J Gastroenterol (2014) 49:502–510
DOI 10.1007/s00535-013-0791-4

The Japanese Society of Gastroenterology

ORIGINAL ARTICLE—LIVER, PANCREAS, AND BILIARY TRACT

Mitochondrial metabolism in the noncancerous liver determine the occurrence of hepatocellular carcinoma: a prospective study

Atsushi Kudo · Kaoru Mogushi · Tadatoshi Takayama · Satoshi Matsumura · Daisuke Ban · Takumi Irie · Takanori Ochiai · Noriaki Nakamura · Hiroshi Tanaka · Naohiko Anzai · Michiie Sakamoto · Shinji Tanaka · Shigeki Arii

Received: 16 January 2013 / Accepted: 4 March 2013 / Published online: 30 March 2013

Abstract
Background Recurrence determines the postoperative prognosis with hepatocellular carcinoma (HCC). It is unknown how the liver dysfunction involving organic anion transporter failure causes the occurrence of HCCs. This study was designed to elucidate the link between liver dysfunction and multicentric occurrence (MO) after radical hepatectomy.
Methods Forty-nine samples of noncancerous liver tissue from HCC patients within the Milan criteria who were treated at our institution between January 2004 and August 2008 were examined as a training set by using genome-wide gene expression analysis. Using the independent 2-institutional cohort of 134 patients between September 2008 and December 2009, we performed a validation study using tissue microarray analysis. Cox proportional hazard regression analyses for MFS were performed to estimate the risk factors.
Results In the Gene Ontology database (GO:0015711), SLC22A7 expression was the best predictor of MO-free survival [MFS] (Fold, 0.726; $P = 0.001$). High SLC22A7 gene expression prevented the occurrence of HCC after hepatectomy (odds ratio [OR], 0.2; $P = 0.004$). Multivariate analyses identified SLC22A7 expression as an independent risk factor (OR, 0.3; $P = 0.043$). In the validation study, multivariate analyses of MFS identified SLC22A7 expression as an independent risk factor (OR, 0.5; $P = 0.012$). As judged by gene set enrichment analysis, SLC22A7 down regulation was associated with mitochondrion ($P = 0.008$) and oxidoreductase activity ($P = 0.006$). Sirtuin 3 as a regulator of mitochondrial metabolism also determined MFS ($P = 0.018$).
Conclusions The mitochondrial pathways may affect SLC 22A7 function to promote the occurrence of HCC. (Word count: 246).

Keywords Hepatocellular carcinoma · Mitochondria · Sirtuin 3 · SLC22A7 · Organic anion transporter

Abbreviations

CI	Confidence interval
FDR	False discovery rate
GSEA	Gene set enrichment analysis
HCC	Hepatocellular carcinoma
HR	Hazard ratio
MO	Multicentric occurrence
NES	Normalized enrichment score
OR	Odds ratio

Accession number of repository for expression microarray data: GSE40873.

Electronic supplementary material The online version of this article (doi:10.1007/s00535-013-0791-4) contains supplementary material, which is available to authorized users.

A. Kudo (✉) · S. Matsumura · D. Ban · T. Irie · T. Ochiai · N. Nakamura · S. Tanaka · S. Arii
Department of Hepatobiliary-Pancreatic Surgery, Graduate School of Medicine, Tokyo Medical and Dental University, 1-5-45 Yushima, Bunkyo-ku, Tokyo 113-8519, Japan
e-mail: kudomsrg@tmd.ac.jp

K. Mogushi · H. Tanaka
Department of Bioinformatics, Medical Research Institute, Tokyo Medical and Dental University, Tokyo, Japan

T. Takayama
Department of Digestive Surgery, Nihon University School of Medicine, Tokyo, Japan

N. Anzai
Department of Pharmacology and Toxicology, Dokkyo Medical University School of Medicine, Mibu, Tochigi 321-0293, Japan

M. Sakamoto
Department of Pathology, Graduate School of Medicine, Keio University, Tokyo, Japan

Introduction

Hepatocellular carcinoma (HCC) is the third most common cause of cancer-related deaths worldwide because of its high fatality (overall ratio of mortality to incidence of 0.93) [1]. In 2008, an estimated 748,000 new cases of HCC and 696,000 deaths associated with HCC occurred [1]. Underlying liver diseases increase the risk of HCC occurrence, although the mechanism by which this occurs is unclear [2]. Multicentric occurrence (MO) is a crucial problem irrespective of anatomic resection to prevent local recurrences. The clinical courses and biological features of MO definitely differ from those of local recurrence and intrahepatic metastases [3, 4].

Various treatments are selected for MO of HCC [5]. Anatomic resection proposed by Makuuchi et al., was implemented to overcome the local recurrence involving micro-dissemination into the portal vein and intrahepatic metastasis [6]. Resection is regarded as a first-line therapy when HCC occurrence is within the Milan criteria. However, even after curative treatments involving anatomic resection, considerable risk of MO has been reported [7]. There is no benefit of anatomic resection in HCC with MO [8]. Noncancerous liver tissue with oncogenic potential may explain the risk of MO after hepatectomy [3]. The criteria for MO are defined in the classification of the Liver Cancer Study Group of Japan [4].

This study was designed to elucidate whether noncancerous liver function involving transporter activity influences the MO of early-stage HCC. This study excluded patients beyond the Milan criteria to reduce malignant factors of the primary tumor. Genome-wide gene expression analysis was used to elucidate the link between this hepatocellular function and MO of HCC. Hepatocellular organic anion transporters exchange materials that are indispensable for mitochondrial metabolism, and they detoxify the sinusoidal microcirculation. Xenobiotics transported through organic anion transporters, are detoxified in hepatocytes and excreted into bile [9]. In the organic anion transporter genes according to the Gene Ontology database (GO:0015711) as a hepatocellular function, the best predictor for MO of HCC was SLC22A7. According to recent reports, SLC22A7 expressed on the hepatocellular sinusoidal membrane takes up orotic acid [10, 11]. An experimental study reported that exposure to dietary orotic acid with hepatectomy promotes liver carcinogenesis [12]. In this study, we present evidence indicating that decreased SLC22A7 expression associated with mitochondrial disability might play a causative role in liver carcinogenesis and that it will be a biomarker for predicting MO even after curative hepatic resection.

Methods

Training set

Between January 2004 and August 2008, 231 curative hepatectomies for HCC were performed at Tokyo Medical and Dental University Hospital (Tokyo, Japan). In total, 69 of 115 patients within the Milan criteria were ethically informed according to the guidelines of our institutional review board. This study excluded "beyond Milan", a contraindication for anatomic resection and liver transplantation. Trans-arterial embolization, radiofrequency ablation, and systemic chemotherapy are available options when tumor recurrence is evident after the first hepatectomy. Serum alpha-fetoprotein (AFP) and des-gamma-carboxy prothrombin (DCP) levels were measured monthly, and ultrasonography, computed tomography, and magnetic resonance imaging were performed every 3 months. The criteria for MO of HCC were defined according to the classification of the Liver Cancer Study Group of Japan (the recurrent tumor consists of early HCC occurring in a different hepatic segment with or without dysplastic nodules in peripheral areas, or well differentiated HCCs with peripheral moderately or poorly differentiated HCC) [4]. Any tumor, regardless of the time to recurrence, arising in the same segment as the initial tumor (or within 2 cm from the surgical stump when performing segmentectomy) was considered a "local" recurrence [13].

Genome-wide gene expression analysis

All samples of noncancerous liver tissue obtained from the resected specimens were separately frozen immediately and stored at −80 °C. Total RNA was extracted using an RNeasy kit (Qiagen, Hilden, Germany). Contaminating DNA was removed by digestion with RNase-free DNase (Qiagen). Upon checking the RNA integrity of the samples using the Agilent 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA, USA), 49 samples were given an RNA integrity number exceeding 5.0. After preparing complementary RNA by 1-cycle target labeling and with a control reagents kit (Affymetrix, Santa Clara, CA, USA), hybridization and signal detection of HG-U133 Plus 2.0 arrays (Affymetrix) were performed according to the manufacturer's instructions. The 49 microarray datasets were normalized using the robust multiarray average method of R

statistical software (version 2.12.1) together with the Bio-Conductor package. Estimated gene expression levels were obtained as log2-transformed values, and 62 control probe sets were selected for further analysis.

Selection of organic anion transporter genes for HCC occurrence

First, probe sets corresponding to known genes were selected on the basis of the NetAffx annotation file, version 32 (available at: http://www.affymetrix.com/analysis/index.affx). Next, probe sets of organic anion transporter genes were selected according to GO:0015711 (52 probes). The 35 probes were matched to the criteria. The univariate Cox proportional hazards regression model was used to estimate the relationship between the gene expression pattern and MO for each probe set. Probe sets that had a $P < 0.005$ by the likelihood ratio test were selected.

Validation study on immunohistochemical analysis using tissue microarrays

To validate the clinical significance of SLC22A7 expression, 134 patients who visited Tokyo Medical and Dental University Hospital and Nihon University Hospital between September 2008 and December 2009 were enrolled in the prospective multicenter cohort. The candidate gene was assessed by immunohistochemical staining on tissue microarrays using resected liver samples from patients with HCC within the Milan criteria with an anti-SLC22A7 antibody (provided by Dr. Anzai) at a 1:20 dilution [14] by the use of an automated immunostainer (Ventana XT System; Ventana Medical Systems, Inc., Tucson, AZ, USA) as described previously. The SLC22A7 staining was judged by 2 investigations, and staining of less than 25 % of cells was judged as negative (Fig. 1c, d).

Gene set enrichment analysis (GSEA)

To investigate the biological backgrounds correlated with a particular gene expression pattern, we used GSEA version 2.0.7 with MSigDB gene sets version 3.0. Gene set category C5, which is based on the GO database, was used. Gene sets satisfying both $P < 0.05$ and a false discovery rate (FDR) <0.25 were considered significant. The customized sirtuin 3-related gene set involving 147 genes was constructed to examine the relationship with SLC22A7 according to the supplemental Fig. 1.

Statistical analysis

Univariate and multivariate Cox regression analyses were performed using SPSS 20.0 (IBM, Armonk, NY, USA). The median value was selected for the cut-offs of clinical variables in the training set and the validation set. The MO-free survival (MFS) was evaluated by the Kaplan–Meier method and the log-rank test. Two-sided $P < 0.05$ were considered significant. Values are given as the mean ± SD unless otherwise stated.

Results

Baseline characteristics

All possible curative resections within the Milan criteria (R0) were attempted for the 49 patients in the training set and the 134 patients in a multicenter validation study (Table 1). The mean observation time in the training set and the validation set were 21.1 and 16.4 months, respectively. The mean MFS in the training set and the validation set were 12.3 and 10.6 months. There was no difference in the mean MFS between the training and validation sets ($P = 0.602$), though the mean observation time was longer in the training set ($P = 0.010$). There was no difference between the two studies in age, gender, viral infection (HBV and HCV), serum albumin, total bilirubin, AFP, DCP, platelet count, tumor number, pathological invasion into the portal vein, liver cirrhosis, and MO occurrence rate. There were differences in Child class B ($P = 0.03$), ICG-R15 value ($P = 0.02$), tumor size ($P < 0.0001$), and anatomic resection ($P < 0.0001$), respectively.

The predictive factors for MFS in the training set

As shown in Table 2, low SLC22A7 expression was the best predictor of MFS ($P = 0.001$; fold difference between the mean expression levels of patients with and without MO = 0.726). Table 3 presents the link between MFS and SLC22A7 gene expression in the 49 HCC patients within the Milan criteria. The MO was observed in seventeen patients (35 %). Univariate analyses identified HCV infection, serum albumin levels, serum platelet counts, and SLC22A7 gene expression as risk factors for HCC occurrence. These results led us to determine which risk factors were independently predictive of the prognosis of HCC patients. According to multivariate analysis, SLC22A7 gene expression determined prognosis independently. Figure 1a indicates that the cumulative recurrence-free survivals were significantly associated with SLC22A7 expression. (Log-rank test, $P = 0.001$).

Validation study for SLC22A7 expression among patients within the Milan criteria

Table 4 illustrates the link between MFS and SLC22A7 protein expression as judged by a tissue microarray in 134

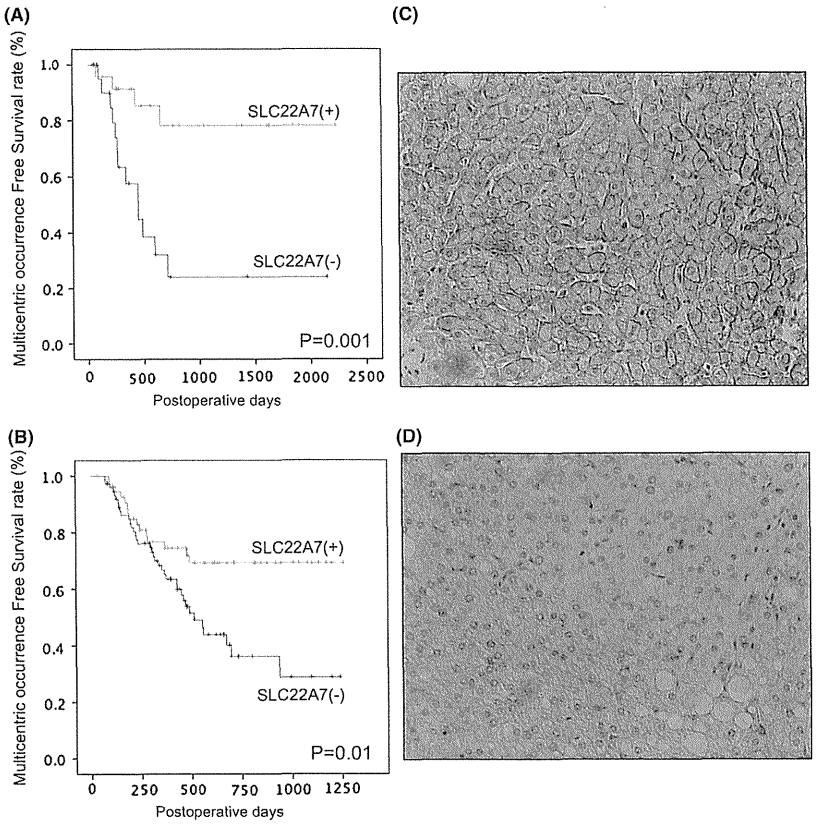


Fig. 1 **a** Training study. MFS of postoperative HCC patients within the Milan criteria with high (SLC22A7 (+) group) and low (SLC22A7 (−) group) SLC22A7 gene expression in noncancerous liver tissue. The median expression level for each gene was used as a cutoff value. The green lines denote the Kaplan–Meier curves for the SLC22A7 (+) group. *Blue line* denotes low SLC22A7 expression in the SLC22A7 (−) group. The prognosis of the SLC22A7 (−) group was significantly worse ($P = 0.001$). **b** Validation study using a prospective multicenter cohort. The *green lines* denote the Kaplan–Meier curves for SLC22A7 (+) group. The *blue line* denotes the MFS in the SLC22A7 (−) group. Note the poor survival of the SLC22A7 (−) group ($P = 0.01$). **c** Immunohistochemical analysis (magnification, ×20) with high expression of SLC22A7. **d** Immunohistochemical analysis (magnification, ×20) with low expression of SLC22A7 protein

HCC patients within the Milan criteria. The SLC22A7 protein was expressed at the hepatocellular sinusoidal membrane in noncancerous tissues (Fig. 1c). The MO was observed in 52 patients (38 %). Low SLC22A7 expression was confirmed when the immunohistochemical staining of the tissue microarray was less than 25 % (Fig. 1d). Univariate analyses identified anatomic resection and low SLC22A7 expression as risk factors for HCC occurrence (Table 4). Other clinicopathological factors did not determine MO. These results led us to determine which factors were independently predictive of prognosis. As shown in Table 4, SLC22A7 expression determined prognosis (OR, 0.5; 95 % CI, 0.3–0.8; $P = 0.012$). In Fig. 1b, the cumulative MFS was significantly associated with SLC22A7 expression ($P = 0.010$). The 1-year MFS in patients with low SLC22A7 expression was 65.2 %, compared with 76.7 % in those with high SLC22A7 expression.

GSEA evaluation of SLC22A7 expression in HCC

The dataset had 54,675 native features. After collapsing the features into gene symbols, 20,606 genes were identified.

Gene set size filters (min = 15, max = 500) resulted in the filtering out of 446/1454 gene sets. The remaining 998 gene sets containing 7,605 genes were used in the analysis. The p value of SLC22A7 was ranked at 261st out of 21,050 genes included in the HG-U133 Plus 2.0 array. In total, 552 of 998 gene sets were positively correlated with SLC22A7 expression. Seventy-seven gene sets were significant at FDR of <25 %. Thirty-six gene sets were significantly enriched at a nominal $P < 1$ %. Conversely, 446 of 998 gene sets were negatively correlated with SLC22A7 expression. No gene sets were significantly enriched at FDR < 25 %. Two gene sets were significantly enriched at a nominal $P < 1$ %. As shown in Fig. 2, mitochondrion ($P = 0.008$; FDR = 0.199; NES = 1.804), oxidoreductase activity ($P = 0.006$; FDR = 0.157; NES = 1.854), and fatty acid metabolic process ($P = 0.021$; FDR = 0.177; NES = 1.723) were significantly correlated with SLC22A7 expression. By analyzing the gene expression profiles of 49 samples of noncancerous tissue, GSEA showed that the 27 of the 62 gene sets (44 %) were closely related with mitochondrial genes involving oxidoreductase activity and fatty acid metabolic process at FDR of 20 % with a nominal $P < 0.05$ (Supplementary Table 1). Mitochondrial sirtuin 3, reported as the regulator of fatty acid oxidation, oxidative damage and orotic acid concentrations, prevents deacetylates and stimulates ornithine transcarbamylase (OTC) and modulates amino acid catabolism and β-oxidation [15, 16]. The correlation between SLC22A7 and sirtuin3 expression levels was 0.300 ($P = 0.034$). As shown in Fig. 2d, decreased sirtuin 3 gene expression also determined patient MFS ($P = 0.018$). These results led us to examine whether the customized sirtuin 3-related gene set correlates with SLC22A7. The GSEA revealed a remarkable correlation with SLC22A7 ($P = 0.008$; FDR = .008; NES = 1.786), as shown in Supplemental Figs. 2 and 3. Mitochondrial factor may be confounding factor of the two factors, though the detailed mechanism remains unknown.

