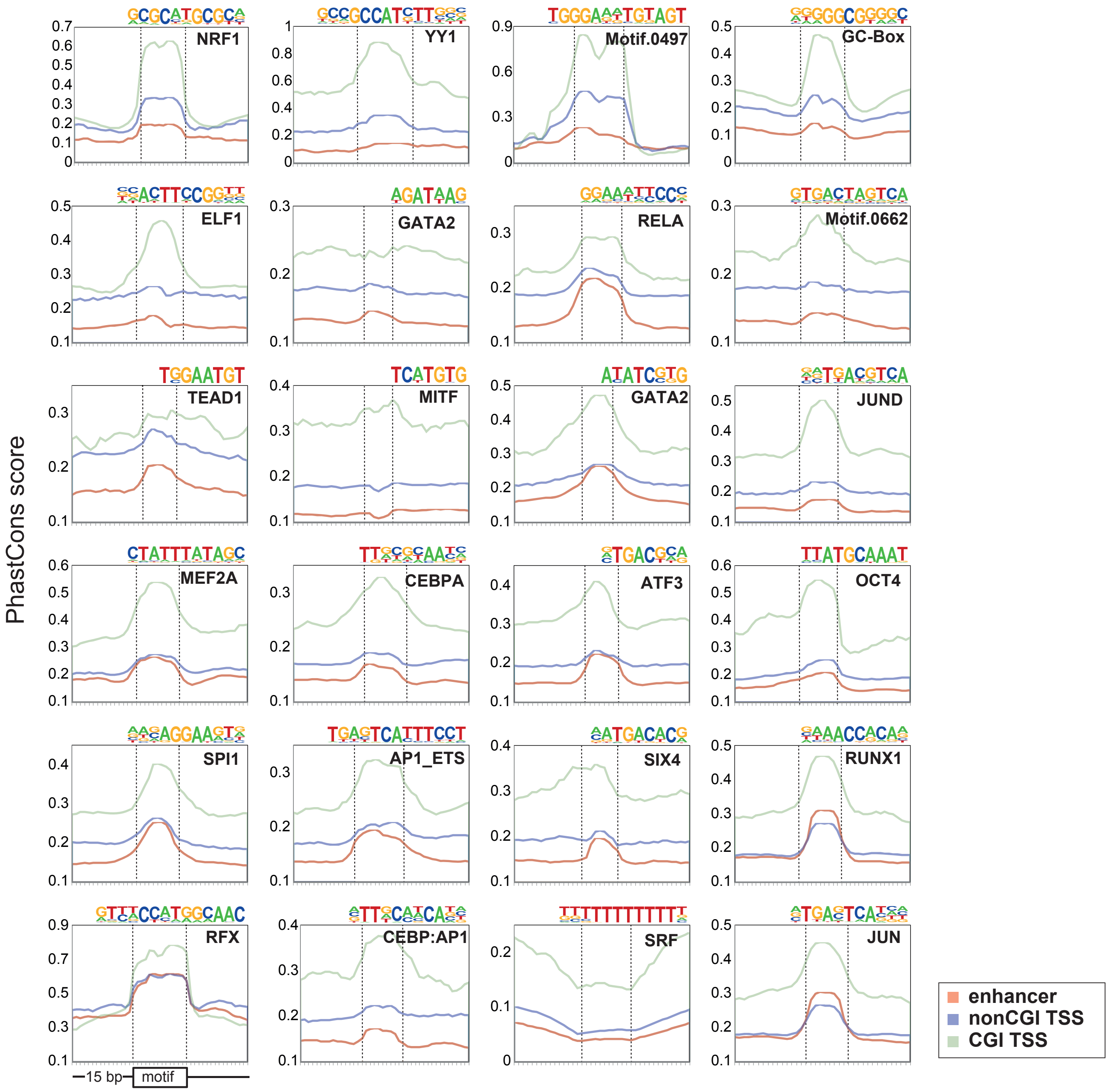
Supplementary Figure 13

a



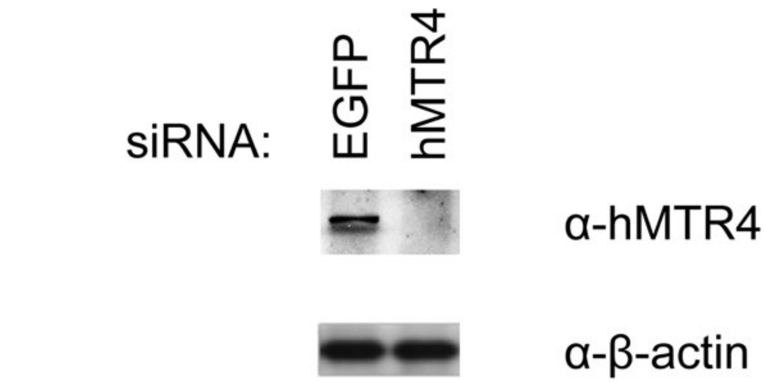
a, Similarity of GC profiles around mRNA TSSs and enhancer TSSs
Mean GC dinucleotide content anchored on RefSeq enhancers overlapping/not overlapping CpG islands, and the dominant peaks of enhancer CAGE tags.

b

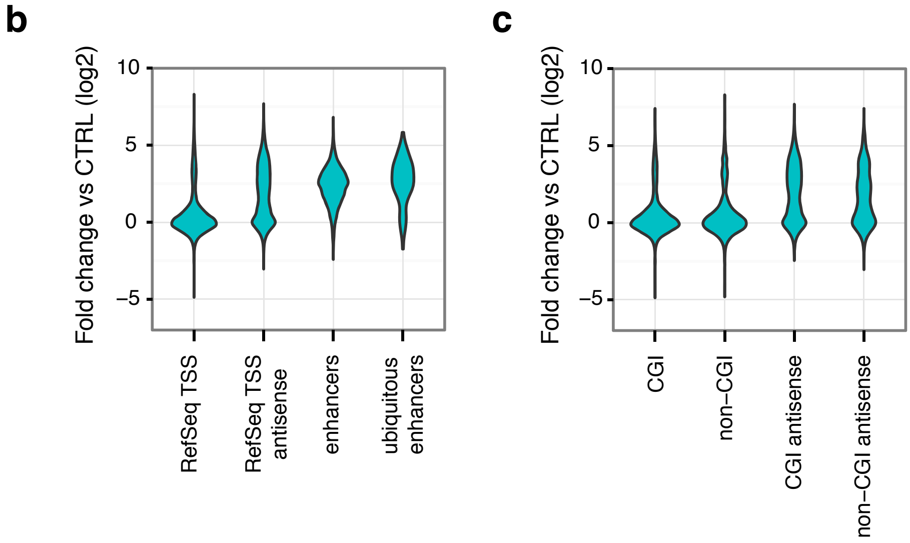


b, Preferential conservation of *de novo* motifs found in enhancers and mRNA TSS regions
Each panel shows PhastCons conservation of the genomic regions around motif hits in enhancers, nonCGI RefSeq TSSs and CpG-overlapping RefSeq TSSs. Transcription factor names below sequence logos indicate closest known motif. Note that JUN, RUNX and RFX are equally or more conserved in CAGE-defined enhancers than in RefSeq TSSs not overlapping CpG islands.

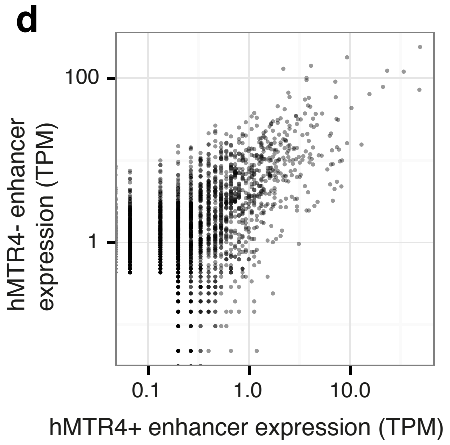
Supplementary Figure 14



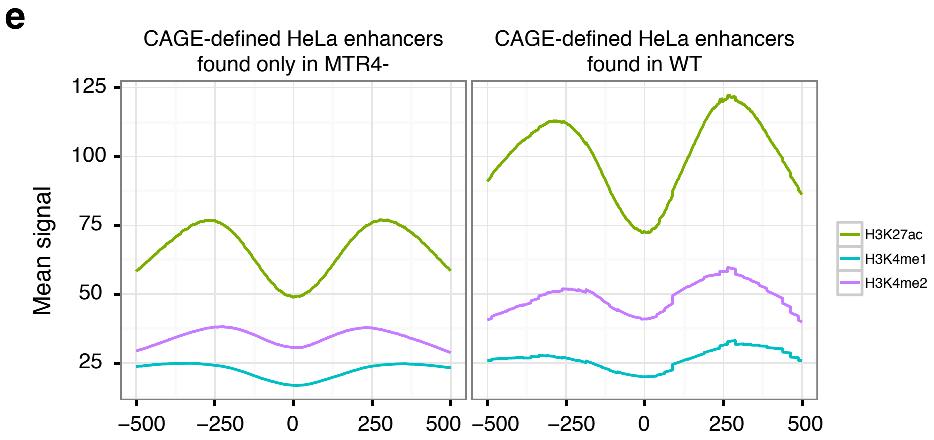
a, Western blotting verification of siRNA mediated depletion of hMTR4.
siRNA against EGFP was used as a knock-down control and β-actin served as a loading control.



b-c, Overall effects of hMTR4 depletion
Distributions of CAGE expression fold changes vs. control (vertical axis) when depleting the exosome (hMTR4 gene) at
b, RefSeq TSSs (sense and antisense), HeLa expressed enhancers and FANTOM5 ubiquitous enhancers and
c, RefSeq TSSs (sense and antisense) overlapping CpG islands (CGI) or not (non-CGI).



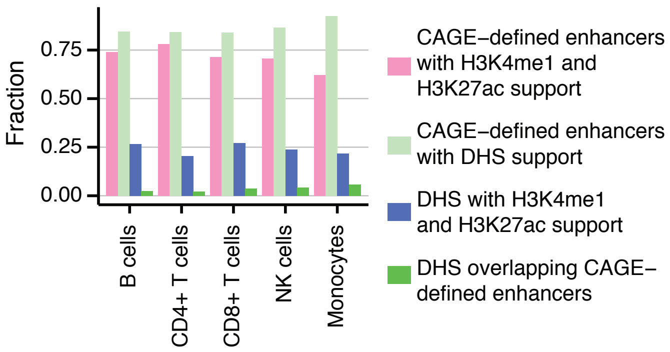
d, Correlation between hMTR4- and hMTR4+ expression
The number of HeLa CAGE tags (TPMs) within enhancers after depletion of hMTR4 (vertical axis, log10 scale) vs. WT control HeLa (horizontal axis, log10 scale).



e, Chromatin features of hMTR4- enhancers
Histone modification strengths at HeLa enhancers detected in WT HeLa (right) and only after MTR4 depletion (left). All data are from HeLa cells.

Supplementary Figure 15

a

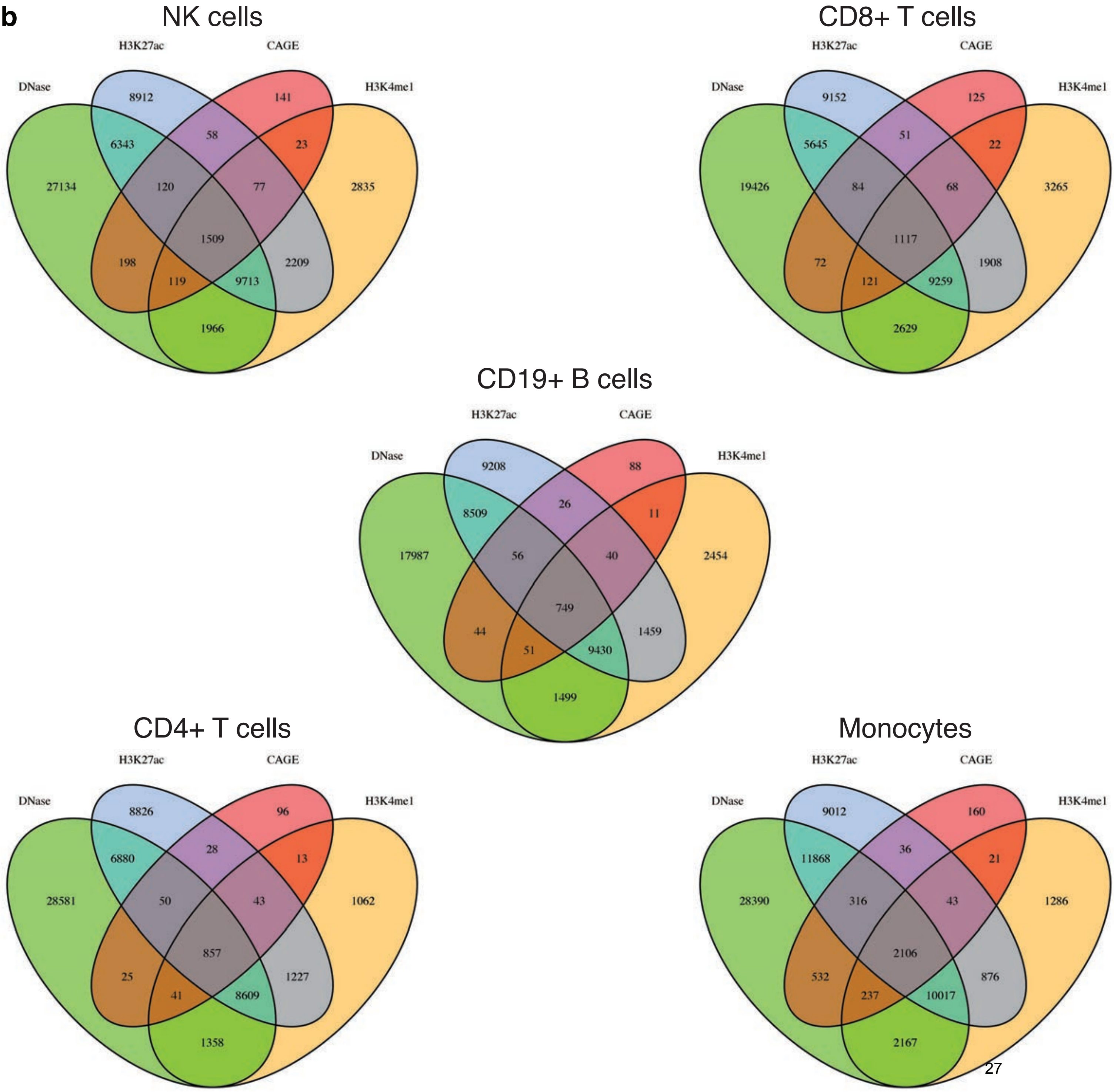


Overlap between CAGE, DHSs and chromatin marks in blood cells The majority of regions characterized by bidirectional CAGE peaks overlap with sites that show the chromatin marks H3K27ac and H3K4me1 and DNase I hypersensitivity (DHS). In contrast, most DHSs do not show enhancer-associated chromatin marks nor do they overlap with bidirectional CAGE tags.