Table 1 Baseline characteristics

Variables	Training set Mean (SD)	Validation set Mean (SD)	P
Age	66.8 ± 10.3	67.3 ± 9	0.941
Male/female	34/15	97/37	0.411
Viral infection			
HBV	11	17	0.136
HCV	31	85	0.363
Laboratory data			
Prothrombin time (%)	83.7 ± 17.9	92.3 ± 12.8	<0.0001*
Albumin (g/dL)	3.9 ± 0.5	4 ± 0.5	0.067
Total bilirubin (mg/dL)	0.8 ± 0.4	0.8 ± 0.4	0.401
Platelet (×10^4/mL)	14.1 ± 6.8	14.4 ± 5.6	0.509
Child-Pugh A/B	42/7	128/6	0.03*
ICG-R15 (%)	20.2 ± 11.9	16.3 ± 10.2	0.015*
Tumor factor			
Diameter (cm)	3.2 ± 1	2.5 ± 1	<0.0001*
Number	1.4 ± 0.9	1.2 ± 0.5	0.159
AFP (ng/mL)	350 ± 1305	175 ± 750	0.272
DCP (mAU/mL)	1053 ± 3389	1345 ± 7086	0.958
Pathological vp (+)	9	22	0.790
Anatomic resection			
Yes	24	28	<0.0001*
Liver background			
NL/CH/LC	3/17/29	6/71/57	0.081
Multicentric occurrence			
+	17	52	0.371

DCP des-gamma-carboxy prothrombin

* $P < 0.05$ was considered significant

Discussion

The retrospective training study and validated prospective multicenter study provided evidence that low SLC22A7 expression promoted the occurrence of HCC after hepatectomy in patients within the Milan criteria. This is the first study to elucidate the link between the occurrence of HCC and SLC22A7 expression in noncancerous liver tissue. Moreover, this gene expression is closely related with the gene sets of mitochondrion in noncancerous liver, as judged by GSEA evaluation in the training set (Fig. 2a). The GSEA showed that the 44 % of SLC22A7-related gene sets were closely related with mitochondrial genes involving oxidoreductase activity and fatty acid metabolic process. Mitochondrial metabolism may also involved fatty acid synthase and oxidoreductase activity.

A previous study demonstrated that the gene expression profiles of the surrounding nontumoral liver tissue, but not the tumor tissues, were highly correlated with survival in the training set of Japanese patients and in validation sets in the United States and Europe ($P = 0.04$) [17]. Anatomic resection was not identified as a prognostic factor in the two studies. The HCV infection, platelet counts, and serum

Table 2 The univariate analyses to estimate the relationship between a gene expression pattern and MO of HCC for each probe set of organic anion transporter genes according to the Gene Ontology database (GO:0015711)

Probe set	Symbol	Title	Fold	P
221661_at	SLC22A7	Solute carrier family 22 (organic anion transporter), member 7	0.726	0.001*
1557918_s_at	SLC16A1	Solute carrier family 16, member 1 (monocarboxylic acid transporter 1)	0.831	0.005
210366_at	SLC10A1	Solute carrier family 10 (sodium/bile acid cotransporter family), member 1	0.936	0.017
207185_at	SLC16A1	Solute carrier family 16, member 1 (monocarboxylic acid transporter 1)	0.907	0.071
202236_s_at	SLCO1A2	Solute carrier organic anion transporter family, member 1A2	0.912	0.074

The best organic anion transporter genes in GO-0015711. † Fold values were calculated by the ratio of mean expression levels in the patients with MO to that in patients without MO

* $P < .005$ was considered significant (Cox's proportional hazard ratio)

Table 3 The risk factors determining MO in 49 HCC patients within the Milan criteria (training set)

Variables	Univariate		P	Multivariate		P
	OR	95 %CI		OR	95 %CI	
Age (years)						
>68	1.7	(0.6–4.4)	0.288			
Gender						
Female	1.2	(0.5–3.3)	0.694			
HCV						
(+)	0.2	(0.1–0.9)	0.027*	0.4	(0.1–1.7)	0.206
HBV						
(+)	2.7	(0.6–11.9)	0.190			
Total bilirubin (×mg/dL)						
≧0.8	1.9	(0.7–5.5)	0.241			
Albumin (g/dL)						
≧4.0	0.2	(0.1–0.6)	0.002*	0.3	(0.1–1.0)	0.056
Prothrombin time (%)						
≧84.4	0.7	(0.3–1.8)	0.426			
Platelet (×10^4/μL)						
≧11.9	0.3	(0.1–0.8)	0.017*	0.8	(0.2–2.8)	0.700
Child-Pugh A vs. B	0.4	(0.1–1.4)	0.140			
ICG-R15 (%)						
≧20	2.3	(0.9–6.1)	0.101			
Tumor diameter (cm)						
>3	0.9	(0.4–2.4)	0.871			
Multiple	1.6	(0.5–5.7)	0.444			
AFP (ng/mL)						
≧12	1.1	(0.9–1.3)	0.337			
DCP (mAU/mL)						
≧38	0.6	(0.2–1.6)	0.284			
Capsule						
(+)	0.4	(0.0–20)	0.309			
Capsule invasion						
(+)	0.5	(0.2–1.3)	0.151			
Pathological vp						
(+)	0.9	(0.2–4.0)	0.909			
CM type vs. SNEG	1.0	(0.3–3.5)	0.972			
SN type vs. SNEG	0.3	(0.1–1.1)	0.061			
Moderately differentiated						
Vs. well	0.8	(0.2–3.5)	0.775			
Poorly differentiated						
Vs. well	0.2	(0.1–2.0)	0.233			
Liver cirrhosis						
(+)	1.7	(0.6–4.6)	0.298			
SLC22A7 expression						
High	0.2	(0.1–0.6)	0.004*	0.3	(0.1–1.0)	0.043*
Anatomic resection						
(+)	0.7	(0.3–1.8)	0.426			

DCP des-gamma-carboxy prothrombin *SN* simple nodular type, *SNEG* simple nodular type with extranodular growth, *MC type* confluent multinodular type

* $P < .05$ was considered significant

Table 4 The risk factors that determine MO in 134 HCC patients within the Milan criteria (validation set)

Variables	Univariate		P	Multivariate		P
	OR	95 %CI		OR	95 %CI	
Age						
≧67	1.08	(0.62–1.89)	0.792			
Gender						
Female	0.88	(0.46–1.68)	0.701			
HCV						
(+)	1.13	(0.64–1.98)	0.681			
Body weight						
≧58.3	1.00	(0.58–1.74)	0.989			
Total bilirubin (mg/dL)						
≧0.7	0.79	(0.45–1.39)	0.418			
Albumin (g/dL)						
≧4.1	0.70	(0.40–1.21)	0.200			
Prothrombin time (%)						
≧97 %	0.79	(0.46–1.37)	0.406			
Platelet (×10^4/μL)						
≧13.8	0.80	(0.46–1.37)	0.413			
Child-Pugh score						
≧6	0.81	(0.41–1.62)	0.550			
ICG-R15 (%)						
≧14.3	1.17	(0.68–2.03)	0.568			
Tumor diameter (cm)						
≧2.4	0.88	(0.51–1.52)	0.644			
Multiple tumor						
(+)	1.28	(0.57–2.84)	0.550			
Pathological vp						
(+)	0.99	(0.48–2.04)	0.985			
AFP (ng/mL)						
≧8.7	1.41	(0.82–2.50)	0.216			
DCP (mAU/mL)						
>36	0.75	(0.43–1.30)	0.302			
Liver cirrhosis						
(+)	0.70	(0.21–2.33)	0.558			
Anatomic resection						
(+)	0.52	(0.24–1.16)	0.110			
SLC22A7 expression						
High	0.46	(0.25–0.84)	0.012*	0.46	(0.25–0.84)	0.012*

DCP des-gamma-carboxy prothrombin

* $P < 0.05$ was considered significant

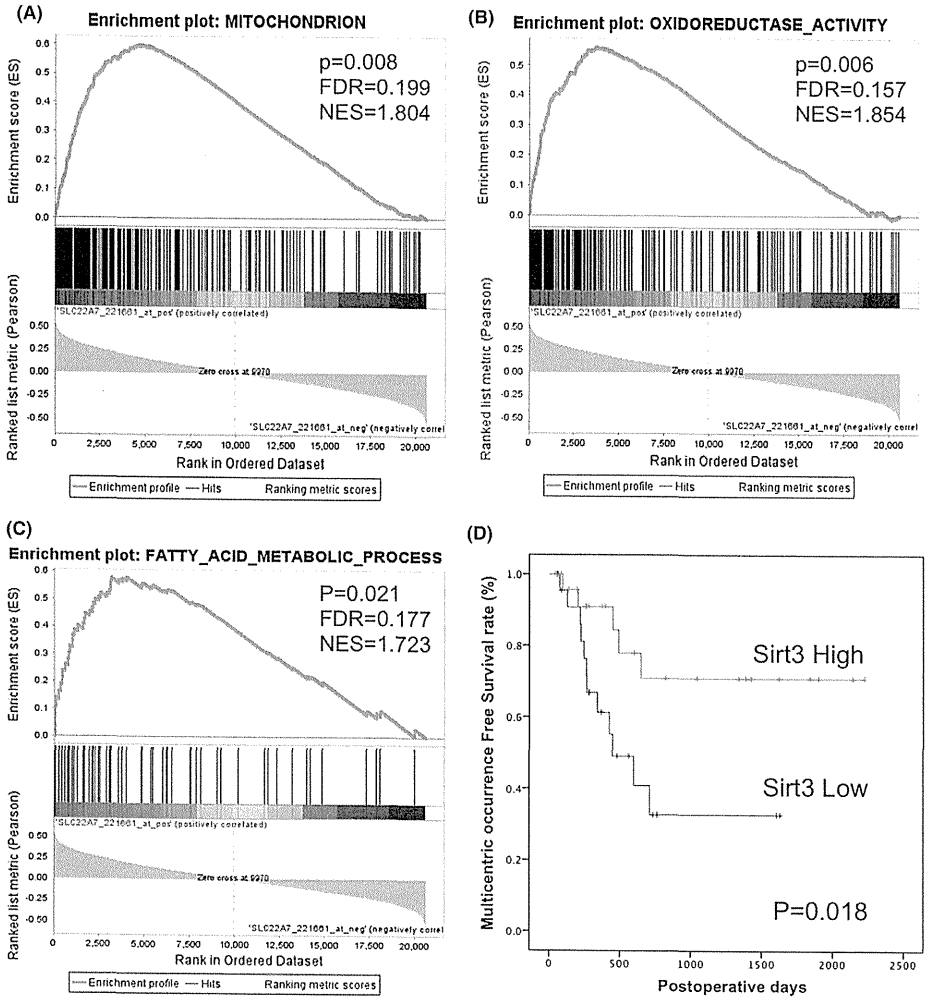


Fig. 2 GSEA evaluation associated with SLC22A7: **a** mitochondrion ($P = 0.008$; FDR = 0.199; NES = 1.804), **b** oxidoreductase activity ($P = 0.006$; FDR = 0.157; NES = 1.854), and **c** fatty acid metabolic process ($P = 0.021$; FDR = 0.177; NES = 1.723). **d** Low expression of sirtuin 3 was associated with a poor prognosis ($P = 0.018$)

albumin levels, which were recognized as risk factors for the occurrence of HCC in univariate analyses of the training set (Table 3), did not decide the clinical outcome in the multicenter validation study (Table 4). Multivariate analysis of the training set and the validation study revealed the low SLC22A7 expression to be the only reliable factor for predicting MO of HCC. The correlation between SLC22A7 gene expression and platelet counts were 0.167 ($P = 0.246$), and 0.134 ($P = 0.355$), respectively (data not shown). There was no difference in SLC22A7 expression between virus negative patients, HBV positive patients and HCV positive patients ($P = 0.439$). The SLC22A7 expression may be independent of platelet count and serum albumin decrease, predicting another aspect of liver functional reserve.

The precise indicator derived from noncancerous liver tissue determined the prognosis, which was not governed by tumor progression. According to the genome-wide gene expression analysis in the training set (Table 1), SLC22A7 best determined the clinical outcome in the organic anion transporter genes (GO:0015711), as judged by Cox regression analysis in ($P = 0.001$). Figure 1a shows the significant difference in tumor-free survival after hepatectomy according to SLC22A7 expression ($P = 0.001$).

The information obtained in the training set was validated in a prospective multicenter study using tissue microarrays (Fig. 1b). To this end, low SLC22A7 expression certainly determined MFS (Table 4). Regarding occurrence-free survival in the present study, de novo HCC may occur at 1 year after hepatectomy in patients with low SLC22A7 expression (Fig. 1). In this context, the prognostic curves appeared to be compatible with clinical observations. Anatomic resection did not decrease the risk for MFS in the multicenter study. It is reasonable that anatomic resection is not available in noncancerous liver with low SLC22A7 expression promoting de novo HCC. The aforementioned criteria were enough to determine the MO of HCC, since the anatomic resection prevents local recurrence within Milan criteria.

Whether oxidative stress resulting from reactive oxygen species in noncancerous tissue or cellular mitochondrial dysfunction promotes hepatocarcinogenesis has been discussed [18, 19]. Antioxidants such as glutathione play an important role and serve as essential components of the detoxification mechanism [9]. Our previous study identified CYP1A2 as an index for HCC recurrence [18]. The CYP1A2 expression is significantly decreased in the steatotic liver induced with orotic acid [20, 21]. The CYP1A2 and SLC22A7 are regulated by interferon-alpha 2b in human primary hepatocytes [22]. Interferon-alpha 2b induced partial remission of hepatoma [23]. These reports imply the possibility that CYP1A2 and SLC22A7 down regulation, are an early alert symptom of MO. In this aspect, decreased SLC22A7 expression may serve as a reliable surrogate biomarker for the prognosis and treatment of HCC.

Organic anion transporters are responsible for the uptake and exclusion of xenobiotics and organic anions [24, 25]. Serum organic anions and xenobiotics are emptied into the Disse's space, taken up by transporters at the hepatocellular sinusoidal membrane, and detoxified in the cytoplasm [26, 27]. The SLC22A7 expressed on the hepatocellular sinusoidal membrane takes up orotic acid [11]. Orotic acid has been regarded as a promoter of liver carcinogenesis, although the detailed mechanisms are unknown [28–30]. Moreover, mitochondrial sirtuin 3 may be involved in the metabolic cycle of orotic acid. Sirtuin 3, by regulating OTC activity to decrease orotic acid, inhibits hepatocellular carcinoma cell growth [31]. Previous research reported that exposure to orotic acid after hepatectomy promotes liver carcinogenesis [12]. Orotic aciduria and encephalopathy were observed in HCC patients without liver cirrhosis [32]. We focused on Sirtuin 3 regulating orotic acid production, because SLC22A7 transports orotic acids. Sirtuin 3 was reported as the regulator of mitochondrial metabolism and the inhibitor of hepatocellular carcinoma cell growth. The present findings indicate that the decreased sirtuin3 expression coinciding with decreased SLC22A7 may regulate hepatocellular orotic acid concentrations to promote MO of HCC (Fig. 2d). The gene expression levels of SLC22A7 correlated with that of sirtuin3. Furthermore, the customized sirtuin 3-related gene set revealed significant correlation with SLC 22A7 expression (Supplementary Figs. 2, 3).

In conclusion, the down regulation of SLC22A7 in noncancerous liver tissue may have promoted the MO of early-stage HCC in the training and multicenter validation studies. Evaluating SLC22A7 expression may be useful for selecting treatment strategies. Further research is required to determine whether the hepatocellular mechanisms involving mitochondrial metabolism increase hepatocellular liver carcinogenesis. Such studies could address whether antioxidant therapy or another therapy to prevent hepatocarcinogenesis would become available in patients with low SLC22A7 expression.

Acknowledgments This work was supported by a Health and Labour Sciences Research Grant (H20-Kanen-Ippan-001) from the Ministry of Health, Labour, and Welfare of Japan and a Grant-in Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science, and Technology of Japan. The authors thank Hiromi Ohnari and Ayumi Shioya for clerical and technical assistance.

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Yang JD, Roberts LR. Hepatocellular carcinoma: a global view. Nat Rev Gastroenterol Hepatol. 2010;7:448–58.
2. Schlitt HJ, Schnitzbauer AA. Hepatocellular carcinoma: agents and concepts for preventing recurrence after curative treatment. Liver Transpl. 2011;17(Suppl 3):S10–2.
3. Utsunomiya T, Shimada M, Imura S, Morine Y, Ikemoto T, Mori M. Molecular signatures of noncancerous liver tissue can predict the risk for late recurrence of hepatocellular carcinoma. J Gastroenterol. 2010;45:146–52.
4. Japan LCSGo: The general rules for the clinical and pathological study of primary liver cancer (in japanese). 5th ed. Tokyo: Kanehara; 2009. p. 43.

5. Arii S, Yamaoka Y, Futagawa S, Inoue K, Kobayashi K, Kojiro M, et al. Results of surgical and nonsurgical treatment for small-sized hepatocellular carcinomas: a retrospective and nationwide survey in japan. The liver cancer study group of japan. Hepatology. 2000;32:1224–9.
6. Hasegawa K, Kokudo N, Imamura H, Matsuyama Y, Aoki T, Minagawa M, et al. Prognostic impact of anatomic resection for hepatocellular carcinoma. Ann Surg. 2005;242:252–9.
7. Imamura H, Matsuyama Y, Tanaka E, Ohkubo T, Hasegawa K, Miyagawa S, et al. Risk factors contributing to early and late phase intrahepatic recurrence of hepatocellular carcinoma after hepatectomy. J Hepatol. 2003;38:200–7.
8. Kobayashi A, Miyagawa S, Miwa S, Nakata T. Prognostic impact of anatomical resection on early and late intrahepatic recurrence in patients with hepatocellular carcinoma. J Hepatobiliary Pancreat Surg. 2008;15:515–21.
9. Kudo A, Kashiwagi S, Kajimura M, Yoshimura Y, Uchida K, Arii S, et al. Kupffer cells alter organic anion transport through multidrug resistance protein 2 in the post-cold ischemic rat liver. Hepatology. 2004;39:1099–109.
10. Sekine T, Cha SH, Tsuda M, Apiwattanakul N, Nakajima N, Kanai Y, Endou H. Identification of multispecific organic anion transporter 2 expressed predominantly in the liver. FEBS Lett. 1998;12:179–82.
11. Fork C, Bauer T, Golz S, Geerts A, Weiland J, Del Turco D, et al. Oat2 catalyses efflux of glutamate and uptake of orotic acid. Biochem J. 2011;436:305–12.
12. Laconi E, Vasudevan S, Rao PM, Rajalakshmi S, Pani P, Sarma DS. The development of hepatocellular carcinoma in initiated rat liver after a brief exposure to orotic acid coupled with partial hepatectomy. Carcinogenesis. 1993;14:2527–30.
13. Takayama T, Makuuchi M, Hirohashi S, Sakamoto M, Yamamoto J, Shimada K, et al. Early hepatocellular carcinoma as an entity with a high rate of surgical cure. Hepatology. 1998;28:1241–6.
14. Enomoto A, Takeda M, Shimoda M, Narikawa S, Kobayashi Y, Yamamoto T, et al. Interaction of human organic anion transporters 2 and 4 with organic anion transport inhibitors. J Pharmacol Exp Ther. 2002;301:797–802.
15. Hallows WC, Yu W, Smith BC, Devries MK, Ellinger JJ, Someya S, et al. Sirt3 promotes the urea cycle and fatty acid oxidation during dietary restriction. Mol Cell. 2011;41:139–49.
16. Hirschey MD, Shimazu T, Goetzman E, Jing E, Schwer B, Lombard DB, et al. Sirt3 regulates mitochondrial fatty-acid oxidation by reversible enzyme deacetylation. Nature. 2010;464:121–5.
17. Hoshida Y, Villanueva A, Kobayashi M, Peix J, Chiang DY, Camargo A, et al. Gene expression in fixed tissues and outcome in hepatocellular carcinoma. N Engl J Med. 2008;359:1995–2004.
18. Tanaka S, Mogushi K, Yasen M, Ban D, Kudo A, Arii S, et al. Oxidative stress pathways in noncancerous human liver tissue to predict hepatocellular carcinoma recurrence: a prospective, multicenter study. Hepatology. 2011;54:1273–81.
19. Marra M, Sordelli IM, Lombardi A, Lamberti M, Tarantino L, Giudice A, et al. Molecular targets and oxidative stress biomarkers in hepatocellular carcinoma: an overview. J Transl Med. 2011;9:171.
20. Su GM, Sefton RM, Murray M. Down-regulation of rat hepatic microsomal cytochromes p-450 in microvesicular steatosis induced by orotic acid. J Pharmacol Exp Ther. 1999;291:953–9.
21. Zhang WV, Ramzan I, Murray M. Impaired microsomal oxidation of the atypical antipsychotic agent clozapine in hepatic steatosis. J Pharmacol Exp Ther. 2007;322:770–7.
22. Chen C, Han YH, Yang Z, Rodrigues AD. Effect of interferon-alpha2b on the expression of various drug-metabolizing enzymes and transporters in co-cultures of freshly prepared human primary hepatocytes. Xenobiotica. 2011;41:476–85.
23. Locker GJ, Mader RM, Steiner B, Wenzl E, Zielinski CC, Steger GG. Benefit of interferon-alpha2b in a patient with unresectable hepatoma and chronic infection with hepatitis c virus. Eur J Gastroenterol Hepatol. 2000;12:251–3.
24. Kudo A, Ban D, Aihara A, Irie T, Ochiai T, Nakamura N, Tanaka S, Arii S. Decreased Mrp2 transport in severe macrovesicular fatty liver grafts. J Surg Res. 2012;178(2):915–21.
25. Ban D, Kudo A, Sui S, Tanaka S, Nakamura N, Ito K, et al. Decreased mrp2-dependent bile flow in the post-warm ischemic rat liver. J Surg Res. 2009;153:310–6.
26. Norimizu S, Kudo A, Kajimura M, Ishikawa K, Taniai H, Suematsu M, et al. Carbon monoxide stimulates mrp2-dependent excretion of bilirubin-ixalpha into bile in the perfused rat liver. Antioxid Redox Signal. 2003;5:449–56.
27. Sui S, Kudo A, Suematsu M, Tanaka S, Ito K, Arii S, et al. Preservation solutions alter mrp2-dependent bile flow in cold ischemic rat livers. J Surg Res. 2010;159:572–81.
28. Rao PM, Nagamine Y, Roomi MW, Rajalakshmi S, Sarma DS. Orotic acid, a new promoter for experimental liver carcinogenesis. Toxicol Pathol. 1984;12:173–8.
29. Laurier C, Tatematsu M, Rao PM, Rajalakshmi S, Sarma DS. Promotion by orotic acid of liver carcinogenesis in rats initiated by 1,2-dimethylhydrazine. Cancer Res. 1984;44:2186–91.
30. Denda A, Laconi E, Rao PM, Rajalakshmi S, Sarma DS. Sequential histopathological analysis of hepatocarcinogenesis in rats during promotion with orotic acid. Cancer Lett. 1994;82:55–64.
31. Zhang YY, Zhou LM. Sirt3 inhibits hepatocellular carcinoma cell growth through reducing mdm2-mediated p53 degradation. Biochem Biophys Res Commun. 2012;423:26–31.
32. Jeffers LJ, Dubow RA, Zieve L, Reddy KR, Livingstone AS, Neimark S, et al. Hepatic encephalopathy and orotic aciduria associated with hepatocellular carcinoma in a noncirrhotic liver. Hepatology. 1988;8:78–81.

ARTICLE

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.

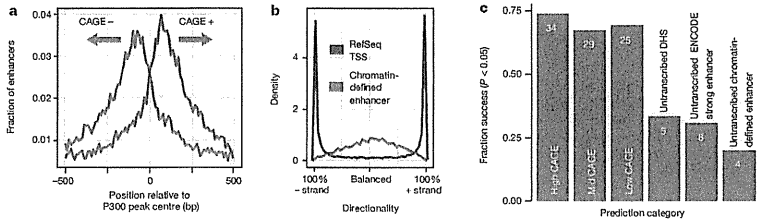


Figure 1 | Bidirectional capped RNAs is a signature feature of active enhancers. a, Enhancers identified by co-occurrence of H3K27ac and H3K4me1 ChIP-seq data[7], centred on P300 binding sites, in HeLa cells were overlaid with HeLa CAGE data (unique positions of CAGE tag 5′ ends, smoothed by a 5-bp window), 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

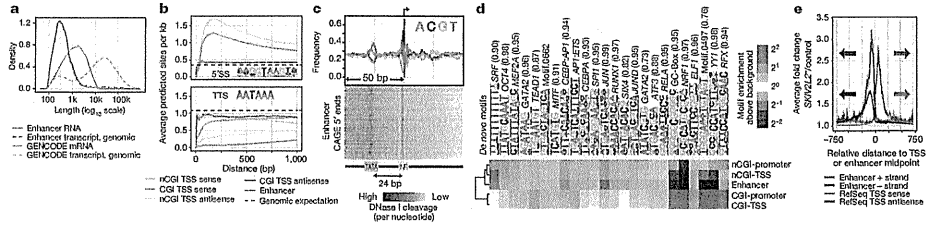


Figure 2 | Features distinguishing enhancer TSSs from mRNA TSSs. a, Densities of the genomic and processed RNA lengths of transcripts starting from enhancer TSSs and mRNA TSSs using assembled RNA-seq reads from 13 pooled FANTOM5 libraries. b, Frequencies of RNA processing motifs (5′ splice motif (5′SS, 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

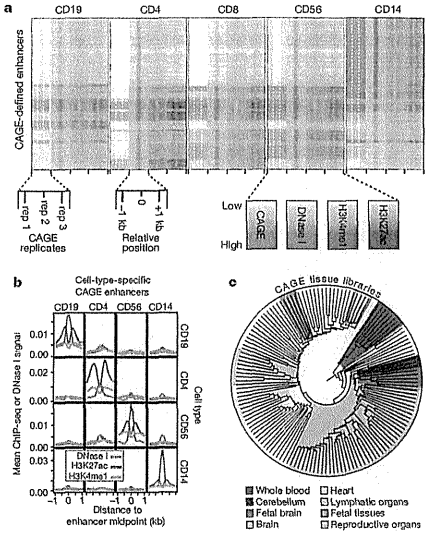


Figure 3 | CAGE expression identifies cell-type-specific enhancer usage. a, Relationship between CAGE and histone modifications in blood cells. Rows represent CAGE-defined enhancers that are ordered based on hierarchical clustering of CAGE expression. Columns for the CAGE tags (pink) represent the expression 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).

identified above) were tested via Tol2-mediated transgenesis in zebrafish embryos. We observed tissue-specific enhancer activity with 3 of 5 fragments, which corresponded to the human enhancer tissue expression (Fig. 4). None of three control fragments without CAGE signal activated the *GATA2* promoter (Supplementary Table 9). Although the sample size is not high enough to reliably estimate the validation rates in zebrafish, the correlation between the enhancer usage profiles in zebrafish to those defined in human by CAGE is notable.

We grouped the primary cell and tissue samples into larger, mutually exclusive cell type and organ/tissue groups (referred to as facets), respectively, with similar function or morphology (Supplementary Tables 10 and 11). Figure 5 summarizes how many enhancers were detected in each facet and the degree of facet-specific CAGE expression (see also Supplementary Fig. 21). From the data we can draw several conclusions:

First, the majority of detected enhancers within any facet are not restricted to that facet. Exceptions, where facets use a higher fraction of specific enhancers, include immune cells, neurons, neural stem cells and hepatocytes amongst the cell-type facets, and brain, blood, liver and testis amongst the organ/tissue facets.

Second, despite their apparent promiscuity, enhancers are more generally detected in a much smaller subset of samples than mRNA transcripts (Supplementary Figs 21 and 22a, b), consistent with cell-line studies[7] and the higher specificity of ncRNAs in general[13]. Facets in which we detect many enhancers typically also have a higher fraction of facet-specific enhancers (Supplementary Fig. 22c, d).

Third, the number of detected expressed enhancers and mRNA transcripts is correlated (Supplementary Fig. 21b), but the number of detected expressed gene transcripts (>1 tag per million mapped reads (TPM)) is 19–34 fold larger than the number of detected enhancers

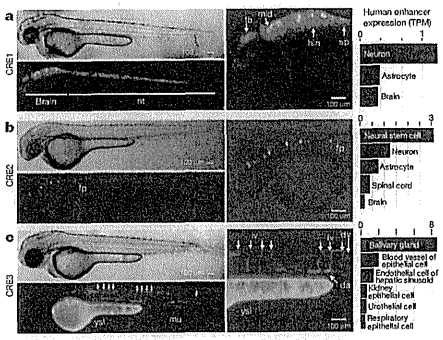


Figure 4 | *In vivo* validation in zebrafish of tissue-specific enhancers. Validations of *in vivo* activity of CAGE-defined human enhancers CRE1, CRE2 and CRE3 in zebrafish embryos at long-pec stage. Each panel shows, from left to right, representative yellow fluorescent protein (YFP) and bright field images of embryos injected with the human enhancer *gata2* promoter reporter gene construct (left), YFP zoom-ins (middle) and CAGE expression in TPM in human tissues/cell types for the enhancer (right). 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. Note the correspondence between zebrafish and human enhancer usage/expression. Supplementary Fig. 20 shows UCSC browser images of each selected enhancer. a, CRE1, ~230 kb 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, 5 kb upstream of the *POU3F2* gene, is active in the floor plate (fp). c, CRE3, 10 kb upstream of the *SOX7* gene TSS, shows specific expression in the vasculature (including intersegmental vessels (iv), dorsal vein (dv) and dorsal aorta (da).

with the cut-offs used. Noteworthy exceptions include blood and immune cells, testis, thymus and spleen, which have high enhancer/gene ratios. Conversely, smooth and skeletal muscle and skin, bone and epithelia-related cells have low ratios. Differential exosome activity between cell types might affect these results, but there was no correlation between *SKIV2L2* mRNA expression and the number of enhancers detected (Supplementary Fig. 22e, f).

As expected, consensus motifs of known key regulators are over-represented in corresponding facet-specific enhancers, 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). Notably, the AP1 motif appears to be enriched across all facets, perhaps associated with a general role for AP1 in regulating open chromatin[19].

Expression clustering reveals ubiquitous enhancers

Hierarchical clustering of enhancers by facet expression revealed a small subset of enhancers (200 or 247, defined by primary cell or tissue facets, respectively) expressed in the large majority of facets (Supplementary Text, Supplementary Figs 24 and 25, and Supplementary Tables 12 and 13). Compared to other enhancers, these ubiquitous (u-) enhancers are 8 times more likely to overlap CGIs and they are twice as conserved (Supplementary Fig. 26a–c). U-enhancers overlap typical chromatin enhancer marks but have higher H3K4me3 signal (Supplementary Fig. 26d). Although they produce significantly longer ncRNAs than other enhancers (median 530 nucleotides, $P < 1.5 \times 10^{-8}$, Mann–Whitney *U* test), the transcripts remain predominantly (~78%) unspliced and significantly shorter ($P < 4.2 \times 10^{-18}$, Mann–Whitney *U* test) than mRNAs (Supplementary Figs 27 and 28), do not share exons with known genes, and are exosome-sensitive (Supplementary Fig. 14b). Therefore, it is unlikely that these are novel mRNA promoters. They are also highly enriched for P300 and cohesin ChIP-seq peaks[20] and RNAPII-mediated ChIA-PET signal[21] compared to other enhancers (Supplementary Fig. 26d). These results indicate that u-enhancers comprise a small but distinct subset of enhancers, which probably has specific regulatory functions used by virtually every human cell.

Linking enhancer usage with TSS expression

A major challenge is to link enhancers to their target genes[21,22]. Uniquely, FANTOM5 CAGE allows for direct comparison between transcriptional activity of the enhancer and of putative target gene TSSs across a diverse set of human cells. Based on pairwise expression correlation, nearly half (40%) of the inferred TSS-associated enhancers (Methods) were linked with the nearest TSS, and 64% of enhancers have at least one correlated TSS within 500 kilobases. Several associations (10,260; 15.3%) are supported by ChIA-PET (RNAPII-mediated) interaction data[21], and the supported fraction increases with the correlation threshold (Supplementary Fig. 29a). The fraction of supported associations is 4.8-fold higher than that of associations predicted from DNase I hypersensitivity correlations[10] (20.6% versus 4.3%, at the same correlation threshold), indicating that transcription is a better predictor of regulatory targets than chromatin accessibility. Conserved sequence motifs and ChIP-seq peaks also co-occurred significantly in associated enhancer-promoter pairs (Benjamini–Hochberg false discovery rate (FDR) < 0.05, binomial test), suggesting an additive or synergistic cooperation between enhancers and promoters at RNAPII foci.

On average, a RefSeq TSS was associated with 4.9 enhancers and an enhancer with 2.4 TSSs and we observed different regulatory architectures around genes (Supplementary Fig. 30). For example, at the beta-globin locus the CAGE expression patterns of four locus control region hypersensitive sites are highly correlated (Pearson's *r* between 0.88 and 0.98) with the expression of known target genes[23,24] *HBG2* and *HBD*, and to some extent *HBG1*.

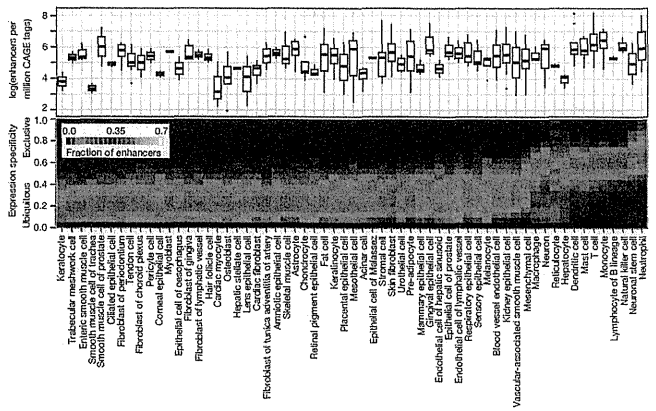


Figure 5 | Enhancer usage and specificity in groups of cells. The upper panel gives the number of detected enhancers per million CAGE tags within each group (facet) of related cell type libraries. The expression specificity of the enhancers is shown as a heat map in the panel below. Colours show the fraction of expressed enhancers in each facet (columns) that are in each specificity range (rows). For corresponding plots on organ/tissue facets and genes, see Supplementary Fig. 21.