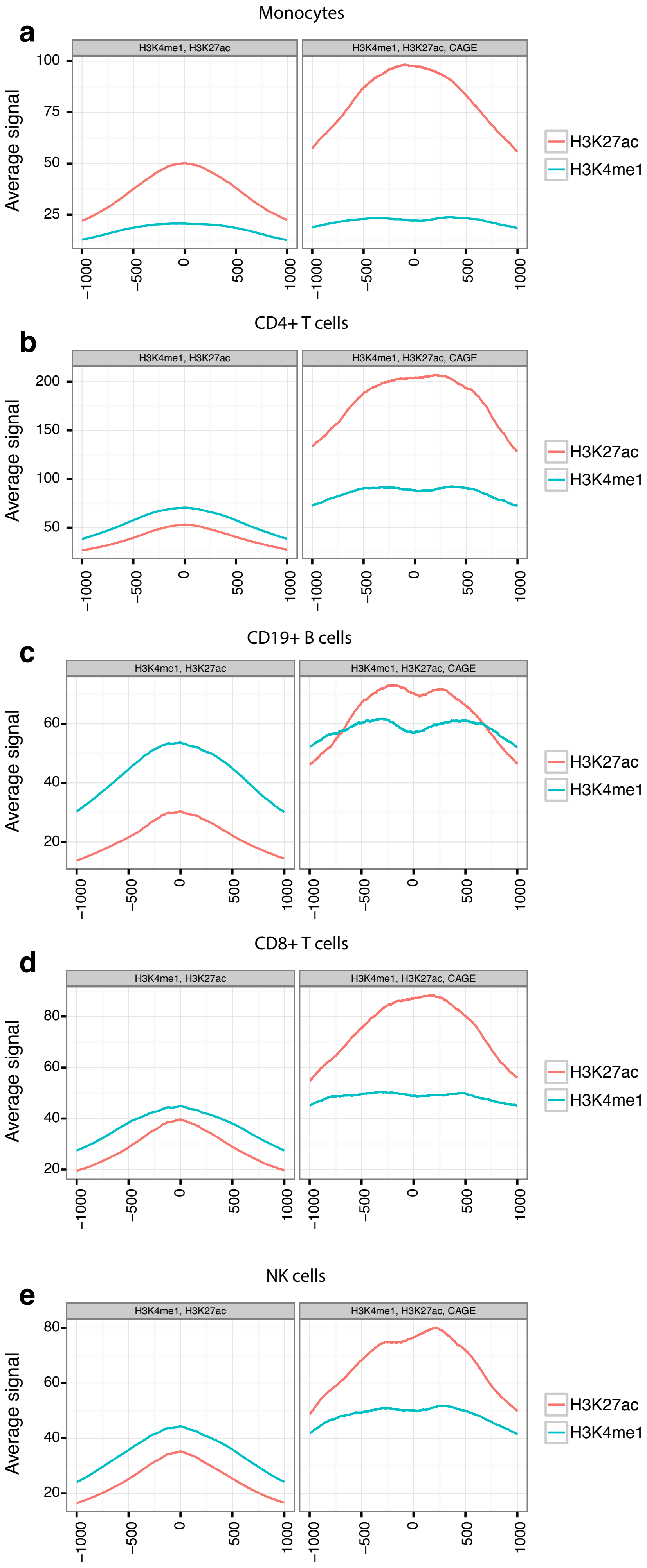
a, Fraction of bidirectional CAGE pairs or DHSs from a given blood cell type supported by H3K27ac and H3K4me1 or DNase/CAGE data from the same cells, filtered for known TSSs, exons and ncRNAs.
b, The Venn diagrams depict the overlap of these four characteristics (CAGE, H3K4me1, H3K27ac, DHS) for five blood cell types.

Peak-calling of DHSs and of H3K27ac and H3K4me1 ChIP-seq signals was performed on pooled data for the five cell types to define peak region boundaries. Subsequent signal quantification and testing for significant signal above background was done per cell type (see Supplementary Methods for details).

b



Supplementary Figure 16

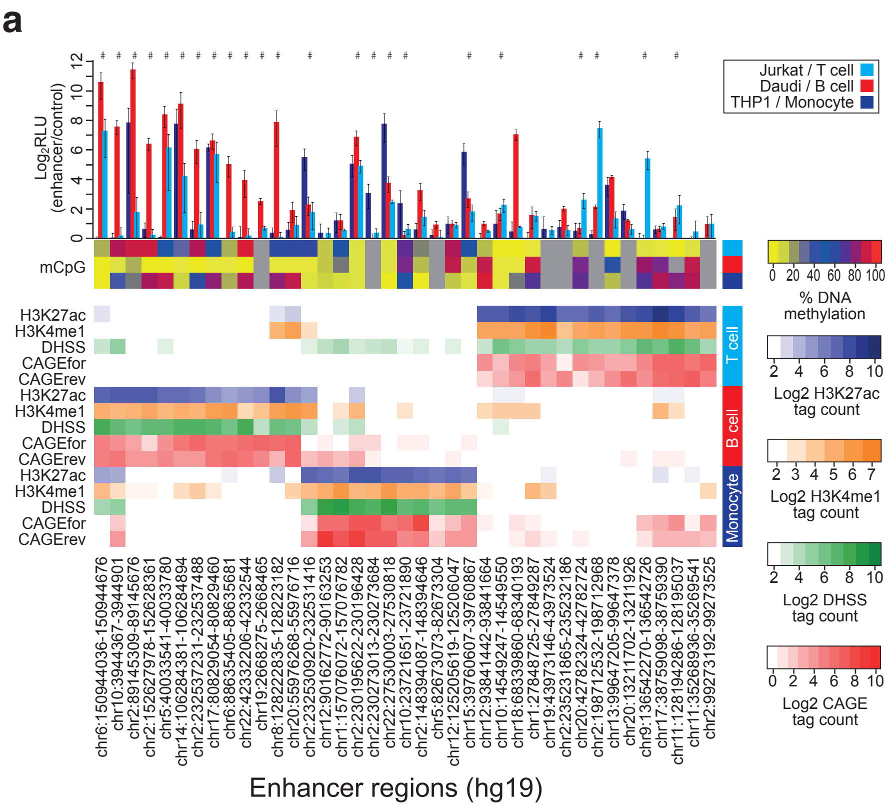


Histone mark signals of enhancers detected by H3K4me1 and H3K27ac, with or without CAGE.

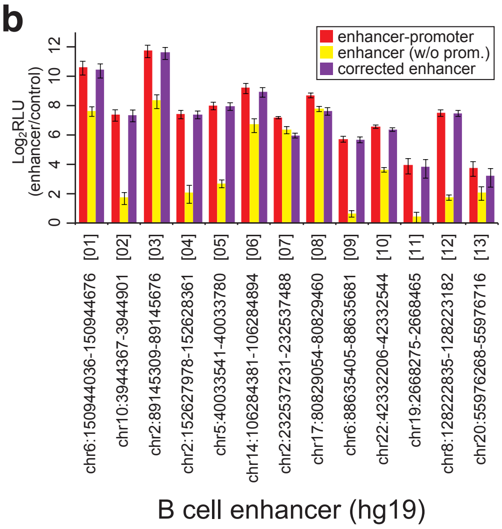
For each blood cell type, we detected enhancers based on H3K4me1 and H3K27ac (see methods) and CAGE independently. We then plotted the average ChIP-seq signal of H3K4me1 and H3K27ac around the center of the enhancers that were detected by only H3K4me1 and H3K27ac (left panels), or by both H3K4me1 and H3K27ac as well as CAGE (right panels) in monocytes (a), CD4+ T cells (b), CD19+ B cells (c), CD8+ T cells (d), and natural killer (NK) cells (e).

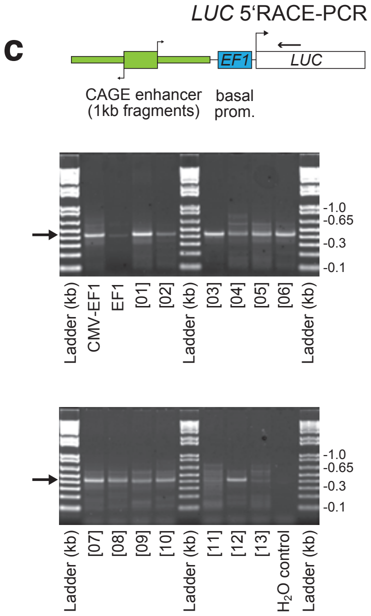
Enhancers only detected by H3K4me1 and H3K27ac have the lowest average ChIP-seq signals. In particular, H3K27ac show big changes between untranscribed and transcribed enhancers. Enhancers detected by both approaches consistently have high enhancer-associated histone modification signals.

Supplementary Figure 17



a, In vitro validation of enhancers detected in blood cells.
Validation of the activity of 39 CAGE-defined blood enhancers in three corresponding cell line models. Bar plots show relative luciferase signals (putative enhancer plus promoter versus EF1 promoter alone), in cell lines corresponding to T cells (Jurkat), monocytes (THP-1), and B cells (Daudi). Relative values are means ± SD (n≥3). Each column represents one enhancer trial in all three cells. Below, methylation of CpGs, CAGE on both strands, DHS, H3K4me1 and H3K27ac signals are shown as a heat map. Enhancer constructs showing at least 4-fold higher activity (compared to the promoter alone) are marked with a hash.



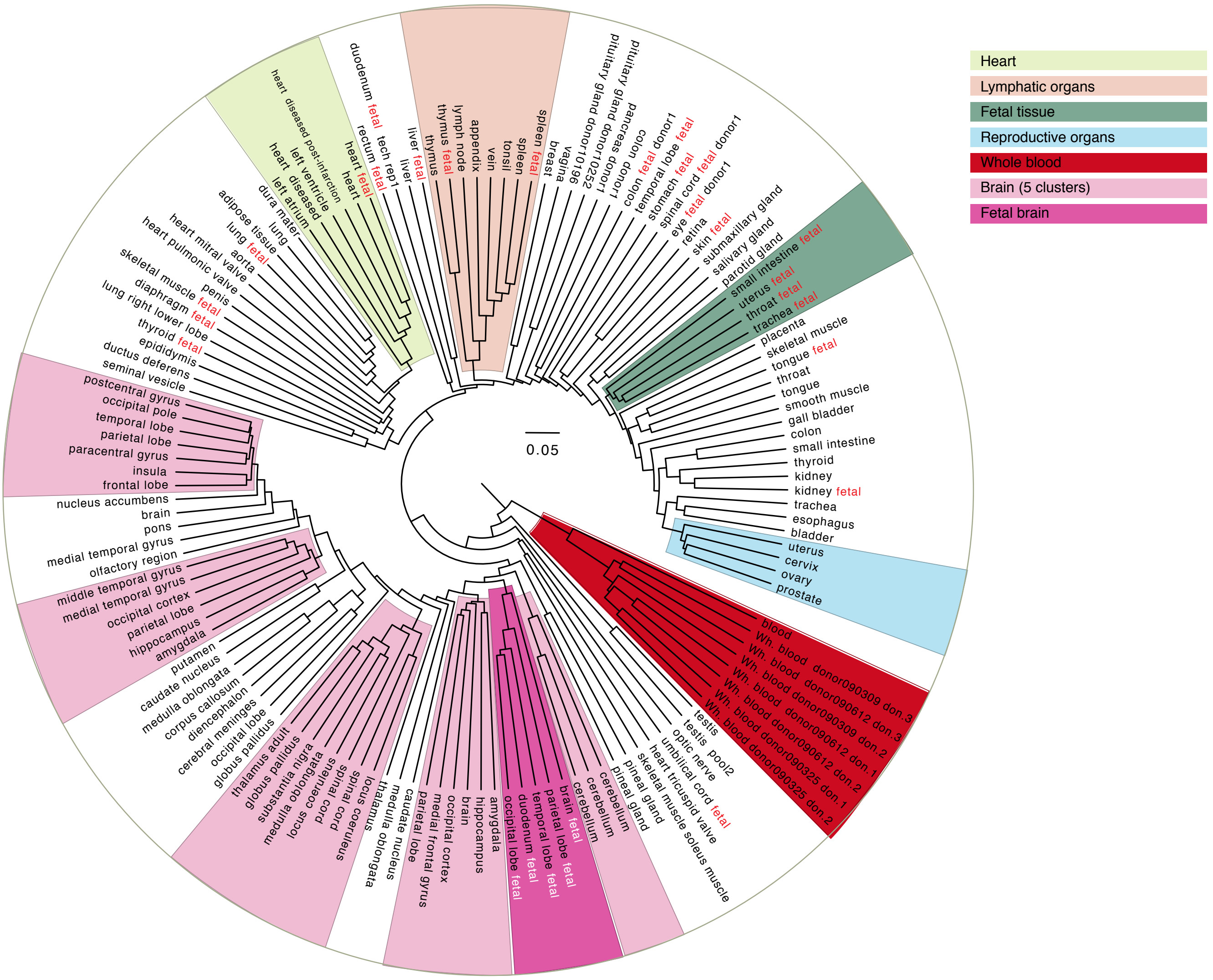


b, Enhancer versus promoter induced reporter gene transcription.
To distinguish enhancer read-through from enhanced promoter-initiated luciferase transcription, B cell enhancer constructs were compared with corresponding constructs lacking the EF1 promoter. Bar plots show relative luciferase signals (values are means ± SD; n=3) for complete enhancer-promoter constructs (red bars), promoter-deleted constructs (yellow bars), and corrected values (complete enhancer-promoter value minus promoter-deleted value, purple bar). Median read-through activity was less than 10% of the total signal, validating the true enhancer activity of these regions.

c, Reporter gene TSS-usage.
To determine TSS usage in enhancer-promoter reporter constructs, we performed 5'RACE PCR using a luciferase-specific primer. RACE products for control vectors (CMV-EF1 and EF1 alone), enhancer constructs [1-13] as well as a negative control (H_2O) were separated on a 2% agarose gel. The major TSS in the majority of samples corresponds to the expected TSS downstream of the EF1 promoter, suggesting that these assays measure 'true' enhancer activity.