These observations call for computational models of enhancer regulation, in which multiple enhancers may work in concert to enhance the expression of a gene. To this end, we focused on 2,206 RefSeq TSSs for which the joint expression of nearby enhancers (the closest ten enhancers within 500 kb) is highly predictive of the gene expression. Model shrinkage showed that in most cases, only one to three enhancers are necessary to explain the expression variance observed in the linked gene, and generally proximal enhancers are more predictive than distal ones (Fig. 6a, Supplementary Fig. 29b–d and Supplementary Text). One hypothesis explaining the function of multiple enhancers driving the

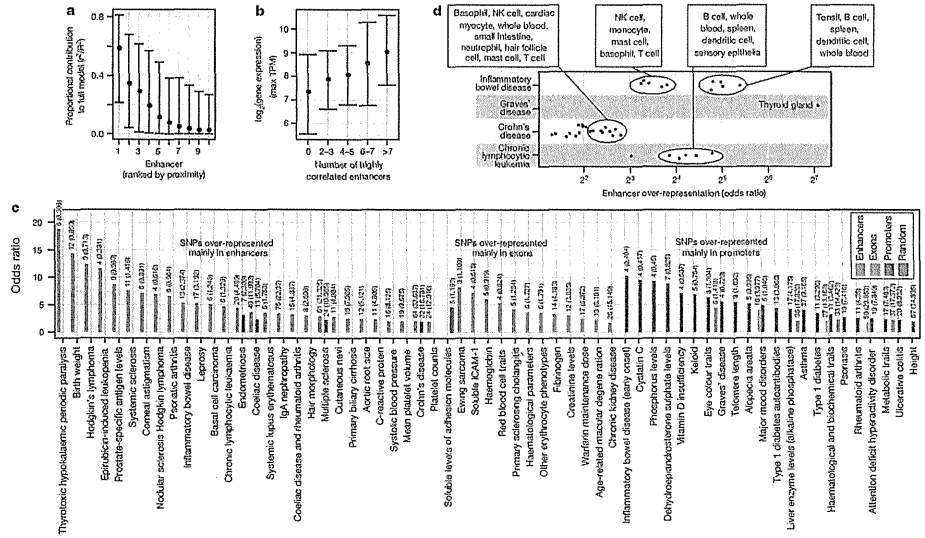


Figure 6 | Linking enhancers to TSSs and disease-associated SNPs. a, The proportional contribution (see Methods) of the 10 most proximal enhancers within 500 kb of a TSS in a model explaining gene expression variance (vertical axis) as a function of enhancer expression. *x* axis indicates the position of the enhancer relative to the TSS: 1 the closest, etc. Bars indicate interquartile ranges and dots medians. **b**, Relationship between the number of highly correlated ('redundant') enhancers per locus (horizontal axis) and the maximal expression (TPM) of the associated TSS in the same model over all CAGE libraries (vertical axis). Error bars as in a. **c**, GWAS SNP sets preferentially overrepresented within enhancers, exons and mRNA promoters. Observed and expected overlaps are shown above bars. The vertical axis gives enrichment odds ratios. The horizontal axis shows GWAS traits or diseases. **d**, Diseases with GWAS-associated SNPs over-represented in enhancers of certain expression facets. The horizontal axis gives the odds ratio as in panel c, broken up by expression facets: each point represents the odds ratio of GWAS SNP enrichment for a disease (vertical axis) in a specific expression facet. Summary annotations of point clouds are shown. See also Supplementary Fig. 31.

same expression pattern is that they might confer higher transcriptional output of a gene[25,26]. Indeed, the number of highly correlated (redundant) enhancers close to TSSs (Supplementary Methods) increased with the observed maximal TSS expression over all libraries (Fig. 6b), implying that these enhancers are redundant in terms of transcription patterns but additive in terms of expression strength. Expression redundancy is also common in genomic clusters of closely spaced enhancers (24% of 815 identified genomic clusters, Supplementary Table 15). These are associated with TSSs of genes involved in immune and defence responses and, as suggested by a previous study[27], have a higher expression than other enhancer-associated genes (eightfold increase on average).

Disease-associated SNPs are enriched in enhancers

Many disease-associated SNPs are located outside of protein-coding exons and a large proportion of human genes display expression polymorphism[28]. Using the NHGRI genome-wide association studies (GWAS) catalogue[29] and extending the compilation of lead SNPs with proxy SNPs in strong linkage disequilibrium (similar to refs 30, 31), we identified diseases/traits whose associated SNPs overlapped enhancers, promoters, exons and random regions significantly more than expected by chance (Fisher's exact test $P < 0.01$, Supplementary Table 16). Disease-associated SNPs were over-represented in regulatory regions to a greater extent than in exons (Fig. 6c). For many traits where enriched disease-associated SNPs were within enhancers, enhancer activity was detected in pathologically relevant cell types (Fig. 6d and Supplementary Figs 31 and 32). Examples include Graves' disease-associated SNPs enriched in enhancers that are expressed predominantly in thyroid tissue, and similarly lymphocytes for chronic lymphocytic leukaemia. As a proof of concept, we validated the impact of two disease-associated regulatory SNPs within enhancers (Supplementary Fig. 33).

Conclusions

The data presented here demonstrate that bidirectional capped RNAs, as measured by CAGE, are robust predictors of enhancer activity in a cell. Transcription is only measured at a fraction of chromatin-defined enhancers and few untranscribed enhancers show potential enhancer activity. This implies that many chromatin-defined enhancers are not regulatory active in that particular cellular state, but may be active in other cells of the same lineage[32] or are pre-marked for fast regulatory activity upon stimulation[33]. Of course, given the relative instability of enhancer RNAs some chromatin-defined sites may be active but fall below the limits of detection of CAGE.

Our results show that position-specific sequence signals upstream of the transcription initiation sites and the production of small, uncapped RNAs immediately downstream is present at both enhancers and mRNA promoters, suggesting similar mechanisms of initiation. Previous studies (for example refs 10, 34, 35) suggested that promoters and enhancers differ in motif composition. This view is not supported by the larger FANTOM5 data set. Instead, the differences reflect the local G+C content because transcribed enhancers tend to harbour low G+C content motifs like non-CGI promoters. Features distinguishing enhancers from mRNA promoters are (1) enhancer RNAs are exosome-sensitive regardless of direction whereas (sense) mRNAs have a longer half-life than their antisense counterpart; (2) enhancer RNAs are short, unspliced, nuclear and non-polyadenylated and (3) enhancers have downstream polyadenylation and 5′ splice motif frequencies at genomic background level similar to antisense PROMPTs, whereas mRNAs are depleted of termination signals and enriched for 5′ splice sites[11,12].

The collection of active enhancers presented here provides a resource that complements the activity of the ENCODE consortium[7] across a much greater diversity of tissues and cellular states. It has clear applications in human genetics, to narrow the search windows for functional association, and for the definition of regulatory networks that underpin the processes of cellular differentiation and organogenesis in human development.

METHODS SUMMARY

Single-molecule HeliScopeCAGE data was generated as described elsewhere[6]. Sequencing and processing of ribosomal RNA-depleted RNAs, short RNAs and H3K27ac or H3K4me1 ChIPs as well as the processing of publicly available DNase-seq data are described in the Methods.

Putative enhancers were identified from bidirectionally transcribed loci having divergent CAGE tag clusters separated by at most 400 bp (described in Supplementary Fig. 6a). We required loci to be divergently transcribed in at least one FANTOM5 sample, defined by CAGE tag 5′ ends within 200 bp divergent strand-specific windows immediately flanking the loci midpoints. The expression of each enhancer in each FANTOM5 sample was quantified as the normalized sum of strand-specific sums of CAGE tags in these windows. A sample-set wide directionality score, D, for each locus over aggregated normalized reverse, R, and forward, F, strand window-expression values across all samples, $D = (F - R)/(F + R)$, were then used to filter putative enhancers to have low, non-promoter-like, directionality scores ($|D| < 0.8$). Further filtering ensured enhancers to be located distant to TSSs and exons of protein- and noncoding genes.

Motif enrichment analyses were done using HOMER[36]. Regulatory targets of enhancers were predicted by correlation tests using the sample-set wide expression profiles of all enhancer-promoter pairs within 500 kb. The regulatory effects of multiple enhancers were modelled using linear regression followed by lasso-based model-shrinkage[37].

Enhancer activity was tested *in vivo* in zebrafish embryos using Tol2-mediated transgenesis[38]. Expression patterns were documented at 48 h post fertilization using >200 eggs per construct. 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 (Supplementary Fig. 9d). Full details are provided in the Methods.

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; accepted 16 October 2013.

1. Bulger, M. & Groudine, M. Enhancers: the abundance and function of regulatory sequences beyond promoters. *Dev. Biol.* **339,** 250–257 (2010).
2. Lenhard, B., Sandelin, A. & Carninci, P. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nature Rev. Genet.* **13,** 233–245 (2012).
3. Banerji, J., Rusconi, S. & Schaffner, W. Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27,** 299–308 (1981).
4. Kim, T.-K. *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465,** 182–187 (2010).
5. Kodzius, R. *et al.* CAGE: cap analysis of gene expression. *Nature Methods* **3,** 211–222 (2006).
6. The FANTOM Consortium and the RIKEN PMI and CLST (DGT). A promoter-level mammalian expression atlas. *Nature* http://dx.doi.org/10.1038/nature13182 (this issue).
7. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).
8. Kheradpour, P. *et al.* Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res.* **23,** 800–811 (2013).
9. Fort, A. *et al.* Deep transcriptome profiling of mammalian stem cells supports a regulatory role for retrotransposons in pluripotency maintenance. *Nature Genet.* (in the press).
10. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489,** 75–82 (2012).
11. Ntini, E. *et al.* Polyadenylation site–induced decay of upstream transcripts enforces promoter directionality. *Nature Struct. Mol. Biol.* **20,** 923–928 (2013).
12. Almada, A. E., Wu, X., Kriz, A. J., Burge, C. B. & Sharp, P. A. Promoter directionality is controlled by U1 snRNP and polyadenylation signals. *Nature* **499,** 360–363 (2013).
13. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489,** 101–108 (2012).
14. Kowalczyk, M. S. *et al.* Intragenic enhancers act as alternative promoters. *Mol. Cell* **45,** 447–458 (2012).
15. Valen, E. *et al.* Biogenic mechanisms and utilization of small RNAs derived from human protein-coding genes. *Nature Struct. Mol. Biol.* **18,** 1075–1082 (2011).
16. Taft, R. J. *et al.* Tiny RNAs associated with transcription start sites in animals. *Nature Genet.* **41,** 572–578 (2009).
17. Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322,** 1845–1848 (2008).
18. Rönnerblad, M. *et al.* Analysis of the DNA methylome and transcriptome in granulopoiesis reveal timed changes and dynamic enhancer methylation. *Blood* http://dx.doi.org/10.1182/blood-2013-02-482893 (in the press).

19. Biddie, S. C. *et al.* Transcription factor AP1 potentiates chromatin accessibility and glucocorticoid receptor binding. *Mol. Cell* **43,** 145–155 (2011).
20. Schmidt, D. *et al.* A CTCF-independent role for cohesin in tissue-specific transcription. *Genome Res.* **20,** 578–588 (2010).
21. Li, G. *et al.* Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* **148,** 84–98 (2012).
22. Chepelev, I., Wei, G., Wangsa, D., Tang, Q. & Zhao, K. Characterization of genome-wide enhancer-promoter interactions reveals co-expression of interacting genes and modes of higher order chromatin organization. *Cell Res.* **22,** 490–503 (2012).
23. Fraser, P., Pruzina, S., Antoniou, M. & Grosveld, F. Each hypersensitive site of the human beta-globin locus control region confers a different developmental pattern of expression on the globin genes. *Genes Dev.* **7,** 106–113 (1993).
24. Dostie, J. *et al.* Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements. *Genome Res.* **16,** 1299–1309 (2006).
25. Barolo, S. Shadow enhancers: frequently asked questions about distributed *cis*-regulatory information and enhancer redundancy. *Bioessays* **34,** 135–141 (2012).
26. Schaffner, G., Schirm, S., Müller-Baden, B., Weber, F. & Schaffner, W. Redundancy of information in enhancers as a principle of mammalian transcription control. *J. Mol. Biol.* **201,** 81–90 (1988).
27. Whyte, W. A. *et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153,** 307–319 (2013).
28. Göring, H. H. H. *et al.* Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nature Genet.* **39,** 1208–1216 (2007).
29. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA* **106,** 9362–9367 (2009).
30. Ward, L. D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* **40,** D930–D934 (2012).
31. Maurano, M. T., Wang, H., Kutyavin, T. & Stamatoyannopoulos, J. A. Widespread site-dependent buffering of human regulatory polymorphism. *PLoS Genet.* **8,** e1002599 (2012).
32. Mercer, E. M. *et al.* Multilineage priming of enhancer repertoires precedes commitment to the B and myeloid cell lineages in hematopoietic progenitors. *Immunity* **35,** 413–425 (2011).
33. Ostuni, R. *et al.* Latent enhancers activated by stimulation in differentiated cells. *Cell* **152,** 157–171 (2013).
34. Rada-Iglesias, A. *et al.* A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470,** 279–283 (2011).
35. Shen, Y. *et al.* A map of the *cis*-regulatory sequences in the mouse genome. *Nature* **488,** 116–120 (2012).
36. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38,** 576–589 (2010).
37. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33,** 1–22 (2010).
38. Gehrig, J. *et al.* Automated high-throughput mapping of promoter-enhancer interactions in zebrafish embryos. *Nature Methods* **6,** 911–916 (2009).

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.H. and a Grant of the Innovative Cell Biology by Innovative Technology (Cell Innovation Program) from the MEXT, Japan, to Y.H. The A.S. group was supported by funds from the European Research Council FP7/2007-2013/ ERC no. 204135, the Novo Nordisk and Lundbeck foundations. Work in the M.R. group was funded by grants from the Deutsche Forschungsgemeinschaft (RE 1310/7,11,13) and Rudolf Bartling Stiftung. F.M. and I.M.E. were supported by "BOLD" Marie Curie ITN and "ZF- Health" Integrated project of the European Commission. We thank S. Noma, M. Sakai and H. Tarui for RNA-seq and sRNA-seq preparation, RIKEN GeNAS for generation and sequencing of the Heliscope CAGE libraries, Illumina RNA-seq and sRNA-seq, the Copenhagen National High-throughput DNA Sequencing Center for Illumina CAGE-seq, M. Edinger, P. Hoffmann and R. Eder for cell sorting, A. Albrechtsen, I. Moltke, W. Wasserman for advice, and the Netherlands Brain Bank for post-mortem human brain material.

Author Contributions R.A., I.H., E.A., E.V., K.L., Y.C., B.L., X.Z., M.J., H.K., T.F.M., T.L., N.B., O.R., A.M.B., J.K.B, C.J.M, N.R., F.O.B., M.R., A.S. made the computational analysis. J.B., M.B., T.L., H.K., N.K., J.K., H.S., M.I., C.O.D, A.R.R.F., P.C., Y.H. prepared and pre-processed CAGE and/or RNA-seq libraries. E.N., P.R.A., T.H.J., J.B., M.B. made the knockdown experiments followed by CAGE. C.G., C.S., L.S., J.R., D.G., M.R. made the blood cell ChIP experiments, methylation assays and *in vitro* blood cell validations. T.S., C.G., Y.I., Y.S., E.F., S.M., Y.N., A.R.R.F., P.C. and H.S. made the HeLa/HepG2 *in vitro* validations. I.M.E., R.A., A.S., F.M. designed and carried out zebrafish *in vivo* tests. R.A., C.G., I.H., C.S., E.A., E.V., F.M., I.M.E., P.C., A.R.R.F, M.B., J.B., A.L., C.D., D.A.H., P.H., M.R., A.S. interpreted results. R.A., C.G., I.H., E.V., I.M.E., J.B., F.M., D.A.H., M.R., A.S. wrote the paper with input from all authors. M.R. and A.S. coordinated and supervised the project.

Author Information The FANTOM5 atlas is accessible from http://fantom.gsc.riken.jp/5. FANTOM5 CAGE, RNA-seq and sRNA data have been deposited in DDBJ/EMBL/GenBank (accession codes DRA000991, DRA001101). Genome browser tracks for enhancers with user-definable expression specificity-constraints can be generated at http://enhancer.binf.ku.dk. Here, processed enhancer expression data, predefined enhancer tracks and motif finding results are also deposited. Blood-cell ChIP-seq data and CAGE data on exosome-depleted HeLa cells have been deposited in the NCBI GEO database (accession codes GSE40668, GSE49834). 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), M.R. (michael.rehli@ukr.de) or A.S. (albin@binf.ku.dk).

METHODS

CAGE data. Single molecule HeliScopeCAGE[39] data was generated as described elsewhere[6]. We used a set of 432 primary cell, 135 tissue and 241 cell line samples that passed quality control measures of >500,000 Q20 Delve (T.L. *et al.*, manuscript in preparation) mapped CAGE tags, RNA integrity and reproducibility (for further details, see ref. 6).

Proof of concept analysis. We defined silent and active enhancers from ENCODE HeLa-S3, GM12878 and K562 broad peaks, downloaded from the UCSC ENCODE repository, according to the co-existence of histone modifications H3K4me1, H3K27ac and H3K27me3. Active enhancers were defined as co-localized H3K4me1 and H3K27ac peaks with no H3K27me3 peak, whereas silent enhancers were considered loci with H3K4me1 and H3K27me3 peaks but no H3K27ac peak. Loci were filtered to be located distant to TSSs (500 bp) and exons (200 bp) of protein-coding genes, multi-exonic noncoding genes and mRNAs (from ENSEMBL, GENCODE (v10), RefSeq and UCSC, downloaded January 12, 2012), and other lncRNAs from a gene-centric set derived from literature[16] as well as manually annotated sense-antisense pairs (coding-noncoding and noncoding-noncoding sense-antisense pairs) with 5′ expressed sequence tag (EST) and complementary DNA support, and 5′ ESTs with no locus with protein-coding capacity. Transcriptional differences between active and silent enhancer sets were determined by comparing the average number of FANTOM5 CAGE tag 5′ ends from the same ENCODE cell lines (pooled triplicates) in a window ± 300 bp around the H3K4me1 peak mid points.

The active enhancer sets of HeLa-S3, GM12878 and K562 cells were then centred on proximal (within 200 bp) P300 ENCODE binding site peaks (joint P300 and GATA1 peaks for K562) to derive centre positions. FANTOM5 CAGE data from the same ENCODE cell lines (pooled triplicates) were then overlaid over these centred enhancer regions and the absence (0) and presence (1) of (one or more) CAGE tag 5′ ends in 10 bp non-overlapping windows were determined and an average profile was calculated to assess the average bidirectional pattern of transcription at chromatin-derived enhancers.

Pooled CAGE data from all FANTOM5 libraries (described above) were further overlaid with these regions and a directionality score based on the aggregate of CAGE tags falling within ± 300 bp from the centre positions were calculated to determine potential strand bias. For comparison, we repeated the same calculations for genomic regions ± 300 bp around TSSs of RefSeq protein coding genes. Directionality was calculated as (F - R) / (F + R), where F and R is the sum of CAGE tags aligned on the forward and reverse strand, respectively. Directionality close to -1 or 1 indicates a unidirectional behaviour while 0 indicates perfectly balanced bidirectional transcription.