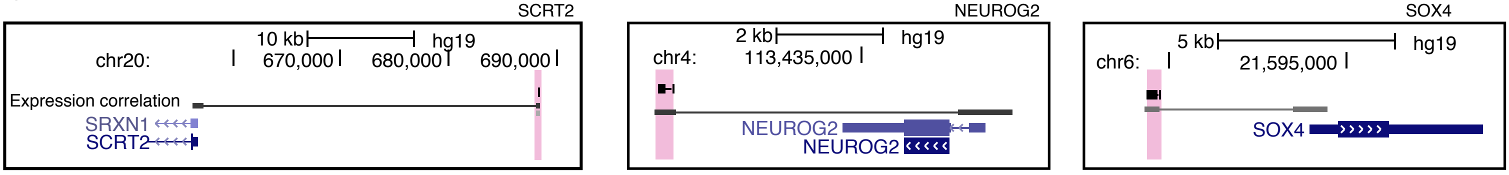
Supplementary Figure 18

a **Clustering of tissue/organ libraries by enhancer expression**



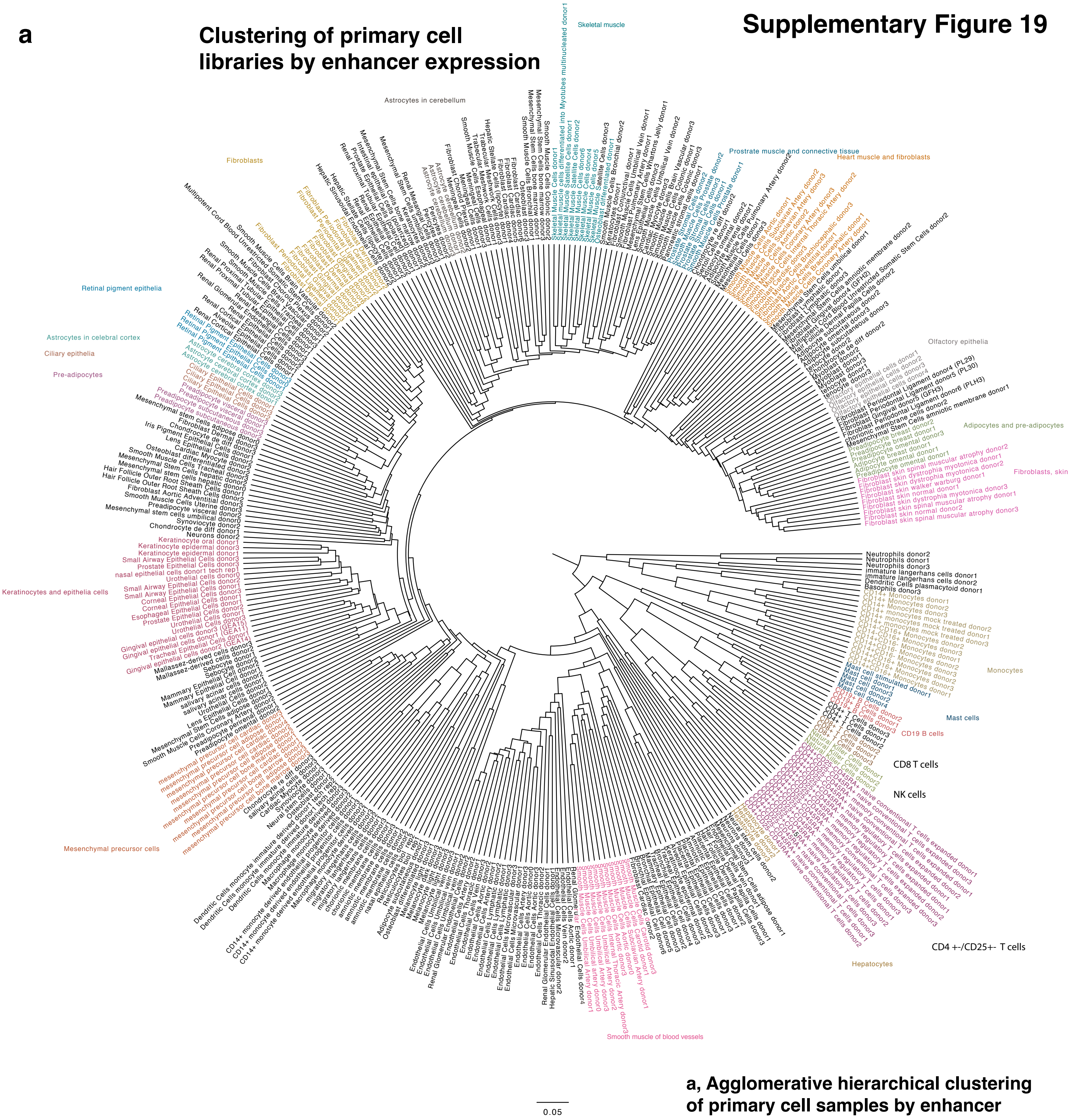
a, Agglomerative hierarchical clustering of tissue samples by enhancer expression
Expansion of the small hierarchical tree in Figure 3c, where actual tissue libraries are shown as leaves. Fetal libraries are indicated by red or white highlights; libraries without highlights are from adult humans.

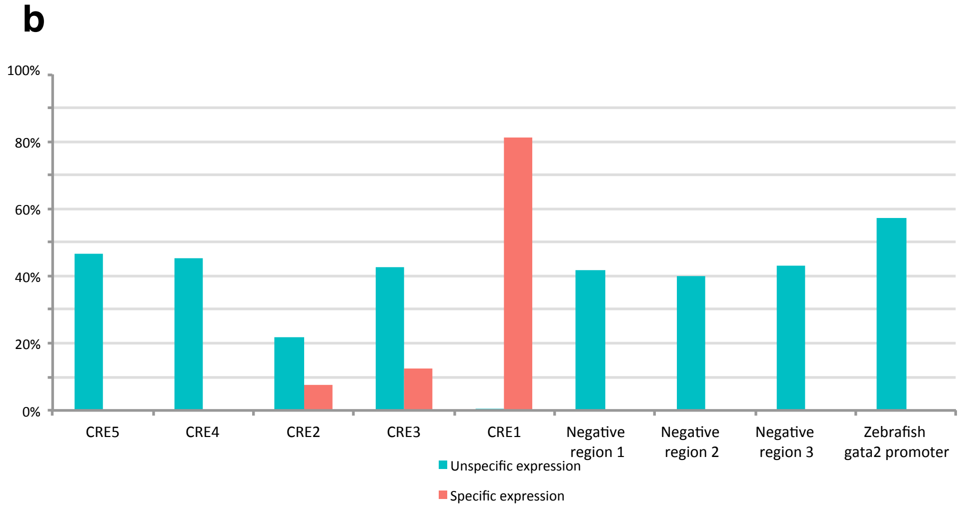
b



b, Fetal brain enhancer genome landscapes
Three examples of enhancers (pink bars) differentially expressed in the brain fetal vs. brain adult subtrees, and the closest genes of these enhancers, which are all known neural development regulators. Grey lines indicate significant expression correlations between TSSs and enhancers, suggesting interactions (see main text). Gene models are from the UCSC gene track.

Supplementary Figure 19

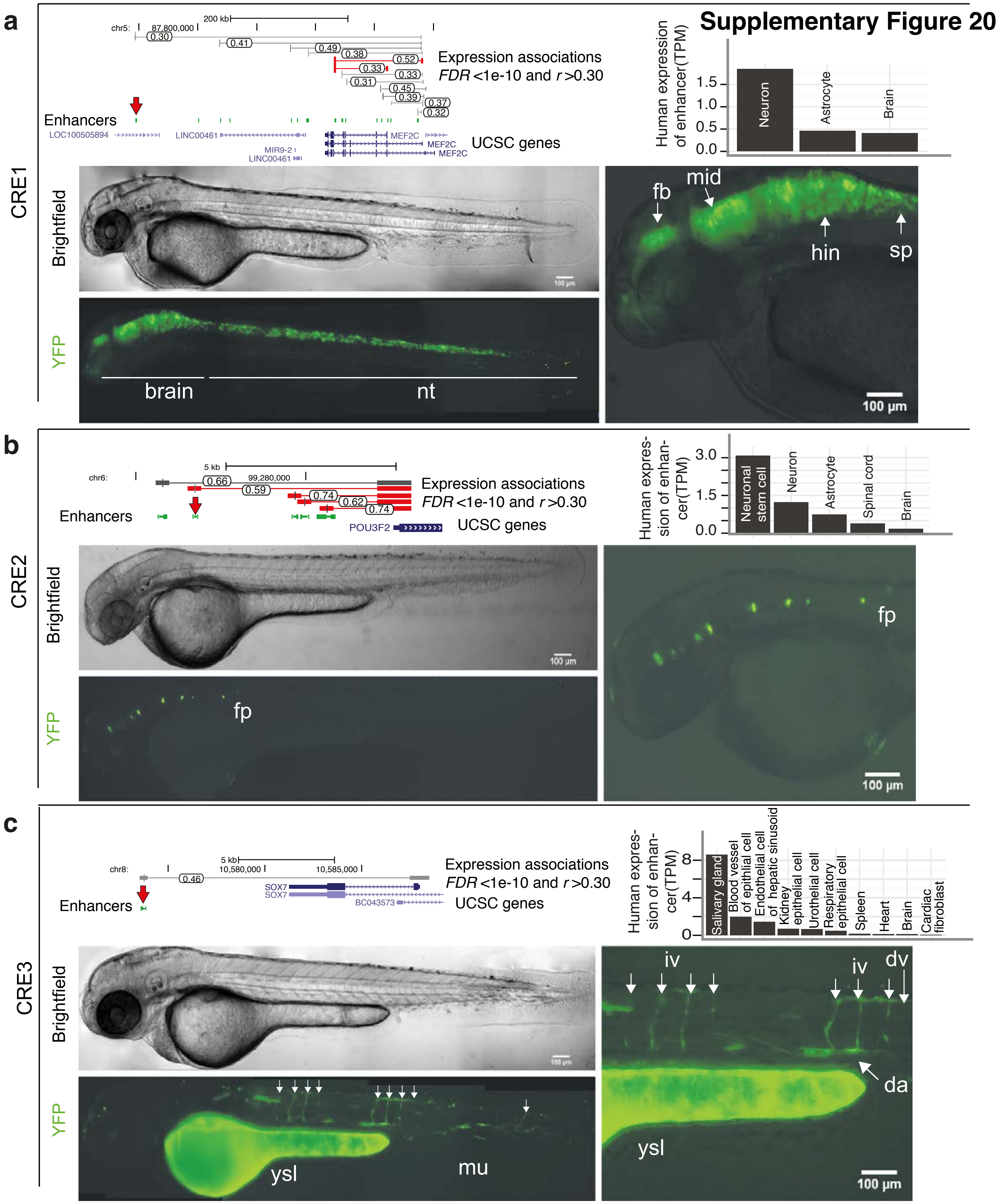

a
Clustering of primary cell libraries by enhancer expression
Skeletal muscle
Astrocytes in cerebellum
Fibroblasts
Prostrate muscle and connective tissue
Heart muscle and fibroblasts
Retinal pigment epithelia
Astrocytes in celebral cortex
Ciliary epithelia
Pre-adipocytes
Olfactory epithelia
Adipocytes and pre-adipocytes
Fibroblasts, skin
Keratinocytes and epithelia cells
Monocytes
Mast cells
CD19 B cells
CD8 T cells
NK cells
Mesenchymal precursor cells
CD4 +-/CD25+- T cells
Hepatocytes
Smooth muscle of blood vessels
0.05



a, Agglomerative hierarchical clustering of primary cell samples by enhancer expression
As in Supplementary Figure 18, but for primary cell libraries. Subtrees of anatomically or functionally related cell types are colored.