Positional cross correlations were calculated between reverse and forward CAGE tag 5′ ends at ChIP-seq-derived active HeLa-S3 and GM12878 enhancer centre positions (as determined by P300 peaks) ± 300 bp (maximum lag 300) to identify their most likely separation. Cross correlations were also calculated in 300 bp windows (maximum lag 150) flanking the enhancer centres between CAGE 5′ ends and ENCODE H2A.Z signals (from the same cell line) for HeLa-S3 and GM12878 as well as between CAGE 5′ ends and 5′ ends of ENCODE GM12878 micrococcal nuclease-digested nucleosome sequencing (MNase-seq) reads (9 pooled replicates). In the latter analysis, correlations were made using reads on the same strand. Pooled, unique CAGE tags (in which only one CAGE tag per bp was counted) were considered in all correlation analyses and enhancers were weighed according to the aggregated signal before subsequent averaging over lags not to make any library or enhancer have an undue influence.

Reporter activity of ENCODE enhancers in relation to transcriptional status. We used published[9] results on a massively parallel reporter assay measuring the activity of ENCODE-predicted enhancers in HepG2 and K562 cells. All results on non-scrambled sequences were considered, regardless of the level of conservation. 198 out of 738 tested K562 enhancers and 307 out of 1,136 tested HepG2 enhancers had significant enhancer reporter activity (as determined by the original publication). We determined the expression in 401 bp windows centred on mid points of ENCODE-predicted enhancers using FANTOM5 CAGE from the same cell lines. We further calculated the false discovery rate after a minimum expression threshold in the interval [0,0.5] TPM, as the fraction of non-significant enhancers among those fulfilling the expression cutoff.

Identification of bidirectionally transcribed loci. Bidirectionally transcribed loci were defined from a set of 1,714,047 forward and 1,597,186 reverse strand CAGE tag clusters (TCs) supported by at least two CAGE tags in at least one sample (TCs defined in ref. 6). Only TCs not overlapping antisense TCs were used. We identified 1,261,036 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), which resulted in 200,171 bidirectional loci (procedure illustrated in Supplementary Fig. 6a). A centre position was defined for each bidirectional locus as the mid position between the rightmost reverse strand 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 centre 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 centre) 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. These filtering steps resulted in 78,555 bidirectionally transcribed loci.

Expression quantification of bidirectionally transcribed loci and prediction of enhancers. We quantified the expression of bidirectional loci for each strand and 200 bp flanking window in each of the 432 primary cell, 135 tissue and 241 cell line samples separately by counting the CAGE tags whose 5′ ends were located within these windows. The expression values of both flanking windows were normalized by converting tag counts to tags per million mapped reads (TPM) and further normalization between samples was done using the relative log expression (RLE) normalization procedure in edgeR[41]. The number of CAGE tags aligned on ChrM was subtracted from the total number of aligned CAGE tags in each library before normalization. The normalized expression values from both windows were used to calculate a sample-set wide directionality score, D, for each enhancer over aggregated normalized reverse, R, and forward, F, strand expression values across all samples (Supplementary Fig. 6a); $D = (F - R)/(F + R)$. D ranges between -1 and 1 and specifies the bias in expression to reverse and forward strand, respectively ($D = 0$ means 50% reverse and 50% forward strand expression, while $|D|$ close to 1 indicates unidirectional transcription). A directionality score calculated from pooled data is a good estimate of sample directionality (Supplementary Fig. 6b). Each bidirectional locus was assigned one expression value for each sample by summing the normalized expression of the two flanking windows.

Bidirectional loci were further filtered to have low, non-promoter-like, directionality scores ($|D| < 0.8$) and to be located distant to TSSs and exons of protein- and noncoding genes (see 'Proof of concept analysis' above for details). This resulted in a final set of 43,011 putative enhancers.

We further tested whether the expression level for each sample and candidate enhancer was significantly greater than the genomic background (see construction of random genomic background regions below). A P value was calculated for each enhancer expression value for each primary cell, tissue and cell line sample by counting the fraction of random genomic regions with greater expression level in the same sample. Enhancers with P values less than 0.001 and Benjamini–Hochberg adjusted FDR < 0.05 was considered transcribed in that sample. This analysis yielded binary expression values, which were used for constructing enhancer sets associated with each sample. In total, 38,554 enhancers were transcribed at a significant expression level in at least one primary cell or tissue sample. Below, we refer to this set as the 'robust set' of enhancers and indicate whenever it was used. For all analyses, we use the whole ('permissive') set of 43,011 enhancers if not otherwise mentioned.

Construction of random genomic background regions. We randomly sampled 100,000 genomic regions of 401 bp that were distal to TSSs and exons of known genes (same as the filtering procedure described above for bidirectionally transcribed loci). These were further filtered to not overlap with our set of 43,011 predicted enhancers, which yielded 98,942 random genomic regions whose expression levels were quantified and normalized in the same manner as described for bidirectional loci (above).

Correlation between ENCODE epigenomic data and CAGE-defined enhancers. Using the UCSC ENCODE repository data (downloaded and pooled 26 March 2012), we assessed the signal of RNA Polymerase II (RNAPII), the pooled transcription factor super track (all TFs), CCCTC-binding factor (CTCF), E1A binding protein P300, DNase I hypersensitive sites (DHSs) and two histone marks: H3K4me1 and H3K27ac around enhancers, TSSs and random genomic sites.

Large scale enhancer reporter validations. We randomly selected 125 CAGE-defined enhancers with significantly higher expression than random genomic regions in at least two out of three HeLa-S3 replicates. These were grouped according to Hela-S3 expression tertiles: low (36), mid-level (41) and strong (46). These could be split up further according to overlap (mid position) with combined ENCODE (release January 2011) segmentations of Segway[42] and ChromHMM[43] chromatin state prediction: 25, 27 and 14 strongly, mid-level and lowly expressed CAGE enhancer overlapped ENCODE state 'E' ('strong enhancer') whereas 21, 16 and 22 strongly, mid-level and lowly expressed CAGE enhancer overlapped ENCODE state 'TSS'.

We further randomly selected 26 and 15 untranscribed (negligible amount of overlapping FANTOM5 HeLa-S3 CAGE tags) 500 bp regions centred on mid

positions of HeLa-S3 E states and HeLa-S3 ENCODE DHSs. Two literature-derived[44] HeLa-S3 positive enhancers and 4 random regions (see 'Construction of random genomic background regions') were used for comparison. For comparison, we also randomly selected 20 manually defined untranscribed HeLa-S3 chromatin-defined active enhancers (see 'Proof of concept analysis').

PCR primers for the amplification of enhancer and control regions were designed using the PerlPrimer tool[45], and purchased from Operon Ltd Primers included BamHI or SalI restriction sites for cloning and sequences are listed in Supplementary Tables 2 and 3. Control fragments ranged between 420 and 1,452 bp. Enhancer fragments usually included a 500 bp window around the mid-point of our predicted enhancers and depending on the availability of unique primer sequences, enhancer fragments ranged between 470 and 840 bp.

We inserted an EF1α basal promoter fragment into HindIII and NheI sites of the multiple cloning site in the promoter-less pGL4.10 (Promega) to construct the pGL4.10EF1α vector. We next removed the BamHI and SalI containing fragment located downstream of the SV40 late poly(A) signal, and re-inserted the fragment at the SpeI site that is located upstream of the synthetic poly(A) signal/transcriptional pause site to generate modified versions of pGL4.10EF1α and pGL4.10 (see Supplementary Fig. 9d).

Enhancer and control regions were PCR-amplified using KOD plus polymerase (Toyobo) from HEK-293T gDNA, digested with BamHI and SalI (Takara Bio), and purified using the E-Gel SizeSelect system (Life Technologies). Five μl of purified PCR products were ligated with 100 ng of the BamHI- and SalI-digested modified pGL4.10EF1α and pGL4.10 plasmids using Ligation-high (Toyobo), and transformed into DH5α competent cells (Toyobo). Correct insertion of the PCR products into the plasmids was checked by colony PCR. Vectors were purified using the QIAGEN Plasmid Plus 96 Miniprep Kit (Qiagen).

HeLa-S3 cells (JCRB Cell Bank) were cultured in MEM (WAKO) supplemented with 10% FBS (Nichirei Bioscience Inc., lot no. 7G0031), 100 Units ml^{-1} penicillin and 100 μg ml^{-1} streptomycin (both Life Technologies). HepG2 Cells (RIKEN BRC) were cultured in DMEM (Life Technologies) supplemented with 10% FBS (Nichirei Bioscience Inc., Lot No. 7G0031), and MEM (WAKO) supplemented with 10% FBS (Nichirei Bioscience Inc., lot no. 7G0031), 100 Units penicillin and 100 μg ml^{-1} streptomycin (Life Technologies). Cell lines were seeded into 96 well plates at a density of 7.5×10^3 cells per well one day before transfection. Firefly luciferase reporter plasmids (190 ng) and 10 ng of pGL4.73 *Renilla* luciferase plasmid (Promega) were co-transfected into HepG2 or HeLa-S3 cells using Lipofectamine (Life Technologies) according to the manufacturer's instruction. Each transfection was independently performed three times. After 24 h, the luciferase activities were measured by GloMax 96 Microplate luminometer (Promega) using the Dual-glo luciferase assay system (Promega) according to the manufacturer's instruction.

Sequence motif analysis on global CAGE enhancer and promoter sets. To compare motif signatures characterizing bidirectionally transcribed enhancers (permissive set) with those of CAGE-defined promoters, we used the set of 184,827 robust human CAGE clusters defined by ref. 6 separated into 61,322 CGI and 123,505 nonCGI-associated clusters. We made further subsets of these CAGE clusters, contingent on their overlap with annotated TSSs from Refseq and Gencode. We merged overlapping extended CAGE clusters (−300, +50; based on the robust cluster set; average size nonCGI: 422 bp; average size CGI: 544 bp) contingent on CGI status and subtracted CAGE cluster regions that overlapped with extended enhancers (mid position ± 200 bp).

This created five sets of regions representing non-overlapping bidirectional enhancers, nonCGI promoters and CGI promoters (annotated and full sets for the two latter ones). Motif enrichment was analysed using HOMER[36] version 3, a suite of tools for motif discovery and next-generation sequencing analysis (http://biowhat.ucsd.edu/homer/). Sequences of the three region sets (enhancers, nonCGI and CGI promoters) were compared to equal numbers of randomly selected genomic fragments of the average region size, matched for GC content and auto-normalized to remove bias from lower-order oligo sequences. After masking repeats, motif enrichment was calculated using the cumulative binomial distribution by considering the total number of target and background sequence regions containing at least one instance of the motif. One hundred motifs were searched for a range of motif lengths (7–14 bp) resulting in a set of 800 *de novo* motifs per set. After filtering redundant motifs, the top 50 motifs resulting from each search were combined, remapped and ranked according to enrichment (depletion) in the enhancer set. In parallel, we also used HOMER to calculate the enrichment of ChIP-seq derived known transcription factor motifs. Motif collections including search parameters are deposited in a web database at http://enhancer.binf.ku.dk. Histograms of PhastCons scores were generated using the annotation tool in HOMER.

Analysis of splice site and termination signals downstream of CAGE enhancer TSSs and promoter TSSs. To identify motifs downstream of TSSs potentially differing between the structurally related bidirectionally transcribed enhancer TSSs and nonCGI-associated promoter TSSs, we extracted 600 bp regions downstream of each TSS and performed comparative *de novo* motif searches using HOMER. Here, we analysed one set using the other set as background (corrected for region size, matched for GC content and auto-normalized) to calculate motif enrichment only on the given strand. The top motif enriched downstream of nCGI promoters was the 5′-splice site motif. Genomic distributions of the enriched splice site motif, as well as the AATAAA termination signal were generated using HOMER.

RNA-seq samples and library preparation. Prior to preparation of sequencing libraries, rRNA was removed by poly(A)$^+$ selection (CD19+ B-cells, CD8+ T-cells, 500 ng) or rRNA depletion (fetal heart, 1 μg). Poly(A)$^+$ selection was done twice by using Dynabeads Oligo$(dT)_{25}$ (Life Technologies) according to the manufacturer's manual. rRNA depletion was done by using Ribo-Zero rRNA removal kit (Epicentre, Illumina) according to the manual. The treated RNA was dissolved in 20 μl water.

The pretreated RNA was then fragmented by heating at 70 °C for 3.5 min in fragmentation buffer (Ambion), followed by immediate chilling on ice and addition of 1 μl of Stop solution. Fragmented RNA was purified with the RNeasy MinElute kit (Qiagen) following the instructions of the manufacturer except 675 μl of 100% ethanol is used in step two, instead of 500 μl. Purified RNA was dephosphorylated in phosphatase buffer (New England Biolabs) with 5 U of Antarctic phosphatase (New England Biolabs) and 40 U of RNaseOut (Life Technologies) at 37 °C for 30 min followed by 5 min at 65 °C. After chilling on ice RNA was phosphorylated by addition of the following reagents; 5 μl of 10 × PNK buffer, 20 U of T4 polynucleotide kinase (New England Biolabs), 5 μl of 10 mM ATP (Epicentre, Illumina), 40 U of RNaseOut, 17 μl of water. The reaction was incubated at 37 °C for 60 min. Phosphorylated RNA was purified with the RNeasy MinElute kit (Qiagen) as described above. Purified RNA was concentrated to 6 μl by vacuum centrifugation on a SpeedVac (Eppendorf). One μl of 2 μM pre-adenylated 3′ DNA adaptor, 5′-App/ATCTCGTATGCCGTCTTCTGCTTG-3′ was added to the concentrated RNA. After incubation at 70 °C for 2 min followed by chilling on ice for 2 min, the following reagents were added to ligate the adaptor at the 3′ end of the RNA; 1 μl of 10 × T4 RNA ligase 2 truncated buffer, 0.8 μl of 100 mM $MgCl_2$, 20 U of RNaseOUT and 200 U of RNA ligase 2 truncated (New England Biolabs). After the incubation at 20 °C for 60 min, 1 μl of heat-denatured 5 μM 5′ RNA adaptor, 5′-GUUCAGAGUUCUACAGUCCGACGAUCGAAA-3′ was ligated with 3′ adaptor ligation products with 20 U of T4 RNA ligase 1 (New England Biolabs) and 1 μl of 10 mM ATP (New England Biolabs) at 20 °C for 60 min. 4 μl of adaptor ligated RNA was mixed with 1 μl of 20 μM RT Primer, 5′-CAAGCAGAAGACGGCATACGA-3′, followed by incubation at 70 °C for 2 min, and immediately kept on ice. The reverse transcription reaction was done with 2 μl 5 × Prime Script buffer, 1 μl of 10 mM dNTP, 20 U of RNaseOUT and 200 U of PrimeScript Reverse Transcriptase (TakaraBIO) at 44 °C for 30 min. The cDNA product was amplified by PCR with 10 μl of 5 × HF buffer, 1.25 μl of 10 mM each dNTP mix, 2 μl of 10 μM FWD primer, 5′-AATGATACGGCGACCACCGACAGGTTCAGAGTTCTACAGTCCGA-3′, 2 μl of RT primer and 1 U of Phusion High-Fidelity DNA Polymerase (New England Biolabs). PCR was carried out in a total volume of 50 μl with the following thermal program; 98 °C for 30 s, 12 PCR cycles of 10 s at 98 °C, 30 s at 60 °C, and 15 s at 72 °C, followed by at 72 °C for 5 min and then kept at 4 °C. Remaining PCR primers were removed twice by using 1.2 volumes of AMPure XP beads (Beckman Coulter). The resulting libraries were checked for size and concentration by BioAnalyzer (Agilent) using the High-Sensitivity DNA Kit (Agilent). Qualified sequencing libraries were loaded on the HiSeq2000 (Illumina) using the custom sequencing primer, 5′-CGACAGGTTCAGAGTTCTACAGTCCGACGATCGAAA-3′.

All RNA-seq samples profiled in this study were also profiled in the FANTOM5 promoterome manuscript and are described in detail there[6]. Briefly all human samples used in the project were either exempted material (available in public collections or commercially available), or provided under informed consent. All non-exempt material is covered under RIKEN Yokohama Ethics applications (H17-34 and H21-14). For the samples profiled by RNA-seq, the human fetal heart RNA was purchased from Clontech (Catalogue no.636583). CD19+ B-cells and CD8+ T-cells were isolated using the pluriBead system (huCD4/CD8 cascade and huCD19 single; PluriSelect). RNA was then extracted using the miRNeasy kit (Qiagen).

RNA-seq mapping and transcript assembly. Single-end 100 bp long reads from libraries originating from the similar cell sources (all six "CD19+ B cells" libraries, all six "CD8+ T cells" libraries and one "Fetal heart" library) were processed together via the Moirai pipeline (Hasegawa, Y. *et al.*, manuscript in preparation). The processing steps implemented within the Moirai pipeline included (1) raw sequenced reads PolyA tail and "CTGTAGGCACCATCAAT" adaptor clipping using FASTQ/A Clipper from FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/), (2) removal of sequenced reads containing "N" and sequences similar to ribosomal RNA using rRNAdust version 1.02 (T.L. *et al.*, manuscript in preparation), and (3) mapping the resulting reads against the hg19 human genome using TopHat[46] (version 1.4.1) using both TopHat *de novo* junction finding mode and known exon-exon junctions extracted from GENCODE V10, with all the other parameters set to their default values. Mapped reads flagged as PCR duplicates were removed and the remaining TopHat aligned reads were then assembled using Cufflinks[47] (version 1.3.0) with Cufflinks parameters set to their default values.

Assessment of lengths of RNAs emanating from enhancers and promoters. All Cufflinks assembled transcripts, whose 5′ends, regardless of strand, were located within the outer boundaries of CAGE enhancers or, on the same strand, within 200 bp (upstream or downstream) of a GENCODE (v10) protein-coding TSS were considered for further analysis. For these Cufflinks transcripts we calculated their (intron-less) RNA length, (possibly intron-containing) genomic length as the genomic distance between their 5′ and 3′ ends, as well as their number of exons. Exons of Cufflinks transcripts with 5′ ends in enhancers were further checked for at least 50% (reciprocal) overlap with exons of GENCODE (v10) known, level 1, protein-coding genes and lincRNAs. We repeated the same analysis specifically for u-enhancers.

Small RNA library preparation and mapping. Short RNA-seq sequencing libraries were prepared as 24-plex using the TruSeq Small RNA Sample Prep Kit (Illumina) following the manufacturer's manual. All starting sources were 1 μg of total RNA. The prepared sequencing libraries were loaded on a HiSeq2000 (Illumina). All samples profiled in this study were also profiled in the FANTOM5 promoterome paper[6] and are described in detail there. Briefly, all human samples used in the project were either exempted material (available in public collections or commercially available), or provided under informed consent. All non-exempt material is covered under RIKEN Yokohama Ethics applications (H17-34 and H21-14). For the samples profiled by sRNA-seq, the human fetal heart RNA was purchased from Clontech (catalogue no.636583). CD19+ B-cells and CD8+ T-cells were isolated using the pluriBead system (huCD4/CD8 cascade and huCD19 single; PluriSelect). RNA was then extracted using the miRNeasy kit (Qiagen).

Short RNAs were profiled using the Truseq protocol from Illumina, using an 8-plex. The 8-plex was first split by barcode and the resulting FASTQ sequences trimmed of the 3′ adaptor sequence. Sequences with low quality base N were removed. Ribsomal RNA sequences were then removed using the rRNAdust program. Remaining reads were then mapped using BWA version is 0.5.9(r16) and multimappers were randomly assigned.

Analysis of small RNAs at enhancer TSSs and promoter TSSs. 5′ and 3′ ends of mapped sRNAs as well as pooled CAGE 5′ ends were overlaid windows of 601 bp centred on forward strand summits of enhancer-defining CAGE tag clusters and sense strand summits in promoters of RefSeq protein-coding genes. The average cross-correlation between CAGE 5′ ends and sRNA 3′ ends were calculated in these windows allowing a max lag of 300. For footprint plots, reads mapping to the same genomic locations were only counted once not to make any library or genomic region have an undue influence.

HeLa cells culturing and *SKIV2L2* depletion. HeLa cells were grown in DMEM medium supplemented with 10% fetal bovine serum at 37 °C and 5% CO_2. siRNA-mediated knockdown of either EGFP(control), and *SKIV2L2* (MTR4) were performed using 22 nM of siRNA and Lipofectamin2000 (Invitrogen) as transfecting agent. A second hit of 22 nM siRNA was given after 48 h. Cells were collected an additional 48 h after the second hit, and protein depletion was verified by western blotting analysis as described elsewhere[48]. The following siRNA sequences were used:

egfp GACGUAAACGGCCACAAGU[dT][dT]
egfp_as ACUUGUGGCCGUUUACGUC[dT][dT]
SKIV2L2 CAAUUAAGGCUCUGAGUAA[dT][dT]
SKIV2L2 _as UUACUCAGAGCCUUAAUUG[dT][dT]

HeLa CAGE library preparations and data processing. CAGE libraries were prepared from 5 μg of total RNA purified from 2×10^6 HeLa cells using the Purelink mini kit (Ambion) with 1% 2-mercaptoethanol (Sigma) and on-column DNase I treatment (Ambion) as recommended by manufacturer. CAGE libraries were prepared as described previously[49]. Prior to sequencing four libraries with different barcodes were pooled and applied to the same sequencing lane. The libraries were sequenced on a HiSeq2000 instrument (Illumina). To compensate for the low complexity in 5′end of the CAGE libraries 30% Phi-X spike-in were added to each sequencing lane as recommended by Illumina. CAGE reads were assigned to their respective originating sample according to identically matching barcodes. Assigned reads were trimmed to remove linker sequences and subsequently filtered for a minimum sequencing quality of 30 in 50% of the bases using the FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). Mapping to the human genome (hg19) was performed using Bowtie[50] (version 0.12.7), allowing for multiple good alignments and subsequently filtering for uniquely mapping reads. Reads that mapped to unplaced chromosome patches or chrM were discarded.

Assessment of degradation rates of RNAs emanating from CAGE enhancers and promoters. Bidirectionally transcribed loci were identified in the same way as with pooled FANTOM5 CAGE libraries (see 'Identification of bidirectionally transcribed loci' and 'Expression quantification of bidirectionally transcribed loci and prediction of enhancers' above) from tag clusters (as defined in[51]) derived from pooled HeLa CAGE (mock treated control, *SKIV2L2*) libraries. From 5,892 bidirectional loci distant to TSSs and exons, 4,196 were predicted to be enhancers based on balanced directionality of transcription of which 3,896 had significantly greater expression than random genomic regions in at least one library. These were then overlapped with the whole set of FANTOM5 CAGE enhancers to estimate the fraction of unseen transcribed enhancers.

The expression fold change of HeLa depleted of *SKIV2L2* compared to mock-treated control were assessed and compared between expressed HeLa CAGE enhancers, promoters of RefSeq protein-coding genes in general and broken up into CpG and non-CpG promoters, and ubiquitous FANTOM5 CAGE enhancers.

We further calculated the average footprints of H3K4me1, H3K4me2 and H3K27ac from ENCODE (Broad, Bernstein) signal files in 601 bp windows centred on mid points of enhancers identified in HeLa cells and those that were novel in *SKIV2L2* $^-$.

Purification of blood cell types. Peripheral blood mononuclear cells were isolated from leukapheresis products of healthy volunteers by density gradient centrifugation over Ficoll/Hypaque (Biochrom AG). Collection of blood cells from healthy donors was performed in compliance with the Helsinki Declaration. All donors signed an informed consent. The leukapheresis procedure and subsequent purification of peripheral blood cells was approved by the local ethical committee (reference number 92-1782 and 09/066c). CD4+ cells were enriched using magnetically labelled human CD4 MicroBeads (Miltenyi Biotec) and the Midi-MACS system (Miltenyi Biotec). The CD4+ fraction was stained with CD4 FITC (fluorescein isothiocyanate; Becton Dickinson, catalogue no. 345768), CD25 phycoerythrin (PE; Becton Dickinson, catalogue no. 341011) and CD45RA allophycocyanin (APC) and CD3+CD4+CD25− T cells were sorted on a FACS-Aria high-speed cell sorter (BD Biosciences). CD8+ cells were enriched using magnetically labelled human CD8 MicroBeads (Miltenyi). The CD8+ fraction was stained with CD3 FITC (Becton Dickinson, catalogue no 345763) and CD8 APC (Becton Dickinson, catalogue no. 345775) and sorted for CD3+CD8+ T cells. CD19+ and CD56+ cells were enriched from the CD8− fraction using magnetically labelled human CD19 and CD56 MicroBeads (Miltenyi). Enriched cells were stained with CD3 FITC (Becton Dickinson, catalogue no. 345763), CD19 PE (Becton Dickinson, catalogue no. 345777) and CD56 APC (Becton Dickinson, catalogue no. 341027) and sorted into CD3+CD19+ B cells and CD3+CD56+ natural killer (NK) cells. Purification of blood monocytes is described elsewhere[52].

Generation of ChIP data for blood cells. Chromatin was obtained from CD4+CD25− T cells, CD8+ T cells, CD19+ B cells, and CD56+ NK cells of two healthy male donors each. Chromatin immunoprecipitation (ChIP) for H3K4me1 and H3K27ac and library construction were done essentially as described elsewhere[52]. Sequence tags were mapped to the current human reference sequence (GRCh37/hg19) using Bowtie[50] and only uniquely mapped tags were used for downstream analyses. H3K4me1 and H3K27ac ChIP-seq data for CD14+ monocytes was generated elsewhere[52]. Complementary DNase hypersensitivity sequencing data was obtained from the Epigenetics Roadmap project (http://www.roadmapepigenomics.org) and mapped as above. Blood cell ChIP-seq data have been deposited with the NCBI GEO database (accession code GSE40668) and UCSC Genome Browser track hub data of the entire blood cell data set can be found at http://www.ag-rehli.de/NGSdata.htm. Also see Supplementary Table 4.

Clustering of blood cell CAGE and epigenetics data. CAGE samples corresponding to CD4+, CD8+, B cells, NK cells and monocytes were selected in triplicates from among the set of primary cell samples. Based on the total set of 43,011 permissive enhancers, a subset of 6,609 blood-expressed enhancers was defined as being significantly expressed above genomic background (described above) in at least two of the triplicate samples for at least one blood cell type. This subset of enhancers was clustered for heat map visualization using complete linkage agglomerative hierarchical clustering based on enhancer usage per cell type (binary matrix) and Manhattan distance.

Enhancers were defined as being specifically expressed in one blood cell type if having a pairwise $\log_2$ fold change > 1.5 with respect to the other four blood cell types. The fold change was calculated based on the mean expression over triplicate samples per cell type. Footprints for DNase I hypersensitivity (DHS), H3K4me1 and H3K27ac were calculated per cell-type-specific enhancer set and cell type by extension of reads to 200 bp and overlap aggregation for a window of ± 1 kb around enhancer midpoint as the mean TPM signal over all enhancers in that specific subset.

©2014 Macmillan Publishers Limited. All rights reserved

Peak-calling was done using MACS2[53] on pooled data for DHS, H3K4me1 and H3K27ac. Per cell type, peaks were regarded as significant if the peak summit fell within the upper 1 percentile of the background signal (max values in 92,604 random 1 kb non-TSS non-enhancer regions). DHS regions were defined as ± 500 bp around peak summits. Since ChIP-Seq signals for H3K4me1 and H3K27ac often form bimodal peaks around enhancer sites, peak regions were defined as merged regions resulting from overlapping ± 500 bp regions around MACS2 called peak summits.

Transient enhancer-reporter assays in blood cells. Selected blood cell-type-specific enhancer regions (ranging from 800–1,200 bp) were PCR-amplified from human genomic DNA and cloned directly into the CpG-free pCpGL-CMV/EF1 vector[54,55]replacing the CMV enhancer with the DMR regions. Primer sequences are given in Supplementary Table 5. All inserts were verified by sequencing. For transient transfections, plasmids were isolated and purified using the EndoFree Plasmid Kit (Qiagen). Each luciferase construct was transiently transfected into three model cell lines (the monocytic THP-1 cell line, the Jurkat T cell line, and the B cell lymphoma cell line DAUDI). THP-1 and DAUDI cells were transfected using DEAE-dextran with 200 ng reporter plasmid and 10 ng *Renilla* control vector essentially as described[56]. Jurkat cells were transfected as described elsewhere[54]. The transfected cell lines were cultivated for 48 h, collected, and cell lysates were assayed for firefly and *Renilla* luciferase activity using the Dual Luciferase Reporter Assay System (Promega) on a Lumat LB9501 (Berthold). Firefly luciferase activity of individual transfections was normalized against *Renilla* luciferase activity. Transfections correspond to at least three independent experiments measured in duplicates.

To correct enhancer activity for the amount of read-through that is potentially generated from the enhancer TSS, we additionally generated constructs lacking the basal EF1α promoter for all B cell-specific constructs. Relative luciferase activities generated by read-through activity were subtracted from the activity of enhancer/EF1 constructs to reveal 'true' enhancer activities of individual regions. To further determine the position and activity of reporter TSS, 5′ rapid amplification of cDNA ends (RACE)-PCR for the luciferase gene was performed as follows: RNA of transfected DAUDI cells was reverse transcribed using the SMARTer RACE cDNA Amplification Kit (Clontech) according to the manufacturers' instructions. Rapid Amplification of luciferase 5′ cDNA ends (5′ RACE) was performed with the Advantage 2 Polymerase System (Clontech) and a LUC specific primer (5′-CATGGCTTCTGCCAGCCTCACAGACATC-3′) using the recommended touchdown-PCR program. 15 μl of the PCR products were analysed by agarose gel electrophoresis (2.5%). In addition, fragments were cloned using the StrataClone PCR cloning Kit (Agilent) according to the manufacturers' instructions and sequenced (Life Technologies).

Mass spectrometry analysis of bisulphite-converted DNA. For the set of genomic regions that were also used in transient enhancer-reporter assays, PCR primers were designed using the MethPrimer web tool[57] and purchased from Sigma-Aldrich (for sequences see Supplementary Table 7). Sodium bisulphite conversion was performed using the EZ DNA methylation kit (Zymo Research) using 200–1,000 ng of genomic DNA from CD4+CD25− T cells, CD8+ T cells, CD14+ monocytes, CD19+ B cells, and CD56+ NK cells (two donors each) and an alternative conversion protocol. Amplification of target regions was followed by SAP treatment, reverse transcription and subsequent RNA base-specific cleavage (MassCLEAVE) as previously described[58]. Cleavage products were loaded onto silicon chips (SpectroCHIP) and analysed by MALDI-TOF mass spectrometry (MassARRAY Compact MALDI-TOF, Sequenom). Methylation was quantified from mass spectra using the Epityper software (Sequenom), and averaging methylation levels of CpG dinucleotides located in the central DNase hypersensitive (nucleosome-free) region that is flanked by CAGE clusters. The methylation data for individual CpGs are provided in Supplementary Table 8.

Definition of expression facets and differentially expressed 'specific' facets. Cell and UBERON ontology term mappings were extracted from the FANTOM5 sample ontology[6] for primary cell and tissue samples, respectively, using indirect and direct 'is_a' and 'part_of' relationships. Ontology terms were manually selected to construct groups (facets) of samples that were mutually exclusive and to cover as broad histological and functional annotations as possible. 362 primary cell samples and 138 tissue and whole blood samples were grouped into 69 cell type facets and 41 organ/tissue facets, respectively (the groupings of samples into facets are provided in Supplementary Tables 10 and 11). A few samples were ignored because they were difficult to assign to a facet with certainty, which means that the number of samples within facets is slightly lower than the total number of samples.

For each facet, we defined a set of robustly expressed enhancers from the union of significantly expressed enhancers (see calculation of expression significance above) associated with each contained sample.

For motif search (see below), we identified the set of robust enhancers that were significantly deviating between facets using Kruskal–Wallis rank sum tests (Benjamini–Hochberg FDR < 0.05) and performed pair-wise post-hoc tests (Nemenyi–Damico–Wolfe–Dunn (NDWD) test[59,60] using the R coin package[61] to identify enhancers with significant differential expression (Bonferroni single-step adjusted $P < 0.05$) between facets. Cell type facets and tissue/organ facets were analysed separately. Each enhancer was considered differentially expressed in a facet with at least one pair-wise significant differential expression and overall positive standard linear statistics. This procedure means that we, for each robust enhancer, selected the facets, if any, with strong overall differential expression compared to all other facets. It should be noted that differential expression in this sense is not equivalent to facet-specific (exclusive) expression.

Specificity and usage level analysis. For each robust enhancer, we calculated a 'specificity' score across cell type and organ/tissue facets. The specificity score was defined to range between 0 and 1, where 0 means unspecific (ubiquitously expressed across facets) and 1 means specific (exclusively expressed in one facet).

In detail, $\text{specificity}(X) = 1 - (\text{entropy}(X)/\log_2(N))$, where X is a vector of sample-average expression values for an enhancer over all facets (cell types and organs/tissues were analysed separately) and N its cardinality ($|X|$, the number of facets). The same calculations were done for TPM and RLE normalized CAGE-derived expression levels of RefSeq protein-coding gene promoters (TSS ± 500 bp).

To visualize the complexity and specialization of facets according to usage and specificity score of enhancers and genes, we counted the frequency of facet-used enhancers (significantly expressed in at least one contained sample) and gene promoters (≥1 TPM in at least one sample) with a specificity score in any of 20 bins distributed between 0 and 1. The number of robustly expressed enhancers and genes per sample were normalized to enhancers and genes per million mapped tags, using the total number of mapped CAGE tags in each sample, and further log-transformed. The counts per million mapped tags were visualized in box plots split by facet (only facets with more than one contained sample were considered).

Motif analysis on differentially specific enhancer sets. To identify and compare motif signatures characterizing facet-specific enhancers (permissive set) we applied *de novo* motif analyses. Motif enrichment was analysed using HOMER. Enhancer regions (400 bp) were compared to ~50,000 randomly selected genomic fragments of the same region size, as described above. Twenty-five motifs were searched for a range of motif lengths (7–14 bp) resulting in a set of 200 *de novo* motifs per set, which was further filtered to remove redundant motifs. In parallel, we also used HOMER to calculate the enrichment of ChIP-Seq derived motifs. Motif collections including search parameters for all facets are deposited in the web database at http://enhancer.binf.ku.dk. Known transcription factor motifs were used to compare motif enrichment between facets.

Hierarchical clustering of samples. Tissue and primary cell samples mapped to ontology facets were clustered by complete linkage agglomerative hierarchical clustering based on Jensen–Shannon divergence[59]. In detail, expression values for all enhancers in the permissive set were normalized to sum to 1 for each sample and the square root (proper distance metric) of all pair wise Jensen–Shannon divergences between samples was calculated. Manually selected clades of samples were analysed for differential expression in a similar way as was done for facets (see above). In summary, differentially expressed enhancers (robust set) were identified by Kruskal–Wallis rank sum tests (Benjamini–Hochberg FDR < 0.05) and subsequent NDWD post-hoc tests were performed to find all significant pair-wise differences (Bonferroni single-step adjusted $P < 0.05$) between clades.

Hierarchical clustering of enhancers. We used matrices describing each enhancer expression in TPMs for each facet (primary cell facets and tissue facets were clustered independently) and clustered these by complete linkage agglomerative hierarchical clustering using Euclidean distances, as implemented in the gputools R package[62], and ran these in parallel on a GTX960 Nvidia GPU. Due to limited memory in the GPU, we reduced the matrices to enhancers with total expression >2.5 TPM in the primary cell set and >0.6 TPM in the tissue/organ set, resulting in sets of roughly 22,500 enhancers each. To make sure these results were stable, we also explored normalization using fold change versus background in each facet instead of TPM normalization, which resulted in very similar results (data not shown).

We then used the cutree method to select 5 sub-clusters in each tree, starting from the root. Enhancers in each set were then extended ± 300 nucleotides from their midpoints, and CpG islands and observed/expected CpG ratios were calculated. The resulting sub-clusters broke up enhancers into 201 and 247 ubiquitous enhancers (u-enhancers) defined by cell type and tissue facets, respectively, (these sets intersect by 106 enhancers) and non-ubiquitous enhancers. To summarize the features of u-enhancers in terms of expression width and variance, identified in a single plot, we used those enhancers falling into u-enhancer group from the tissue clustering. We then plotted the mean TPM over all tissue facets, as well as the coefficient of variation (expression variance over all tissue facets scaled by mean expression). Then we repeated this for the remaining enhancers (non-u-enhancers).