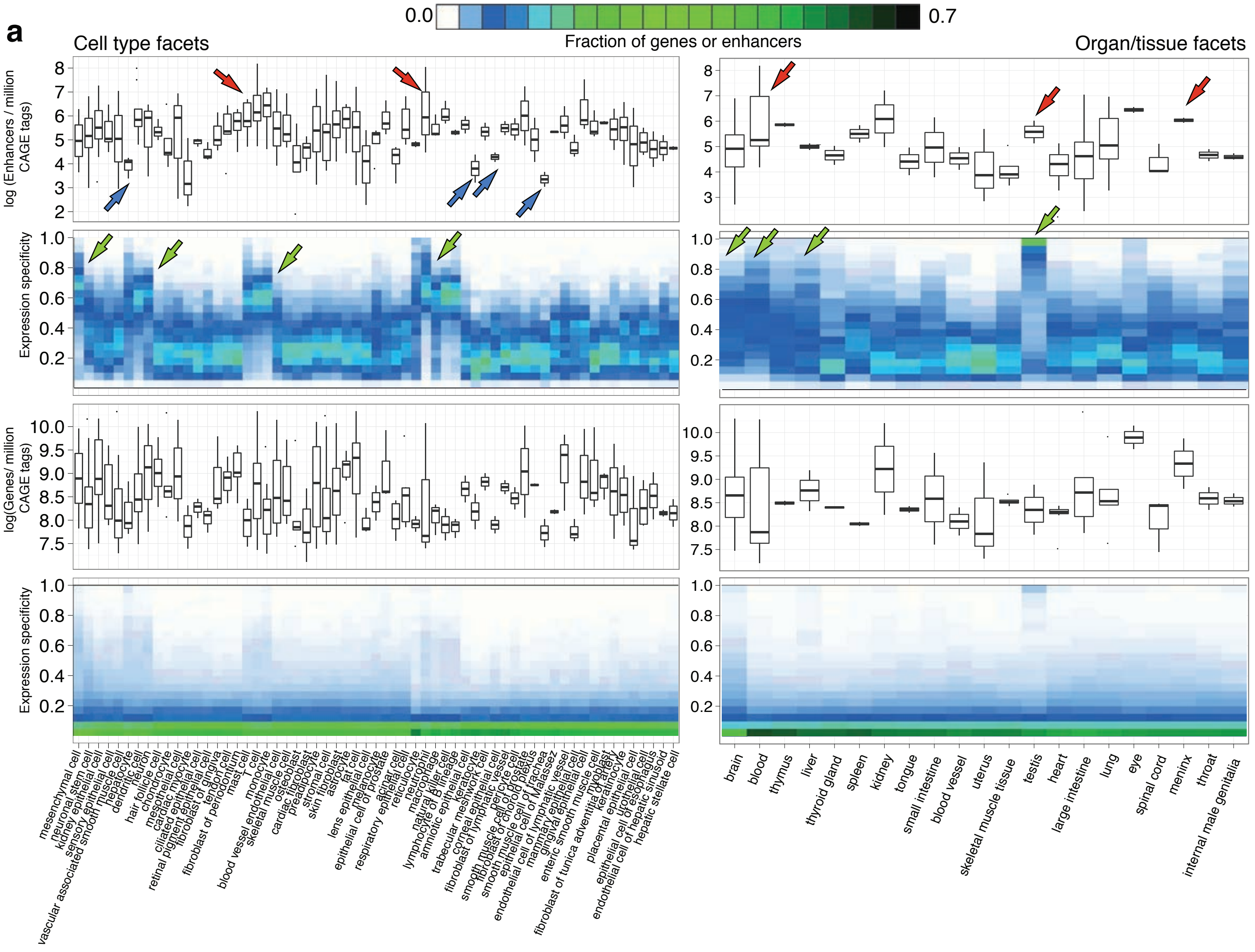
b, Enhancer activity of reporter constructs in mosaic transgenic zebrafish embryos
Percentage of zebrafish injected embryos at 48 hpf showing tissue specific expression (driven by human CRE1-3, red bars) and unspecific expression coming from the enhancer-less gata2 promoter containing vector (blue bars). For detailed expression patterns refer to Supplementary Table 9.

Supplementary Figure 20

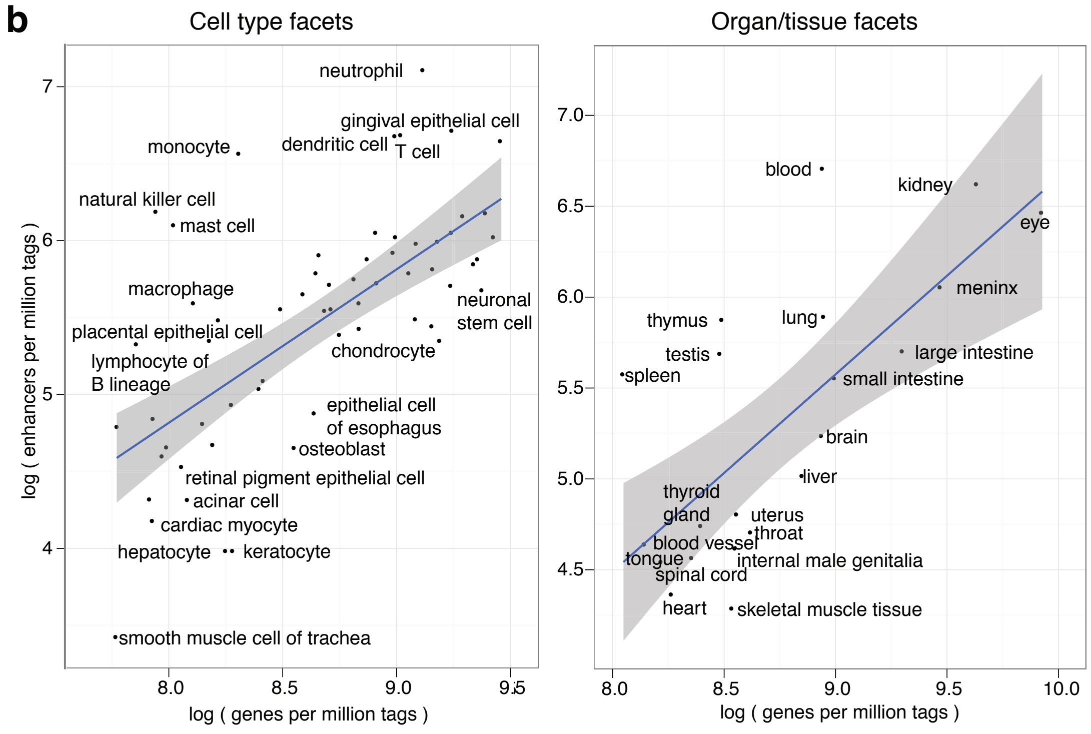


a-c) Validations of in vivo activity of CAGE-defined human enhancers CRE1-3 in zebrafish embryos at long-pec stage.
The image extends Figure 4 with UCSC browser sub-panels.
Each collection of panels show:
i) Top left, a UCSC genome browser image depicting the genome landscape around the validated enhancer (indicated by a red arrow) in human, with enhancer-TSS expression correlations shown as horizontal bars with the Pearson correlation coefficient in the middle circle. Red lines indicate interactions supported by ENCODE (RNAPII mediated) ChIA-PET interaction data.
ii) Top right: CAGE expression in TPM in human tissues/cell types for the enhancer. Note the correspondence between zebrafish and human enhancer usage/expression in the two subpanels below.
iii) Below: representative YFP and brightfield images of embryos injected with the human enhancer gata2 promoter reporter gene construct. Muscle (mu) and yolk syncytial layer (ysl) activities are background expression coming from the gata2 promoter-containing reporter construct. All images are lateral, head to the left. Right image shows YFP zoom-ins described below.
a, CRE1, ~230kb upstream of the MEFC2 gene, drives highly robust expression in the brain (brain) and neural tube (nt). Right panel gives zoom-in overlay image showing expression in the forebrain (fb), midbrain (mid), hindbrain (hin) and spinal cord (sp).
b, CRE2, 5kb upstream of the POU3F2 gene, is active in the floor plate (fp). Right panel is a zoom-in overlay image.
c, CRE3, 10kb upstream of the SOX7 gene TSS, shows specific expression in the vasculature (including intersegmental vessels (iv), dorsal vein (dv) and dorsal aorta (da)). Details are shown in the right panel.

Supplementary Figure 21

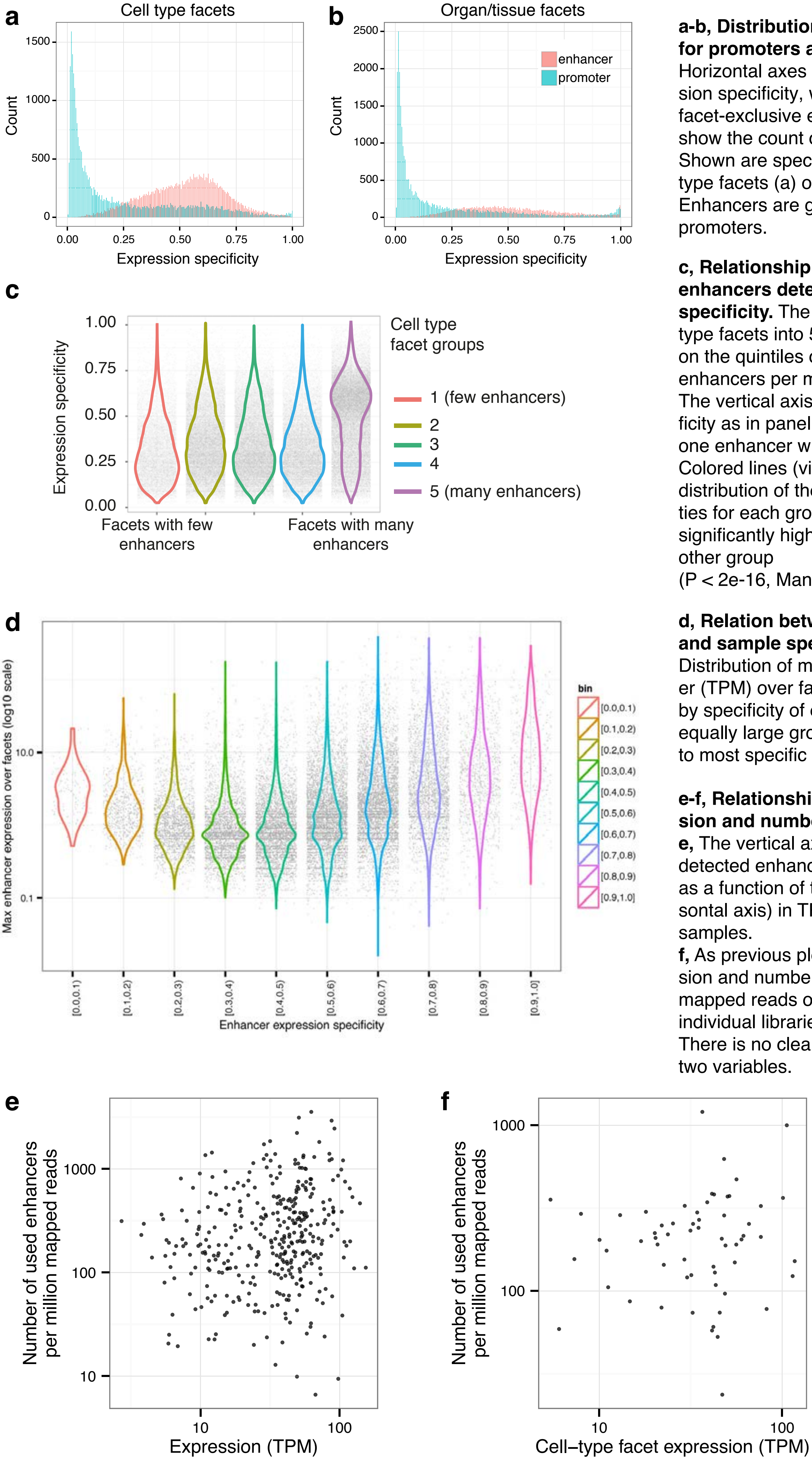


a, Relation between expression specificity and number of enhancers and genes
The number of detected enhancers per million mapped CAGE tags within each group of related libraries ("facets") is shown as box plots in the upper panel. The specificity of the enhancers is shown as a heatmap in the next panel below, where 1 indicates facet-exclusive expression. Colors show the fraction of expressed enhancers that are in each specificity range. Corresponding plots are shown for CAGE-detected RefSeq TSSs in the two lowest panel rows. Green arrows indicate samples that have enhancers with higher sample-specificity than others, blue and red arrows show samples with unusually high or low number of detected enhancers.



b, Correlation between gene and enhancer counts
Relationship between number of detected enhancers (vertical axes) and genes per million mapped tags (horizontal axes), in cell type (left panel) and tissue/organ (right panel) facets. Facets outside of the 90% confidence region (grey) of the regression line (blue) are labeled in the cell type plot; all facets are labeled in the organ/tissue plot.

Supplementary Figure 22



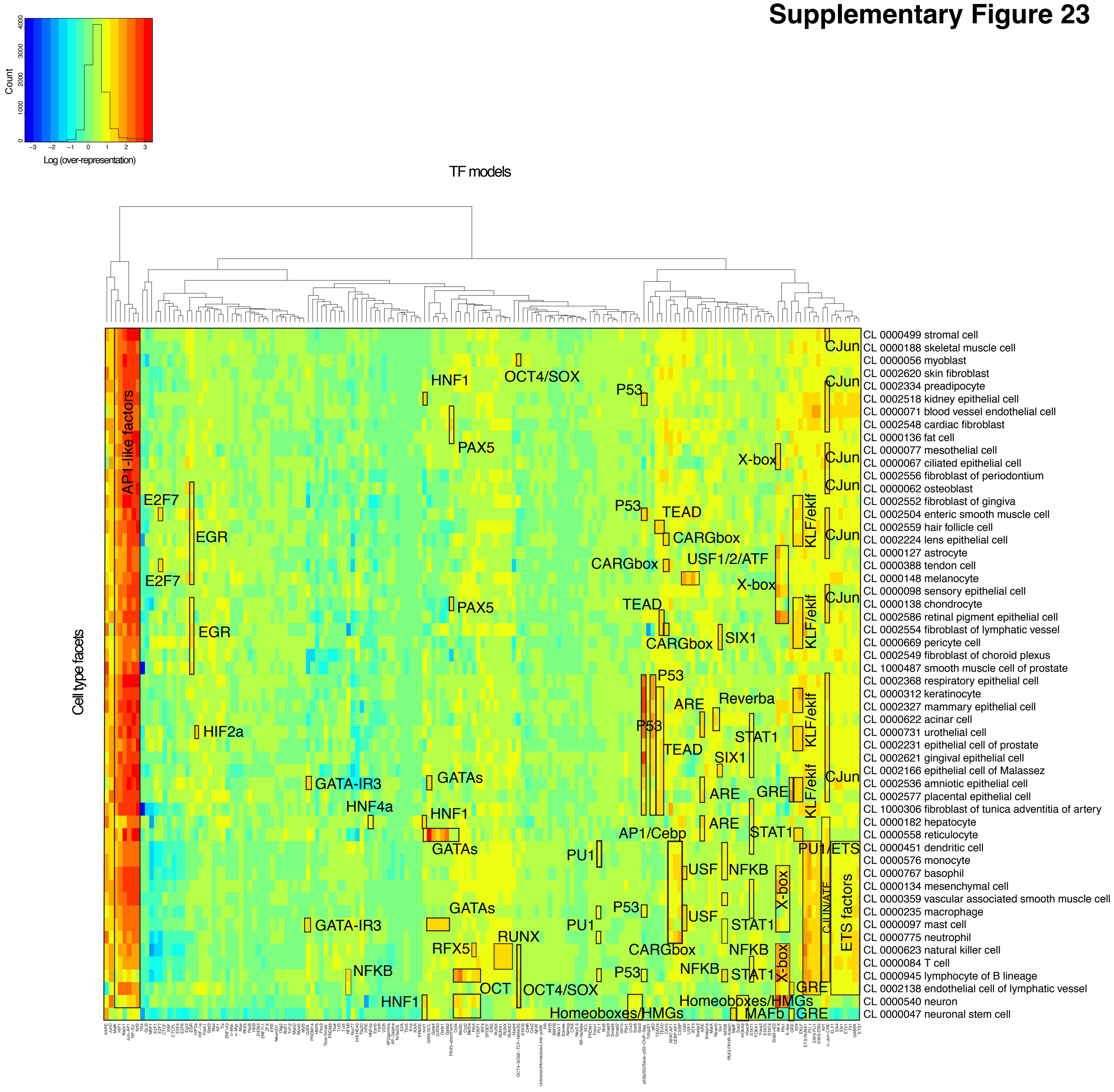
a-b, Distribution of expression specificity for promoters and enhancers. Horizontal axes show the normalized expression specificity, where 0 is ubiquitous and 1 is facet-exclusive expression. Vertical axes show the count of enhancer/promoters. Shown are specificities calculated from cell type facets (a) organ/tissue facets (b). Enhancers are generally more specific than promoters.