Zebrafish reporter transgenesis experiments. We selected enhancers for validation based on human–zebrafish conservation (>70% sequence identity over 100 nucleotides, hg19 vs DanRer7) to take into account the large evolutionary separation between the two species, and selected enhancers that were only expressed in a subset of tissues/cells. We did not take epigenetic data (ChIP/DHS etc.) into consideration. We also selected three negative control regions, chosen randomly from the human genome with the following constraints: low conservation with zebrafish and no other enhancer-selective feature, that is, no DNase hypersensitivity, no H3K4me1 or H3K27ac signals and CAGE signal only at noise levels.

Selected human enhancers (CRE1-5) were amplified from human genomic DNA using primers (Supplementary Table 9). PCR products were purified using NucleoSpin Gel and PCR Clean-up Kit (Macherey Nagel) and were digested using appropriate enzymes (listed in Supplementary Table 9). Human enhancers were cloned into EcoRV/SpeI or HindIII/EcoRI sites of pDB896 vector (gift from D. Balciunas) upstream of zebrafish *gata2* promoter[63,64] and *YFP* reporter gene. Plasmid DNA was purified using NucleoBond Xtra Midi Kit (Macherey Nagel) and quality checked by sequencing before injections.

Zebrafish stocks (*Danio rerio*) were kept and used according to Home Office regulations (UK) at the University of Birmingham. For these experiments wild-type fish (AB* strain) were used. Adults were crossed pairwise and eggs were collected 10-15 min after fertilization. Microinjection solutions contained 30 ng μl^{-1} of plasmid DNA, 0,2% of phenol red (Sigma) and 15 ng μl^{-1} of Tol2 mRNA transcribed *in vitro* from pCS2:Tol2 plasmid using mMESSAGE machine SP6 Kit (Ambion). Injections were performed through the chorion and into the cytoplasm of zygotes using an analogue pressure-controlled microinjector (Tritech Research). More than 200 eggs were injected per construct and experiments were replicated at least three times. Embryos were kept according to ref. 65 in E3 Medium containing 50 ng ml^{-1} of gentamicin (Fisher Scientific) and 0.03% phenylthiourea (PTU, Sigma) in an incubator at 28.5 °C.

Injected embryos were screened during the first 5 days post-fertilization using a Nikon SMZ1500 fluorescence stereomicroscope. Specific expression patterns were documented at 48 hpf and levels of expression were quantified by counting the number of embryos showing enhancer-specific expression. To control for overall background activity from the construct (that is, promoter, backbone) an empty pDB896 vector containing *gata2* zebrafish promoter linked to the reporter gene but lacking an enhancer sequence was used. Any tissue-specific enrichment shown by enhancer-containing vectors over the activity shown by the empty control vector was considered enhancer-specific. Additionally, three negative regions were also cloned to check the specificity of the enhancer selection process. These regions were chosen randomly from the human genome to have low conservation with zebrafish and no other enhancer-selective feature, that is, no DNase I hypersensitivity, no H3K4me1 or H3K27ac signals and CAGE signal only at noise levels. In parallel, 5 selected human enhancers were also analysed. See Supplementary Table 9 for a summary of zebrafish validations, including expression patterns, signal strengths and primers.

Analysis of cohesin data. We used MCF7 cell ChIP experiments with antibodies targeting STAG1 and RAD21 proteins, downloaded from the Short Read archive (accession nos ERR011980, ERR011982). These were mapped using Bowtie[50] with standard settings but discarding non-unique hits, and peak-called using MACS[53] with default settings. We then used the intersection between peak sets as proxy binding sites for the cohesin complex.

Linking TSSs and enhancers by expression correlations. We identified all intra-chromosomal enhancer-promoter pairs (470,315 cases, permissive set of enhancers and unique locations of RefSeq protein-coding gene transcript TSSs ± 500 bp) within 500 kb, in which the TSS was expressed >1 TPM in at least one sample, and performed Pearson correlation tests between the expression of such pairs: 64% of enhancers had at least one significant association (Benjamini–Hochberg FDR ≤ 10^{-5}) within that distance. On average, a TSS was associated with 4.9 enhancers and an enhancer with 2.4 TSSs.

Next, we identified which predicted associations were supported by ENCODE ChIA-PET (via RNAPII (MMS-126R)) interaction data[21] from four ENCODE cell lines (HCT-116, HeLa-S3, K562, MCF-7) by requiring an overlap of both enhancer and promoter in both (and different) sites of a ChIA-PET interaction pair. An association was considered supported if it overlapped in this way with any cell line replicate of interactions.

For comparison, the fraction of 1,672,958 published[10] predicted enhancer-promoter associations derived from DNase data supported by ENCODE ChIA-PET interaction data was calculated.

Analysis of genomic clusters of densely positioned enhancers. By pairwise distance calculations between CAGE enhancers, we identified clusters of densely positioned enhancers (mid points separated by <2 kb) in the genome. 815 regions of length ≥ 2 kb containing >2 enhancers were identified. Of these, 198 regions contained enhancers whose average pairwise expression correlation (Pearson's r) were ≥ 0.75. The expression of associated Refseq genes (see 'Linking TSSs and enhancers by expression correlations') as well as their enrichment of gene ontology biological process terms (via the DAVID tool[66]) were compared to that of genes associated with non-clustered enhancers.

Inferring regulatory architectures by multiple linear regression. Multiple linear regression was performed for all 25,144 expressed (max TPM >1) RefSeq TSSs with at least ten FANTOM5 CAGE-defined enhancers within 500 kb. Enhancers were ranked by proximity to the TSS and the expression values across all samples of the ten closest were used as predictor variables in a model with the TSS expression as response variable. The expression data of enhancers and TSSs were centred and rescaled. 2,206 TSS models, considering in total 11,386 enhancers, with $R^2 \geq 0.5$ were considered for further analyses. We also fitted a simple linear regression model using each enhancer as predictor variable on their own, in order to compare the predictive power of a single enhancer to the power of using all ten. We defined a new measure of 'proportional contribution' to the variance explained as the ratio between simple linear regression r^2 and multiple linear regression R^2, for each enhancer among the ten considered for each TSS. This measure yielded highly similar ranking results of enhancers as the R^2 contribution averaged over orderings among regressors[67,68] and R^2 decorrelation decomposition[67,69] (data not shown), implemented in the 'relaimpo' R package[67,69] (lmg and car methods, respectively). We used ranking of enhancers according to proportional contribution and within-model enhancer-enhancer correlations to identify TSSs with different enhancer architectures. Redundant enhancers were identified for TSSs that had enhancers that were, by proportional contribution, ranked second and onwards with at least some proportional contribution (> 0.2) and high correlation (Pearson's $r > 0.7$) with any other of the nine enhancers in the model. Patterning architectures were considered for enhancers in non-redundant models that were, by proportional contribution, ranked second and onwards with at least some proportional contribution (> 0.2) and low correlation (Pearson's $r < 0.3$) with all other of the nine enhancers in the model.

Penalized lasso-based regression was used to reduce the number of enhancers in the models. The optimal models were selected using 100-fold cross validation and the largest value of lambda such that the mean squared error was within one standard error of the minimum, using the R package glmnet[29,37]

SNP analysis. The NIH NHGRI catalogue of published genome-wide association studies[29] (GWAS catalogue, downloaded 7 May 2012) contained 7,899 SNP-disease/trait associations. We extended this set to 190,356 autosomal associations by propagating disease/trait associations to proxy SNPs using the SNAP proxy search tool[70] (http://www.broadinstitute.org/mpg/snap/) based on linkage disequilibrium ($r^2 > 0.8$) between SNPs (within 250 kb) in any of the three populations in the 1000 genomes project pilot[71] data. The 1000 Genome data coordinates were in hg18 coordinates and were mapped to hg19 using the UCSC liftOver tool[72].

For robust enhancers (centre ± 200 bp), promoters (unique locations of RefSeq protein-coding gene transcript TSSs ± 200 bp), exons (unique locations of RefSeq protein-coding gene transcript inner exons), and random regions (described above), we calculated the number of overlapping and non-overlapping GWAS SNPs associated with each disease/trait in the extended GWAS catalogue. Non-associated SNPs were extracted from the NCBI single nucleotide polymorphism database (dbSNP, build 135). For each genomic feature and disease/trait with an odds ratio > 1, we tested whether the observed overlap was significantly greater than expected (Fisher's exact test $P < 0.01$). Only diseases/traits with more than three SNPs overlapping were tested. The same analysis was repeated for each set of significantly expressed enhancers associated with each facet. For ease of visualization and interpretation, only odds ratios for which the filtering criteria on both significance and overlap number were met are shown. Lists of enhancer-overlapped GWAS SNPs are in S16.

Statistical tests, visualization and tools used. Statistical tests were done in the R environment (http://www.R-project.org). Graphs were made using lattice, ggplot2 and gplots R packages. Cluster trees were generated by the APE[73] R package and visualized using the FigTree software (http://tree.bio.ed.ac.uk/software/figtree/). Intersections of and distances between various genomic features were calculated using BEDTools[74].

39. Kanamori-Katayama, M. *et al.* Unamplified cap analysis of gene expression on a single-molecule sequencer. *Genome Res.* **21,** 1150–1159 (2011).
40. Khalil, A. M. *et al.* Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc. Natl Acad. Sci. USA* **106,** 11667–11672 (2009).
41. 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).
42. Hoffman, M. M. *et al.* Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature Methods* **9,** 473–476 (2012).

43. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods* **9,** 215–216 (2012).
44. Heintzman, N. D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459,** 108–112 (2009).
45. Marshall, O. J. PerlPrimer: cross-platform, graphical primer design for standard, bisulphite and real-time PCR. *Bioinformatics* **20,** 2471–2472 (2004).
46. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25,** 1105–1111 (2009).
47. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnol.* **28,** 511–515 (2010).
48. Preker, P. *et al.* RNA exosome depletion reveals transcription upstream of active human promoters. *Science* **322,** 1851–1854 (2008).
49. Takahashi, H., Lassmann, T., Murata, M. & Carninci, P. 5′ end-centered expression profiling using cap-analysis gene expression and next-generation sequencing. *Nature Protocols* **7,** 542–561 (2012).
50. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).
51. Carninci, P. *et al.* Genome-wide analysis of mammalian promoter architecture and evolution. *Nature Genet.* **38,** 626–635 (2006).
52. Pham, T. H. *et al.* Dynamic epigenetic enhancer signatures reveal key transcription factors associated with monocytic differentiation states. *Blood* **119,** e161–e171 (2012).
53. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9,** R137 (2008).
54. Schmidl, C. *et al.* Lineage-specific DNA methylation in T cells correlates with histone methylation and enhancer activity. *Genome Res.* **19,** 1165–1174 (2009).
55. Klug, M. & Rehli, M. Functional analysis of promoter CpG methylation using a CpG-free luciferase reporter vector. *Epigenetics* **1,** 127–130 (2006).
56. Rehli, M. *et al.* PU.1 and interferon consensus sequence-binding protein regulate the myeloid expression of the human Toll-like receptor 4 gene. *J. Biol. Chem.* **275,** 9773–9781 (2000).
57. Li, L. C. & Dahiya, R. MethPrimer: designing primers for methylation PCRs. *Bioinformatics* **18,** 1427–1431 (2002).
58. Ehrich, M. *et al.* Quantitative high-throughput analysis of DNA methylation patterns by base-specific cleavage and mass spectrometry. *Proc. Natl Acad. Sci. USA* **102,** 15785–15790 (2005).
59. Lin, J. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory* **37,** 145–151 (1991).
60. Hollander, M. & Wolfe, D. A. *Nonparametric Statistical Methods* (Wiley-Interscience, 1999).
61. Hothorn, T., Hornik, K., Van De Wiel, M. A. & Zeileis, A. A Lego system for conditional inference. *Am. Stat.* **60,** 257–263 (2006).
62. Buckner, J. *et al.* The gputools package enables GPU computing in R. *Bioinformatics* **26,** 134–135 (2010).
63. Ellingsen, S. *et al.* Large-scale enhancer detection in the zebrafish genome. *Development* **132,** 3799–3811 (2005).
64. Meng, A., Tang, H., Ong, B. A., Farrell, M. J. & Lin, S. Promoter analysis in living zebrafish embryos identifies a cis-acting motif required for neuronal expression of GATA-2. *Proc. Natl Acad. Sci. USA* **94,** 6267–6272 (1997).
65. Westerfield, M. *The Zebrafish Book. A Guide for the Laboratory Use of Zebrafish (Danio rerio).* (Univ. Oregon Press, 1995).
66. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols* **4,** 44–57 (2008).
67. Zuber, V. & Strimmer, K. High-dimensional regression and variable selection using CAR scores. *Stat. Appl. Genet. Mol. Biol.* **10,** 1–27 (2011).
68. Chevan, A. & Sutherland, M. Hierarchical partitioning. *Am. Stat.* **45,** 90–96 (1991).
69. Groemping, U. Relative importance for linear regression in R: the package relaimpo. *J. Stat. Softw.* **17,** 1–27 (2006).
70. Johnson, A. D. *et al.* SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics* **24,** 2938–2939 (2008).
71. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (2010).
72. Rhead, B. *et al.* The UCSC Genome Browser database: update 2010. *Nucleic Acids Res.* **38,** D613–D619 (2010).
73. Paradis, E., Claude, J. & Strimmer, K. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20,** 289–290 (2004).
74. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).

doi:10.1038/nature13182

A promoter-level mammalian expression atlas

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

Regulated transcription controls the diversity, developmental pathways and spatial organization of the hundreds of cell types that make up a mammal. Using single-molecule cDNA sequencing, we mapped transcription start sites (TSSs) and their usage in human and mouse primary cells, cell lines and tissues to produce a comprehensive overview of mammalian gene expression across the human body. We find that few genes are truly 'housekeeping', whereas many mammalian promoters are composite entities composed of several closely separated TSSs, with independent cell-type-specific expression profiles. TSSs specific to different cell types evolve at different rates, whereas promoters of broadly expressed genes are the most conserved. Promoter-based expression analysis reveals key transcription factors defining cell states and links them to binding-site motifs. The functions of identified novel transcripts can be predicted by coexpression and sample ontology enrichment analyses. The functional annotation of the mammalian genome 5 (FANTOM5) project provides comprehensive expression profiles and functional annotation of mammalian cell-type-specific transcriptomes with wide applications in biomedical research.

The mammalian genome encodes the instructions to specify development from the zygote through gastrulation, implantation and generation of the full set of organs necessary to become an adult, to respond to environmental influences, and eventually to reproduce. Although the genome information is the same in almost all cells of an individual, at least 400 distinct cell types[1] have their own regulatory repertoire of active and inactive genes. Each cell type responds acutely to alterations in its environment with changes in gene expression, and interacts with other cells to generate complex activities such as movement, vision, memory and immune response.

Identities of cell types are determined by transcriptional cascades that start initially in the fertilised egg. In each cell lineage, specific sets of transcription factors are induced or repressed. These factors together provide proximal and distal regulatory inputs that are integrated at transcription start sites (TSSs) to control the transcription of target genes. Most genes have more than one TSS, and the regulatory inputs that determine TSS choice and activity are diverse and complex (reviewed in ref. 2).

Unbiased annotation of the regulation, expression and function of mammalian genes requires systematic sampling of the distinct mammalian cell types and methods that can identify the set of TSSs and transcription factors that regulate their utilization. To this end, the FANTOM5 project has performed cap analysis of gene expression (CAGE)[3] across 975 human and 399 mouse samples, including primary cells, tissues and cancer cell lines, using single-molecule sequencing[4] (Fig. 1; see the full sample list in Supplementary Table 1).

CAGE libraries were sequenced to a median depth of 4 million mapped tags per sample (Supplementary Methods) to produce a unique gene expression profile, focused specifically on promoter utilization. CAGE has advantages over RNA-seq or microarrays for this purpose, because it permits separate analysis of multiple promoters linked to the same gene[13]. Moreover, we show in an accompanying manuscript[4] that the data can be used to locate active enhancers, and to provide numerous insights into cell-type-specific transcriptional regulatory networks (see the FANTOM5 website http://fantom.gsc.riken.jp/5). The data extend and complement the recently published ENCODE[5] data, and microarray-based gene expression atlases[6] to provide a major resource for functional genome annotation and for understanding the transcriptional networks underpinning mammalian cellular differentiation.

The FANTOM5 promoter atlas

Single molecule CAGE profiles were generated across a collection of 573 human primary cell samples (~3 donors for most cell types) and 128 mouse primary cell samples, covering most mammalian cell steady states. This data set is complemented with profiles of 250 different cancer cell lines (all available through public repositories and representing 154 distinct cancer subtypes), 152 human post-mortem tissues and 271 mouse developmental tissue samples (Fig. 1a; see the full sample list in Supplementary Table 1). To facilitate data mining all samples were annotated using structured ontologies (Cell Ontology[7], Uberon[8], Disease Ontology[9]). The results of all analyses are summarized in the FANTOM5 online resource (http://fantom.gsc.riken.jp/5). We also developed two specialized tools for exploration of the data. ZENBU, based on the genome browser concept, allows users to interactively explore the relationship between genomic distribution of CAGE tags and expression profiles[10]. SSTAR, an interconnected semantic tool, allows users to explore the relationships between genes, promoters, samples, transcription factors, transcription factor binding sites and coexpressed sets of promoters. These and other ways to access the data are described in more detail in Supplementary Note 1.

CAGE peak identification and thresholding

To identify CAGE peaks across the genome we developed decomposition-based peak identification (DPI; described in Supplementary Methods; Extended Data Fig. 1). This method first clusters CAGE tags based on proximity. For clusters wider than 49 base pairs (bp) it attempts to decompose the signal into non-overlapping sub-regions with different expression profiles using independent component analysis[11]. Sample- and genome-wide, DPI identified 3,492,729 peaks in human and 2,088,255 peaks in mouse. To minimize the fraction of peaks[3] that map to internal exons (which could exist due to post-transcriptional cleavage and recapping of RNAs[12]), and enrich for TSSs, we applied tag evidence thresholds

*Lists of participants and their affiliations appear at the end of the paper.