c, Relationship between the number of enhancers detected and expression specificity. The horizontal axis divides cell type facets into 5 equally large groups based on the quintiles of the number of detected enhancers per million mapped CAGE tags. The vertical axis gives the expression specificity as in panel a. Each grey dot indicates one enhancer within respective group. Colored lines (violins) show the overall distribution of the specificity scores as densities for each group. The fifth group has significantly higher cell specificity than any other group
($P < 2e-16$, Mann-Whitney U test).

d, Relation between maximal expression and sample specificity
Distribution of maximal expression of enhancer (TPM) over facets (vertical axis), broken up by specificity of enhancer (horizontal axis: 10 equally large groups starting from lowest (left) to most specific (right)) .

e-f, Relationship between MTR4 expression and number of detected enhancers
e, The vertical axis shows the number of detected enhancer per million mapped reads as a function of the MTR4 expression (horisontal axis) in TPM over all primary cell samples.
f, As previous plot, but using MTR4 expression and number of enhancers per million mapped reads over cell type facets instead of individual libraries.
There is no clear relationship between these two variables.

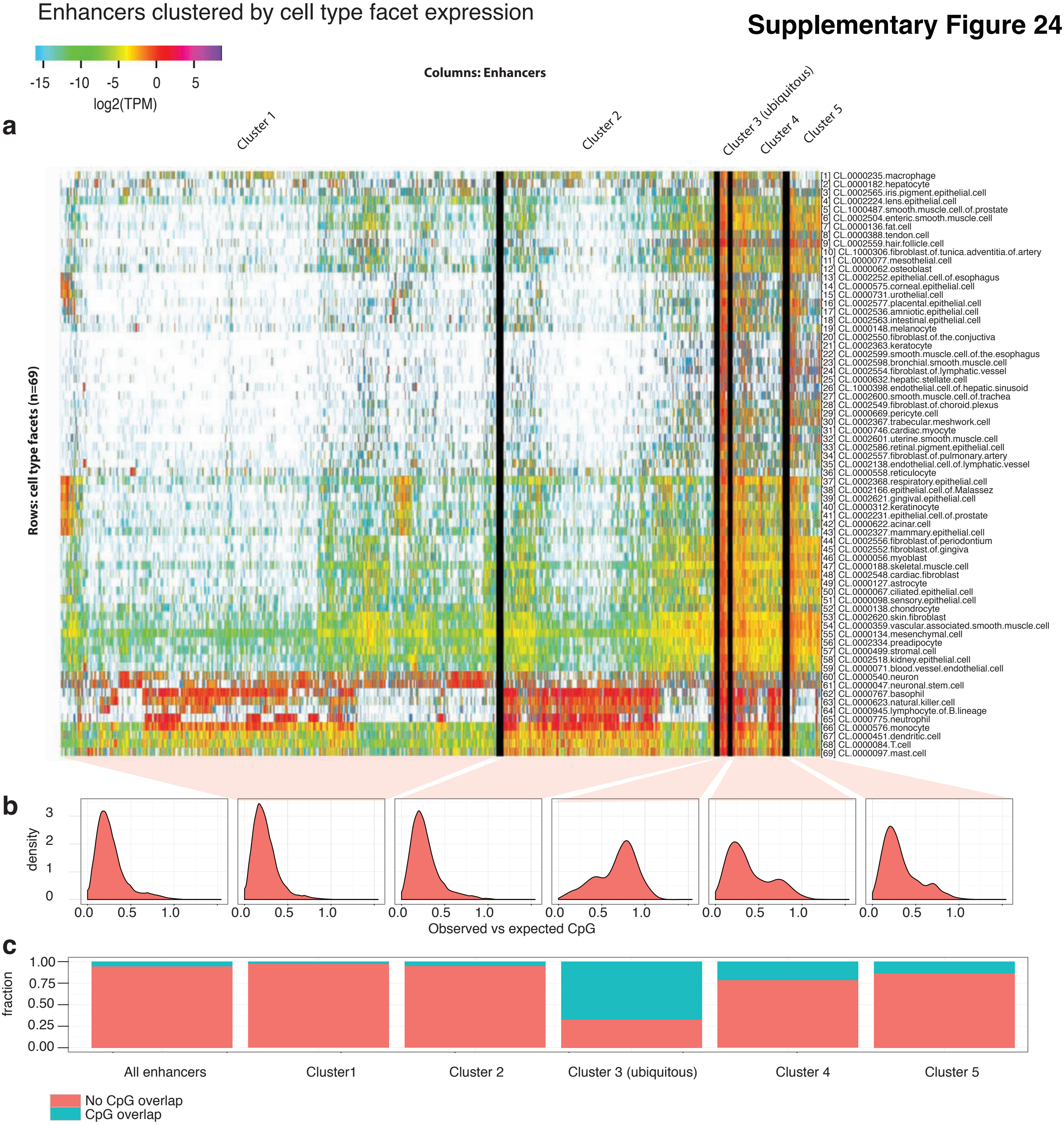
Supplementary Figure 23



Comprehensive agglomerative hierarchical clustering of motif over/under-representation in facet-specific enhancers, using Euclidean distance and complete linkage

We used HOMER on each facet-over-represented enhancer set which had at least 100 enhancers. Columns represent different TF models (motifs) that are scanned over the region. Colors indicate the number of hits vs background, log-scaled. Boxes show one or several TFs being over-represented in one or more facets, labeled by the TF name. Specific blocks of over-represented motifs are indicated with rectangles. Note the general over-representation of AP1-like motifs in almost all facets.

Supplementary Figure 24



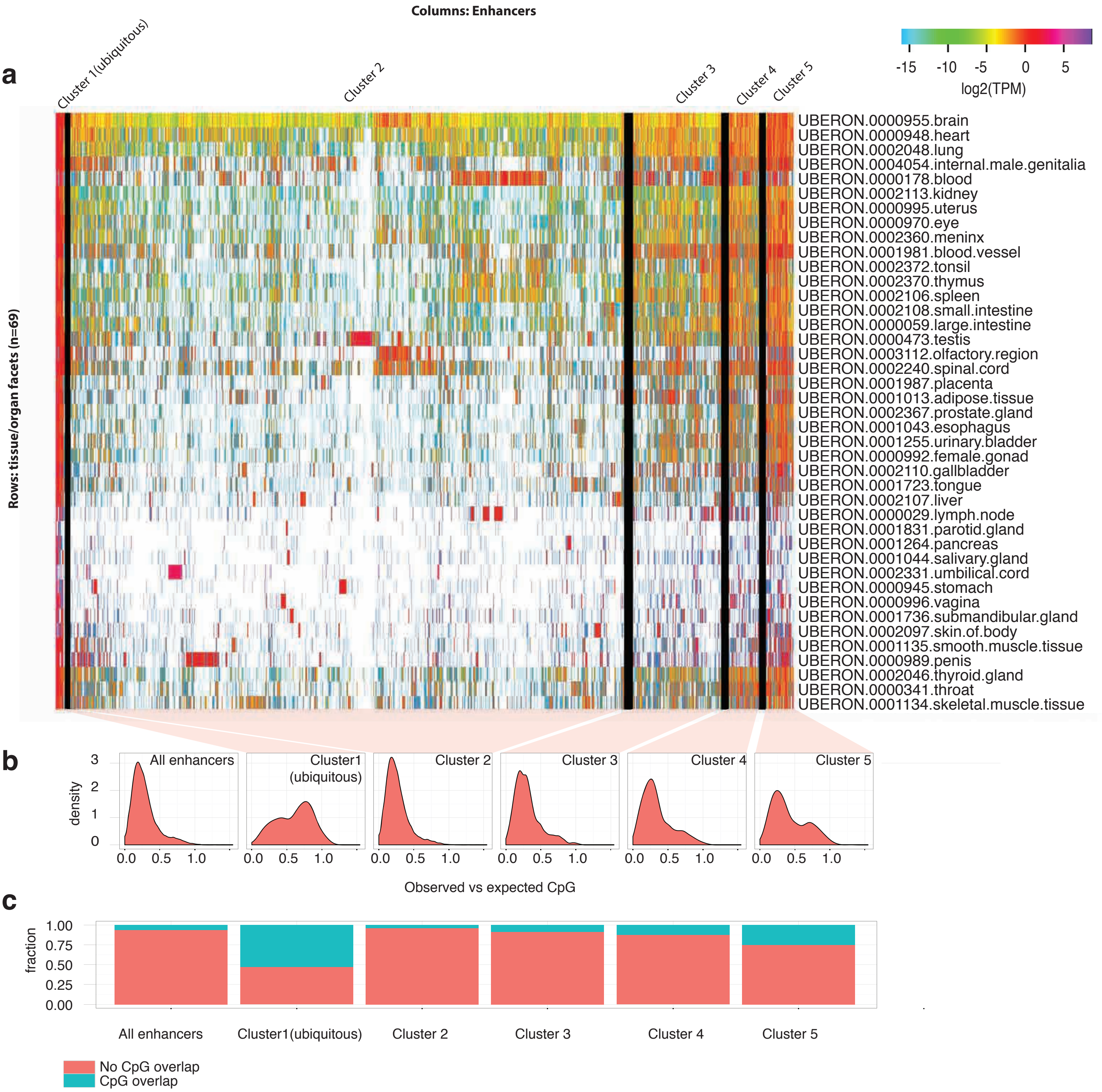
a, Global clustering of individual enhancer expression over all cell type facets.
Heat map ordered by hierarchical clustering of enhancers based on cell type facet data. Columns represent ~22,500 robust enhancers (the whole set of robust enhancers are not used due to computational limitations), rows represent cell type facets. Black columns separate five subclusters (based on the cutree method). The third cluster shows high expression over almost all cell types.

b, CG content of subgroups identified in clustering
Distribution of observed vs expected CG dinucleotide content of the +/- 300 bp region from each enhancer midpoint in each cluster. Note that all enhancer groups are CpG depleted except cluster 3.

c, CpG island overlap of subgroups identified in clustering
As in panel b, but plotting the fraction of overlap with CpG islands.

Supplementary Figure 25

Enhancers clustered by tissue/organ facet expression

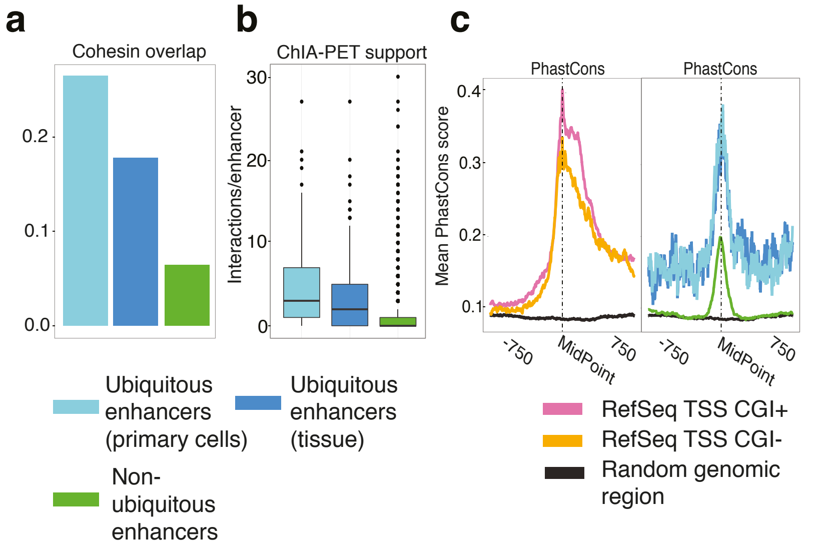


a, Global clustering of individual enhancer expression over all organ/tissue facets.
Heat map showing hierarchical clustering of enhancers based on organ/tissue facet data. Columns represent ~22,500 robust enhancers (the whole set of robust enhancers are not used due to computational limitations), rows represent organ/tissue facets. Black columns separate five subclusters (based on the cutree method). The first cluster shows high expression over almost all tissues.

b, CG content of subgroups identified in clustering
Distribution of observed vs expected CG dinucleotide content of the +/- 300 bp region from each enhancer midpoint in each cluster. Note that all enhancer groups are CpG depleted except cluster 1.

c, CpG island overlap of subgroups identified in clustering
As in panel b, but plotting the fraction of overlap with CpG islands.

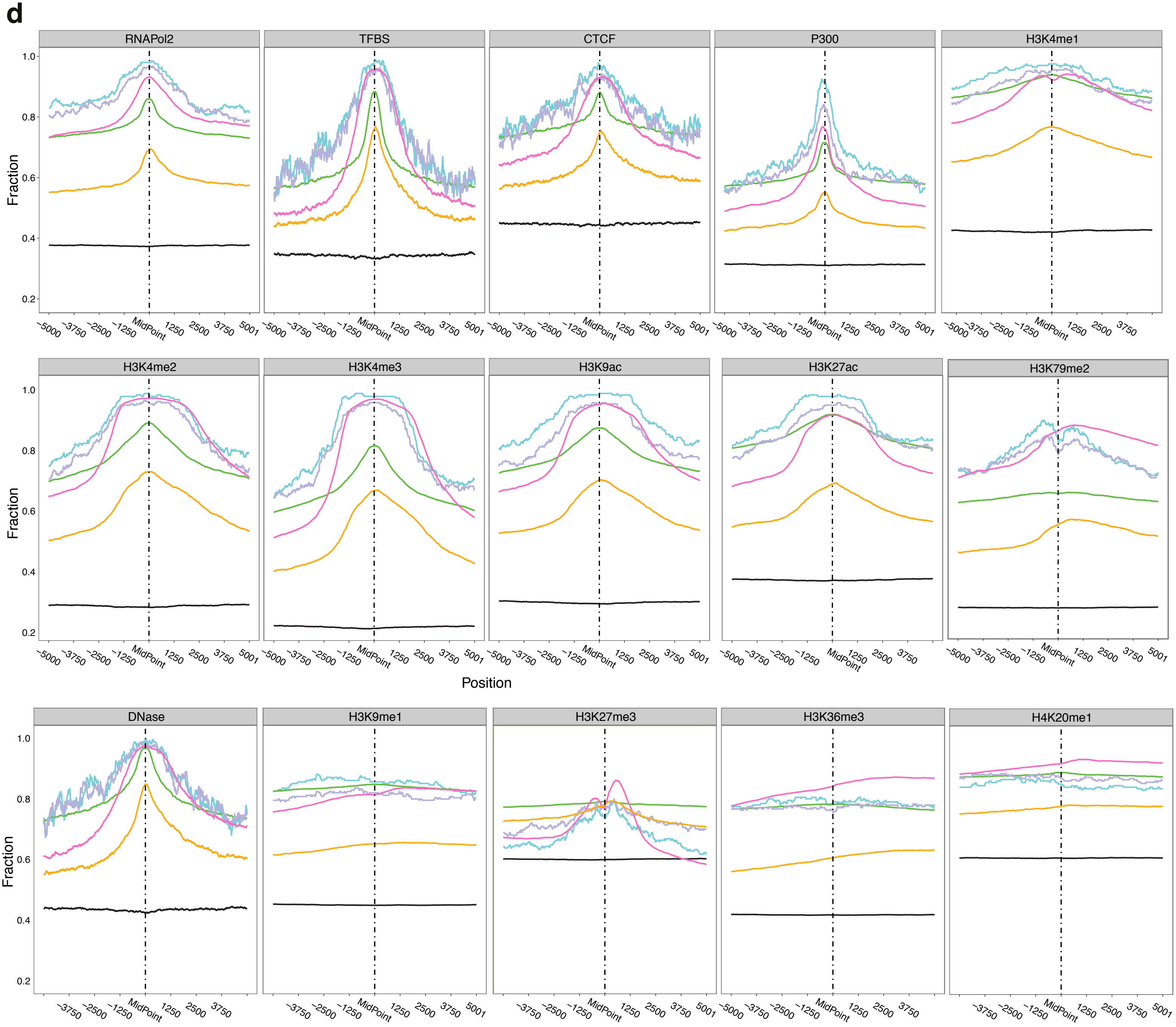
Supplementary Figure 26



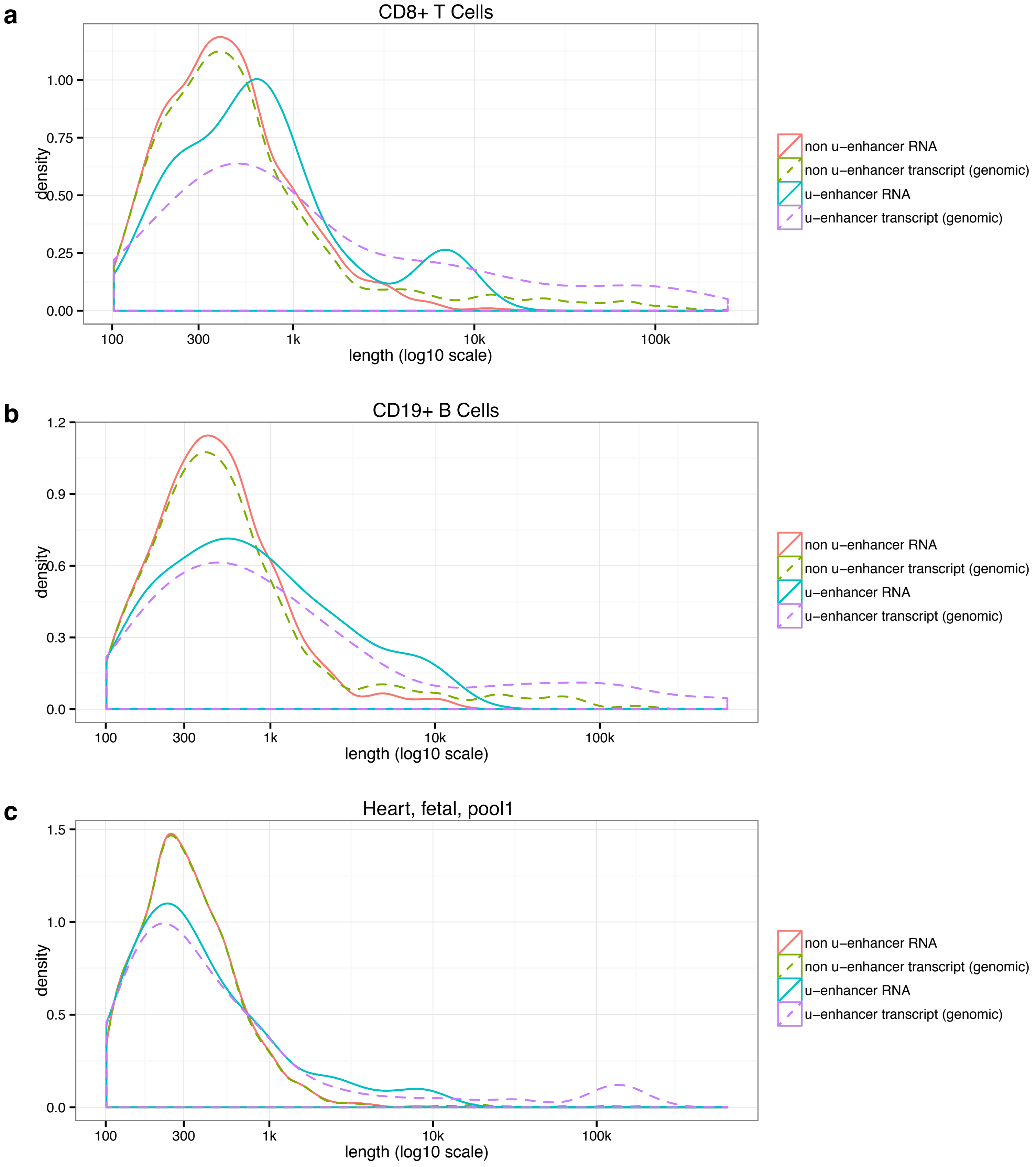
a, Overlap with cohesin ChIP
Overlap between ubiquitous and non-ubiquitous enhancers defined by tissue or cell type facets and cohesin (RAD21 and STAG1) ChIP-seq peaks (+-300 bp).

b, Overlap with ChIA-pet interactions
Number of ChIA-pet interactions to promoters per enhancer. Non-ubiquitous enhancers are shown for reference.

c, Conservation of u-enhancers
Average PhastCons43 conservation/bp of the genomic regions around RefSeq TSSs (left), broken up by CpG island overlap, and enhancers (right) broken up by ubiquitous/non-ubiquitous expression.



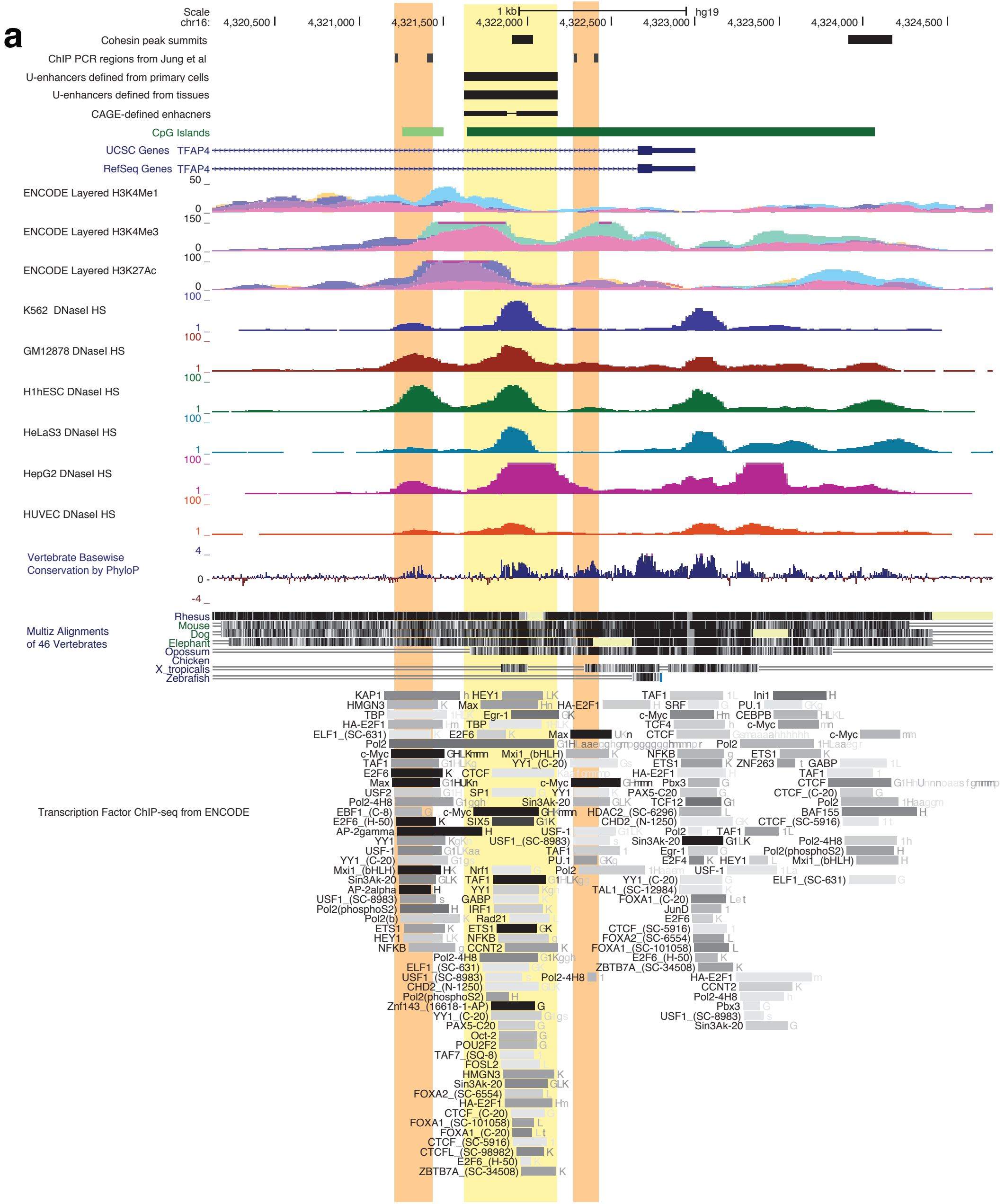
d, Overlap with chromatin features and DNA-binding protein ChIP
Vertical axes show the fraction of enhancers and RefSeq TSS regions (split by ubiquitousness and CpG overlap, respectively) that overlap various ENCODE ChIP-seq peaks (panels). Horizontal axes show the +/- 5000 bp region around the TSS or enhancer center. Ubiquitous enhancers are defined either by tissue of cell type facets (but overlap substantially). Notice the similarity between ubiquitous enhancers and CpG promoters in terms of epigenetic features. The lower intensity of non-CpG promoters may be due to that they are expressed in fewer cells and therefore are not active in the ENCODE cell lines to the same extent.



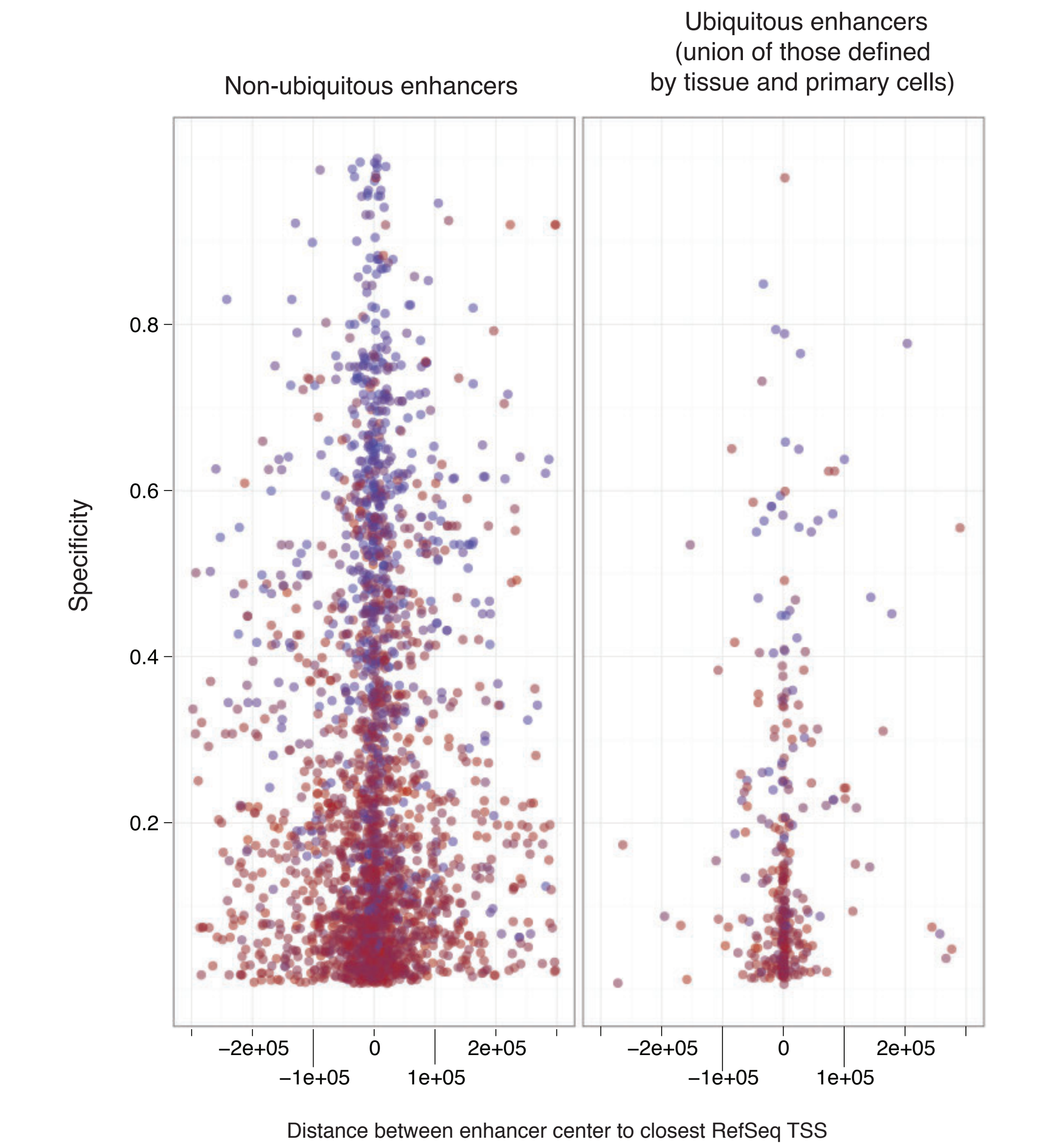
Lengths of transcripts from ubiquitous enhancers

Distributions of genomic lengths of Cufflinks transcripts (dotted lines) and also derived nucleotide lengths of (intron-less) RNAs (solid lines), inferred from RNA-seq (total RNA) in CD8+ T cells (a), CD19+ B cells (b) and fetal heart tissue (c) whose 5' ends originate from CAGE-defined ubiquitous and non-ubiquitous enhancers. RNA-seq was run on the same samples analyzed with CAGE within FANTOM5. While RNAs from ubiquitous enhancers on average are slightly longer than RNA emanating from other enhancers, they are shorter than mRNAs (see main text and Supplementary Figure 11).

Supplementary Figure 28

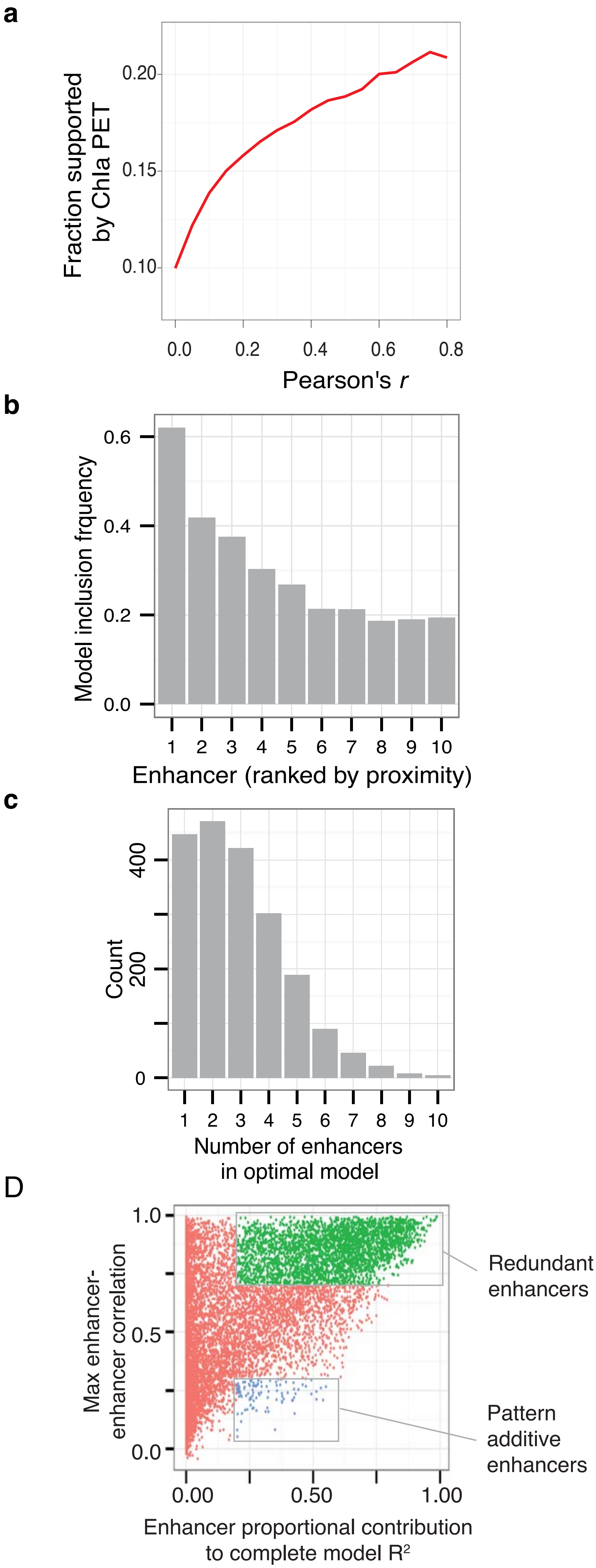


a, Example of a characterized ubiquitous enhancer
Genome-browser image showing a representative ubiquitous enhancer region (light yellow highlight) defined as ubiquitous by tissue and primary cell facets. It overlaps a cohesin ChIP-seq peak, DHSs from multiple cell lines and mutiple transcription factor ChIP-seq peaks. It is located in the first intron of the AP4 (TFAP4) gene, about 1kb from the TSS, inside a wide CpG island. Jung et al (Proc. Natl. Acad. Sci. U.S.A. 105, 15046–15051 (2008)) previously characterized E-box regions that flank this region (~100 bp away; light orange highlights), but judging from the ENCODE ChIP-seq and DHS tracks, most regulatory events are focused at the CAGE-defined enhancer between these.



b, Features of closest TSS to ubiquitous and non-ubiquitous enhancers
Horizontal axes show the distance from enhancer midpoints to the closest RefSeq TSS in any direction. Vertical axes show the expression specificity (as in Figure 5: 1 indicates expression exclusive to one facet) based on cell type facets for those TSSs. The color indicates the CG over-representation in the +/- 300 bp region around these TSSs.

Supplementary Figure 29



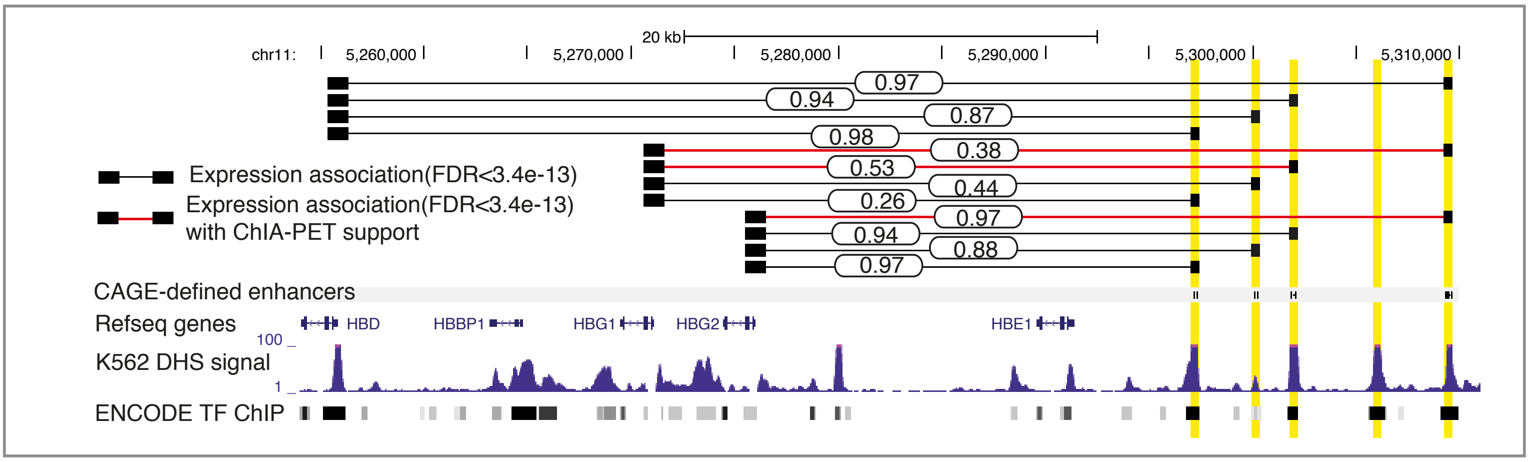
a, Correlation between physical enhancer-TSS interactions and expression association scores
Fraction of predicted enhancer-TSS associations supported by ENCODE ChIA-PET interaction data from four cell lines (vertical axis), as a function of Pearson correlation coefficient cutoff (horizontal axis)

b, Enhancer inclusion in minimal enhancer-TSS models as a function of proximity
Inclusion rate of enhancers in 'optimally' reduced (lasso-based shrinkage) regression models explaining TSS expression with as few enhancers as possible, as a function of enhancer proximity (1 is closest). The first enhancer is included in ~61% of models

c, Number of enhancers included in minimal enhancer-TSS models
Histogram showing how many enhancers that are included in the models in panel b. Most models include 1-3 enhancers out of 10 possible.

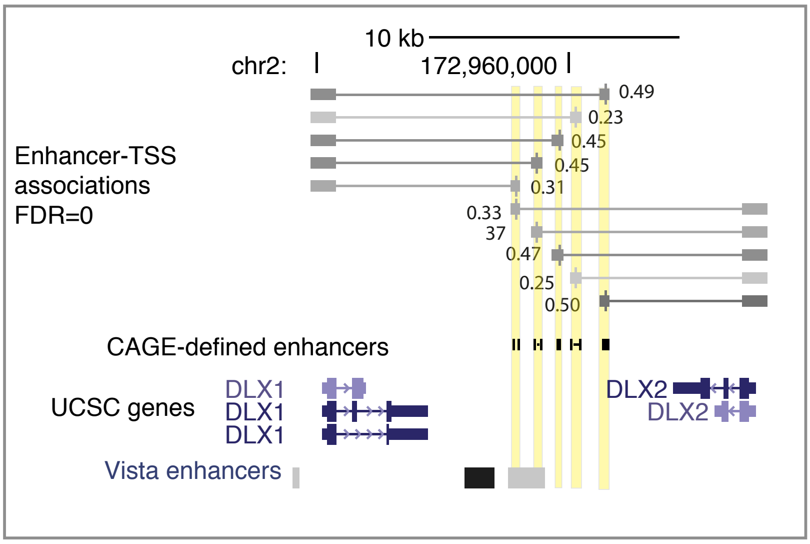
d, Usage of ‘redundant’ enhancers
Proportion of TSS architectures with i) multiple enhancers that have similar expression patterns (redundant enhancers, green) or ii) two or more enhancers that have divergent expression patterns but together explain the TSS expression (blue). Vertical axis shows the highest observed enhancer-enhancer correlation within a model. Horizontal axis shows the proportional contribution of the enhancer to the complete TSS expression model (including all 10 considered enhancers).

a



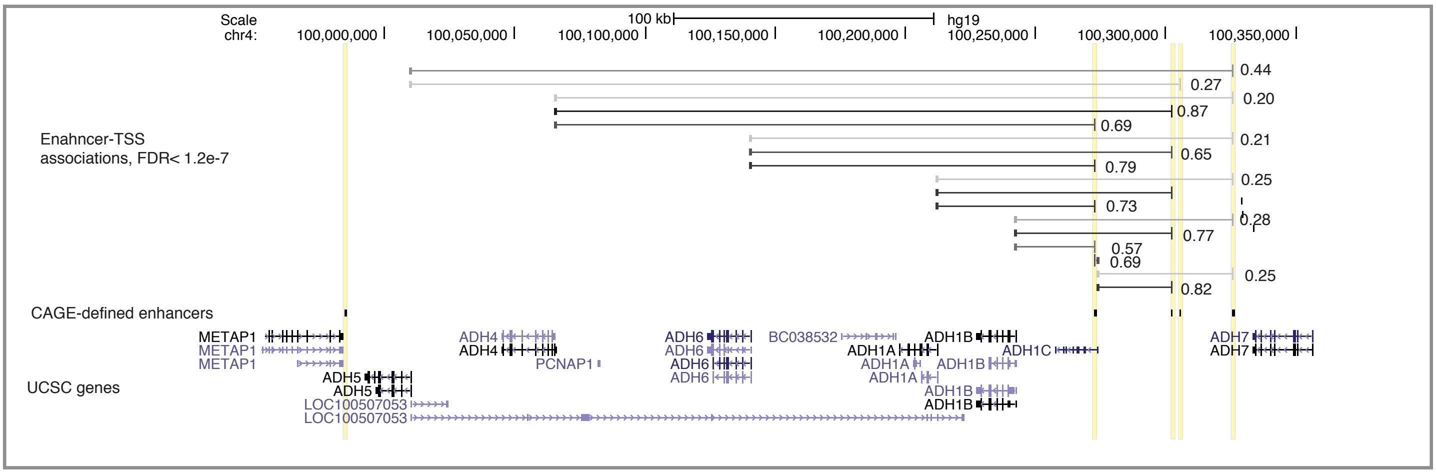
a, Example of recapitulation of many-to-many TSS-enhancer interactions at the beta-globin locus. Enhancers (known LCR HS regions) are indicated in yellow (the fourth from the left overlaps CAGE tags but does not satisfy the balanced bidirectional transcription criterion). Horizontal bars indicate significant correlations between enhancer and TSS CAGE expression; red bars indicate additional ChIA-PET support. Numbers within circles indicate enhancer-TSS correlations (Pearson's r).

b



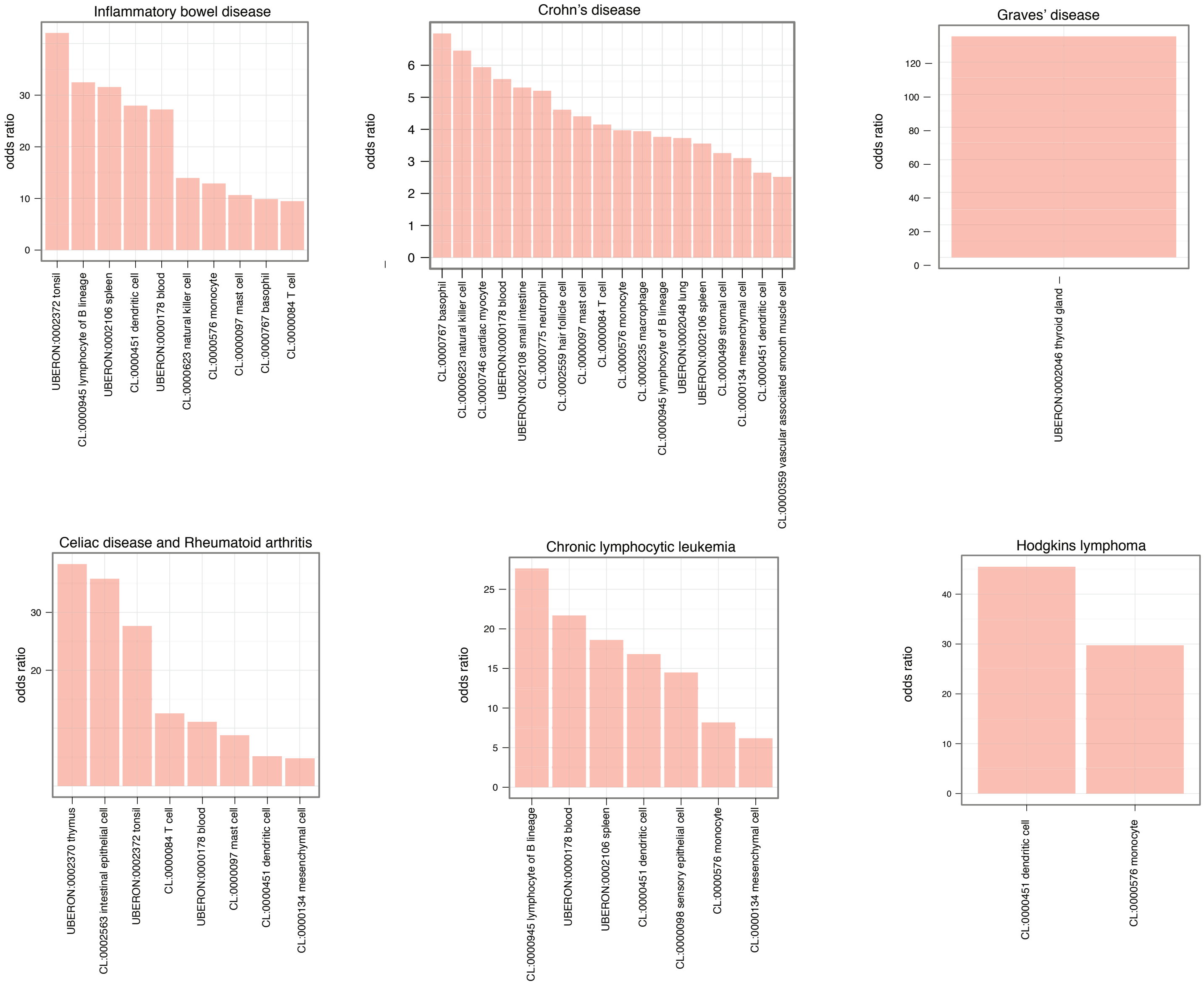
b, Example of enhancer sharing by two related genes at the DLX1-DLX2 locus. The DLX locus exemplifies two related genes that share an array of enhancers in between them.

c



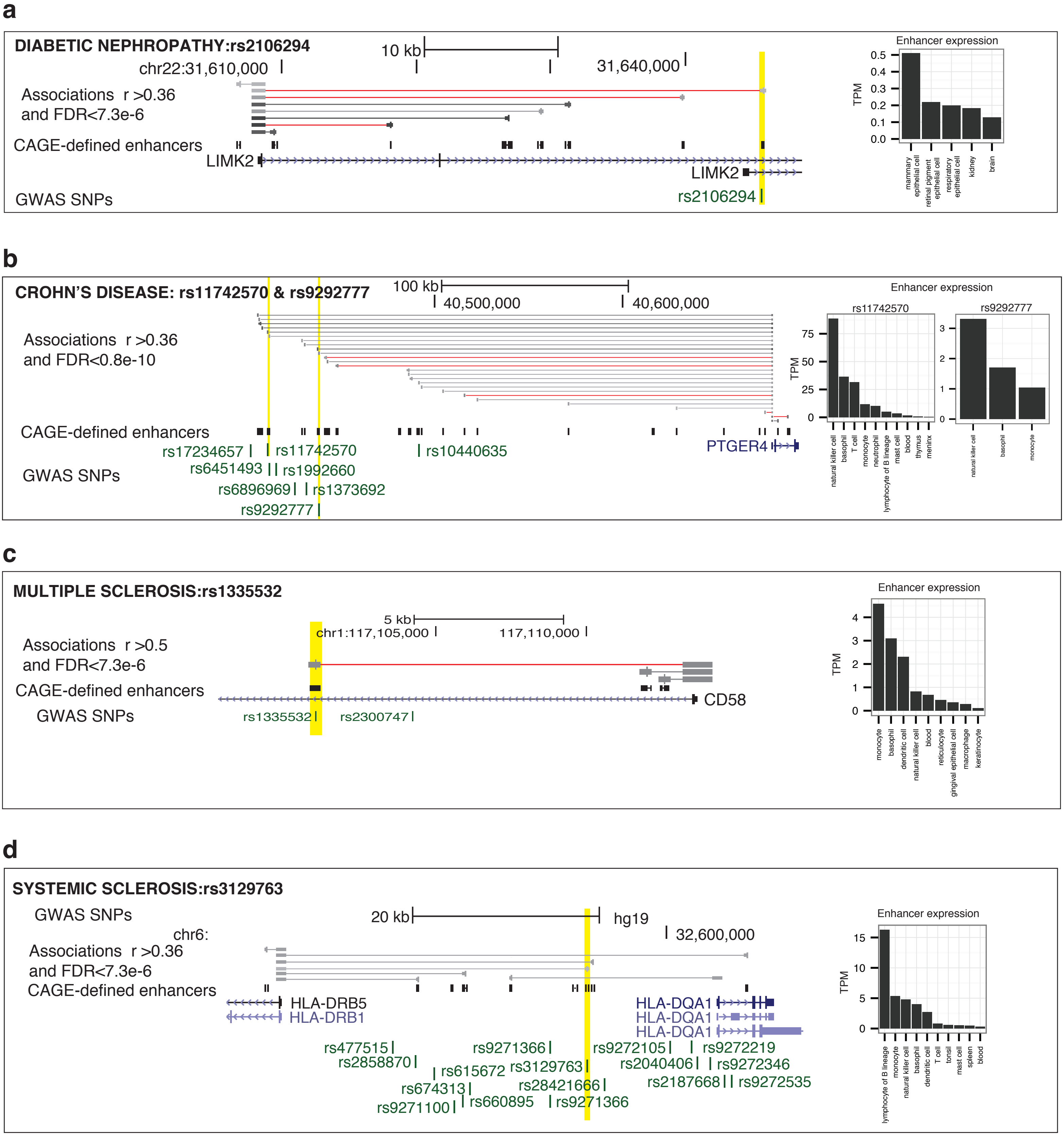
c, Example of multiple genes from a protein complex sharing few enhancers at the ADH gene cluster locus. The alcohol dehydrogenase locus includes multiple genes which work within the alcohol dehydrogenase complex, and are co-expressed. In this locus, the number of genes is greater than the number of enhancers, and the enhancers are linked to many TSSs.

Supplementary Figure 31



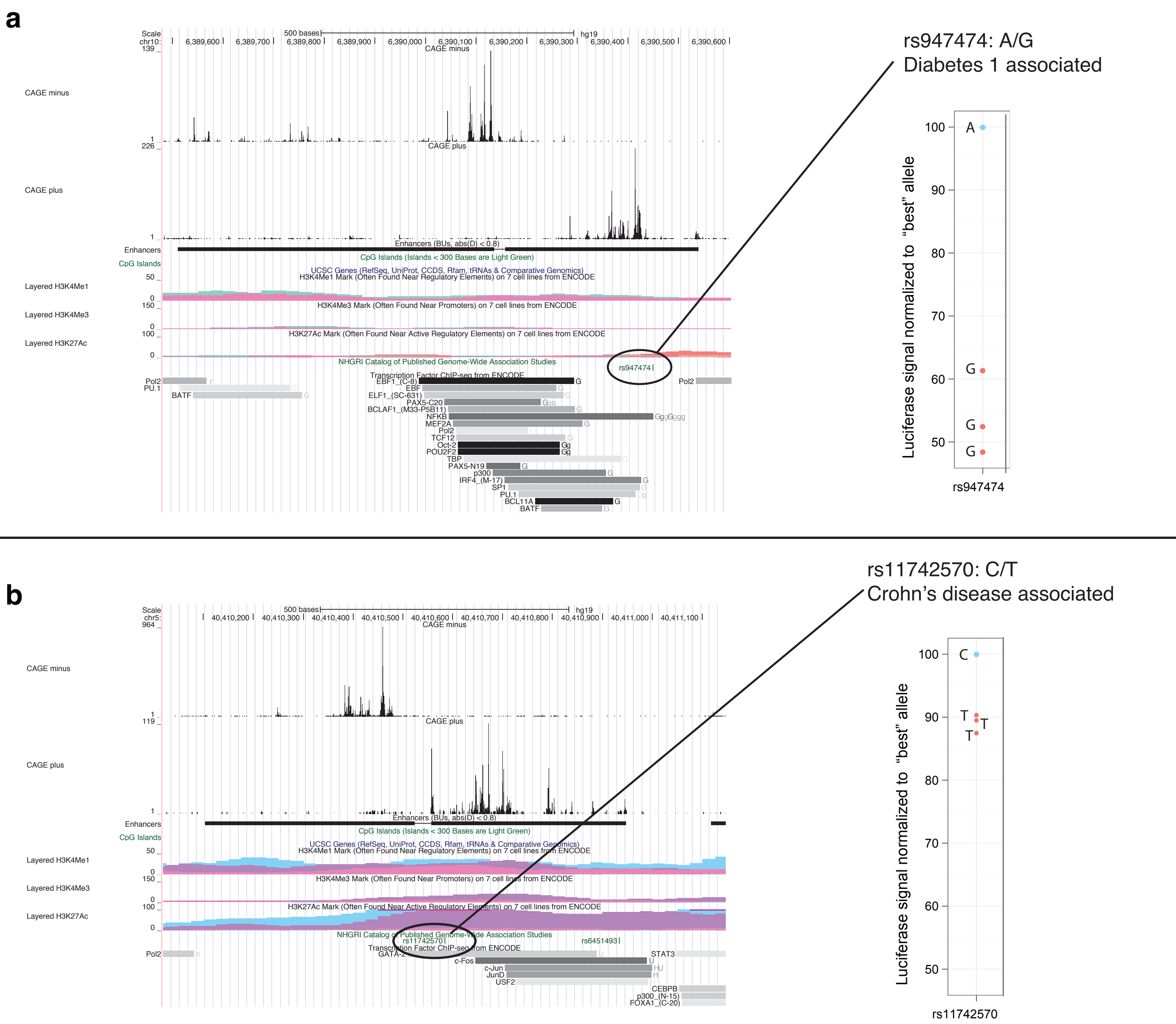
Expression facet over-representation in selected GWAS sets. Odds ratios of observed vs expected overlap of enhancers significantly expressed in facets with GWAS lead and proxy ($r^2 > 0.8$) SNPs associated with six different diseases (panels).

Supplementary Figure 32



Examples of disease-associated SNPs overlapping enhancers. UCSC genome-browser images showing examples of enhancers overlapping disease-associated SNPs (yellow highlights) for diabetic nephropathy (a), Crohn's disease (b), multiple sclerosis (c), and systemic sclerosis (d). In all cases, the enhancer is significantly associated with nearby gene TSSs (grey horizontal bars, red if also supported by ChIA-pet data). These enhancers are expressed in the relevant cell types; bar plots on the right show top facets in which the enhancer is significantly expressed in at least one sample.

Supplementary Figure 33



In vitro validation of regulatory SNPs within enhancers.
Selected enhancers overlapping GWAS SNPs were cloned into plasmids (one plasmid per SNP) and evaluated for *in vitro* enhancer activity, as in Supplementary Figure 17. Each set of panels shows the enhancer as a genome browser picture, and then the relative luciferase signal, normalized so that the allele with the highest activity had signal 100. All of these differences are statistically significant (P<0.05, t-test).

a, The rs947474 SNP is associated with diabetes and the G variant has a ~50% reduction in enhancer activity in THP1 cells .

b, The rs11742570 SNP is associated with Crohn's disease and the T variant has a ~10% reduction in enhancer activity in Daudi cells.