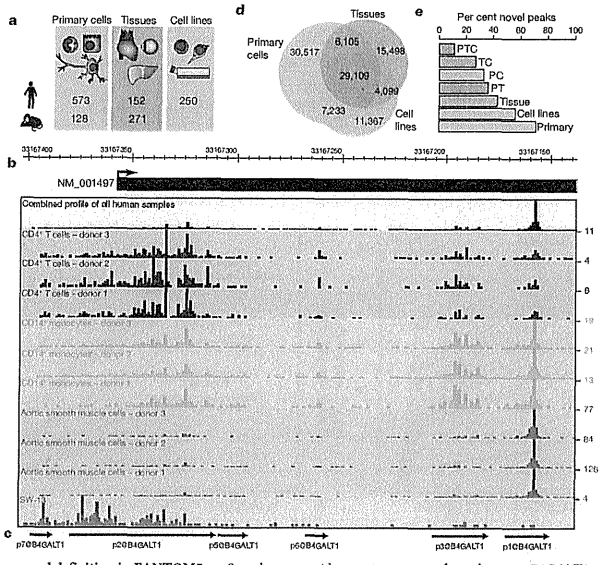


Figure 1 | Promoter discovery and definition in FANTOM5. **a**, Samples profiled in FANTOM5. **b**, Reproducible cell-type-specific CAGE patterns observed for the 266 base CpG island associated *B4GALT1* locus transcription initiation region hg19:chr9:33167138..33167403. CAGE profiles for $CD4^+$ T cells (blue), $CD14^+$ monocytes (gold), aortic smooth muscle cells (green) and the adrenal cortex adenocarcinoma cell line SW-13 (red) are shown. A combined pooled profile showing TSS distribution across the entire human collection is shown in black. Values on the *y* axis correspond to maximum normalized TPM for a single base in each track. **c**, Decomposition-based peak identification (DPI) finds 6 differentially used peaks within this composite transcription initiation region (note: peaks are labelled from p1@B4GALT1 with most tag support through to p7@B4GALT1 with the least tag support; p4@B4GALT1 is not shown and is in the 3′ UTR of the locus at position hg19::chr9:33111241..33111254−). Note in particular one large broad region on the left used in all samples and a sharp peak to the right, preferentially used in the aortic smooth muscle cells. **d**, Venn diagram showing DPI defined peaks expressed at ≥ 10 TPM in primary cells (red), tissues (blue) and cell lines (green). **e**, Fraction of unannotated peaks observed in subsets of **d**. P, primary cells, T, tissues, C, cell lines, PT, TC, PC and PTC correspond to peaks found in multiple sample types, for example, PT, found in primary cells and tissue samples.

to define robust and permissive subsets (described in more detail in Supplementary Methods and summarized in Table 1). Specifically the robust threshold, which is used for most of the analyses presented here, enriched for peaks at known 5′ ends compared to known internal exons by twofold (that is, two-thirds of the peaks hitting known full-length transcript models hit the 5′ end). A flow diagram showing the relationship between samples, peaks, thresholding and subsets used in each analysis is provided in the Supplementary Figure 1. Supporting evidence that the peaks are genuine TSSs, based upon support from expressed sequence tags (ESTs), histone H3 lysine 4 trimethylation (H3K4Me3) marks and DNase hypersensitive sites is provided in Supplementary Note 2.

Figure 1b illustrates the 266 bp spanning transcription initiation region of *B4GALT1*, where 6 independent robust peaks were identified by DPI, each with a unique regulatory pattern (Fig. 1c). A total of 58% of human and 56% of mouse robust peaks occur in such composite transcription initiation regions, defined as clusters of robust peaks within 100 bases of each other. More than half of these contain peaks with statistically significant differences in expression profiles (63% of human and 54% of mouse composite transcription initiation regions; likelihood ratio test, false discovery rate (FDR) <1%, Extended Data Fig. 1d). Supplementary Tables 2 and 3 summarize public domain EST evidence that these independent peaks contained within composite transcription initiation regions give rise to long RNAs.

Known gene coverage in FANTOM5

To provide annotation of the CAGE peaks, the distance between individual peaks and the 5′ ends of known full-length transcripts was determined and then peaks within 500 bases of the 5′ end of known transcript models were assigned to that gene (see Supplementary Methods, Table 1). To provide names for each TSS region, peaks identified at the permissive threshold were ranked by the total number of tags supporting each and then sequentially numbered (for example, p1@GFAP corresponds to the promoter of *GFAP* which has the highest tag support). From these annotations, TSS for 91% of human protein coding genes (as defined by the HUGO Gene Nomenclature Committee) were supported by robust CAGE peaks, and 94% at the permissive threshold (Supplementary Note 3). The atlas also detected signals from the promoters of short RNA primary transcripts, and long non-coding RNAs. In comparison to the previous FANTOM3 and 4 projects, FANTOM5 measured expression at an additional 4,721 human and 5,127 mouse RefSeq genes. The inclusion of primary cells, cell lines and tissues in the atlas provided greater coverage than any of the sample types alone (Fig. 1d) and the primary cell samples in particular were a rich source of unannotated peaks (Fig. 1e).

Mammalian promoter architectures

Mammalian promoters can be classified as broad or sharp types, based upon local spread of TSSs along the genome[13]. The FANTOM5 data

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

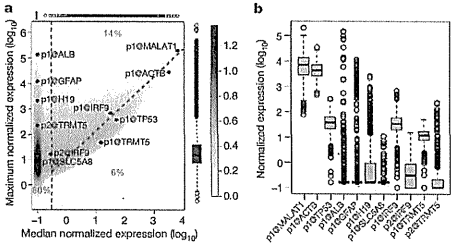


Figure 2 | Cell-type-restricted and housekeeping transcripts encoded in the mammalian genome. **a**, Density plot summarizing the distribution of relative log expression (RLE) normalized maximum and median TPM expression values for the 185K robustly detected human peaks identified by FANTOM5 (colour 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

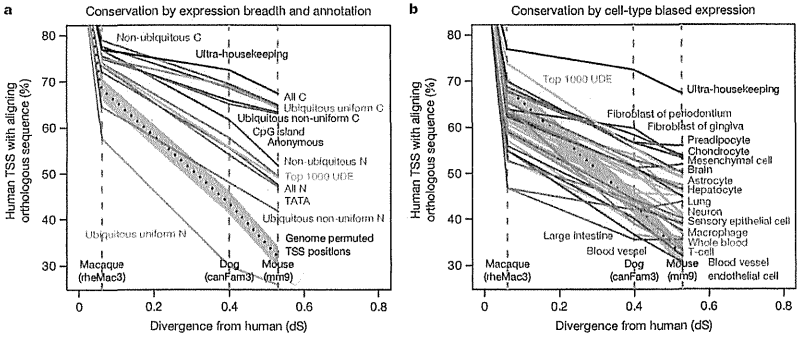


Figure 3 | TSS conservation as a function of expression properties and functional annotation. **a**, **b**, Human robust TSS coordinates were projected through EPO12 whole genome multiple sequence alignments (Supplementary Methods). The y-axis values show the fraction of human TSSs that align to an orthologous 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 = 6.35×10^{-124}, Extended Data Fig. 11h) and bone marrow (UBERON:0002371; P value = 2.7×10^{-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

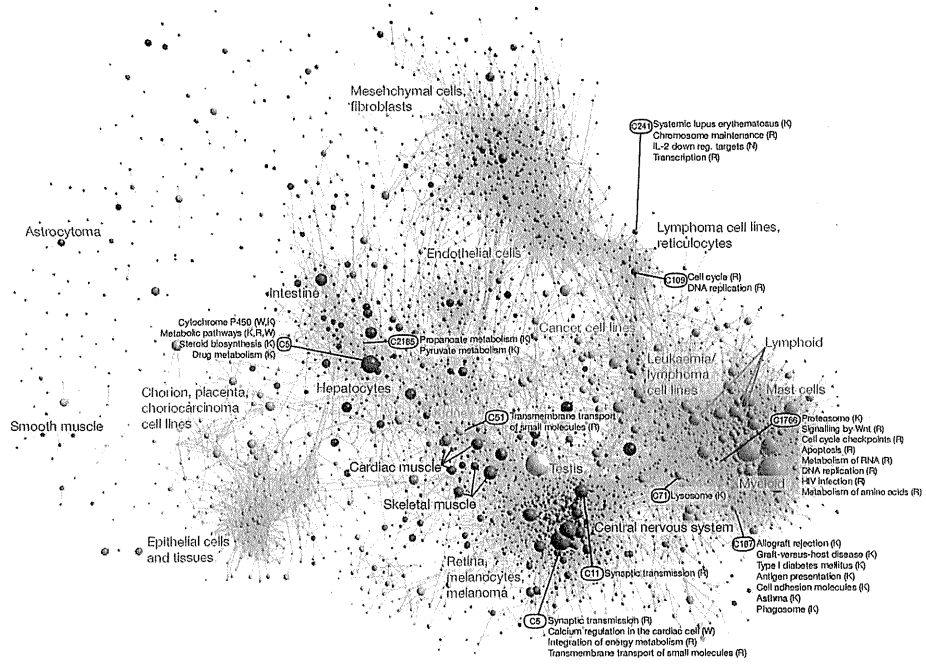


Figure 4 | Coexpression clustering of human promoters in FANTOM5. Collapsed coexpression network derived from 4,882 coexpression groups (one node is one group of promoters; 4,664 groups are shown here) derived from expression profiles of 124,090 promoters across all primary cell types, tissues and cell 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., Love, 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,** e2106 (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.t.H., 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. Clani, 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. (yoshihide@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 Jorgensen[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[29,29], 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[66], 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[73], 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[19], 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[60,85], 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[68]†, Naganari Ohkura[67], Hiroko Ohmiya[1,2], Hiroshi Ohno[51,52], Mitsuhiro Ohshima[69], Mariko Okada-Hatakeyama[51,52], Yasushi Okazaki[60,85], Valerio Orlando[30,37], Dmitry A. Ovchinnikov[29], 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[67], Tetsuro Toyoda[70], Eivind Valen[95], Marc van de Wetering[41], Linda M. van den Berg[63], Roberto Verardo[90], 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, Vavilovstr. 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, Charitéplatz 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, HCIH 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

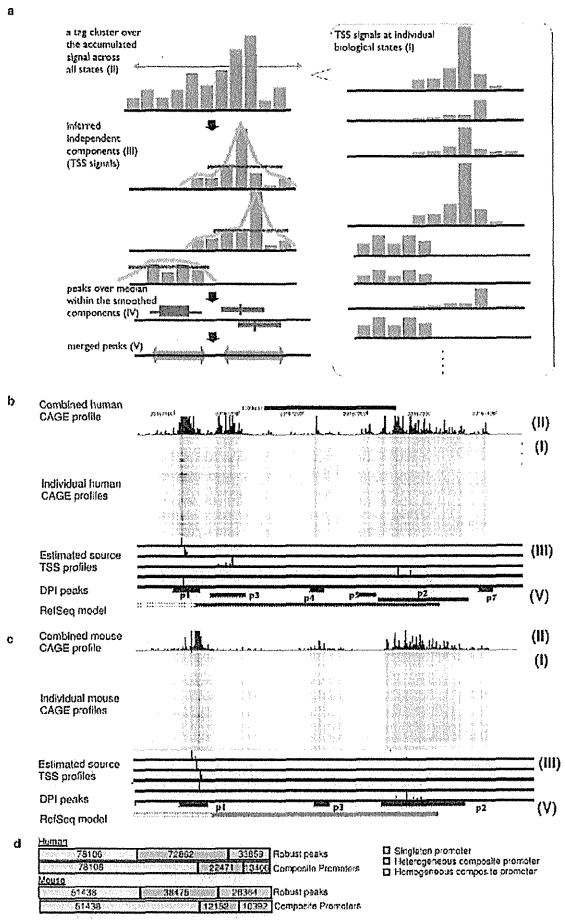
*These authors contributed equally to this work.
†Deceased.

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.).

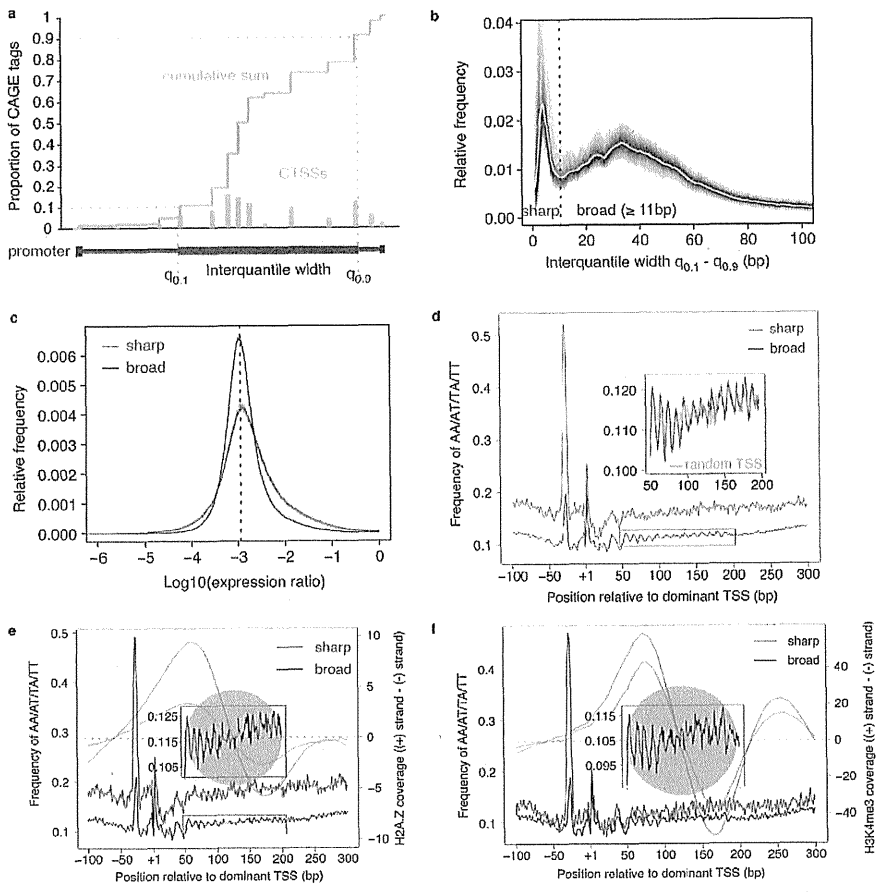
*These authors contributed equally to this work.

51. Pham, T. H. *et al.* Dynamic epigenetic enhancer signatures reveal key transcription factors associated with monocytic differentiation states. *Blood* **119,** e161–e171 (2012).
52. Shulha, H. P. *et al.* Epigenetic signatures of autism; trimethylated H3K4 landscapes in prefrontal neurons. *Arch. Gen. Psychiatry* **69,** 314–324 (2012).
53. Yoneyama, M. *et al.* The RNA helicase RIG-I has an essential function in double-stranded RNA-induced innate antiviral responses. *Nature Immunol.* **5,** 730–737 (2004).
54. Shapira, S. D. *et al.* A physical and regulatory map of host-influenza interactions reveals pathways in H1N1 infection. *Cell* **139,** 1255–1267 (2009).
55. Talukder, A. H. *et al.* Phospholipid scramblase 1 regulates Toll-like receptor 9-mediated type I interferon production in plasmacytoid dendritic cells. *Cell Res.* **22,** 1129–1139 (2012).

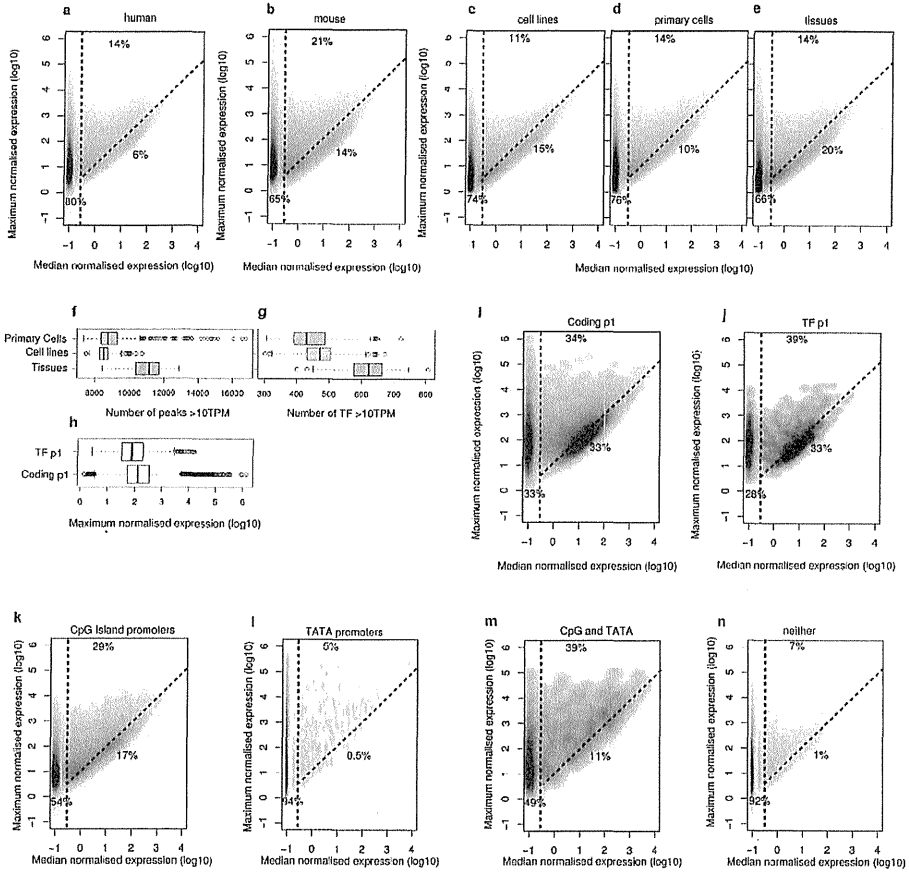
©2014 Macmillan Publishers Limited. All rights reserved



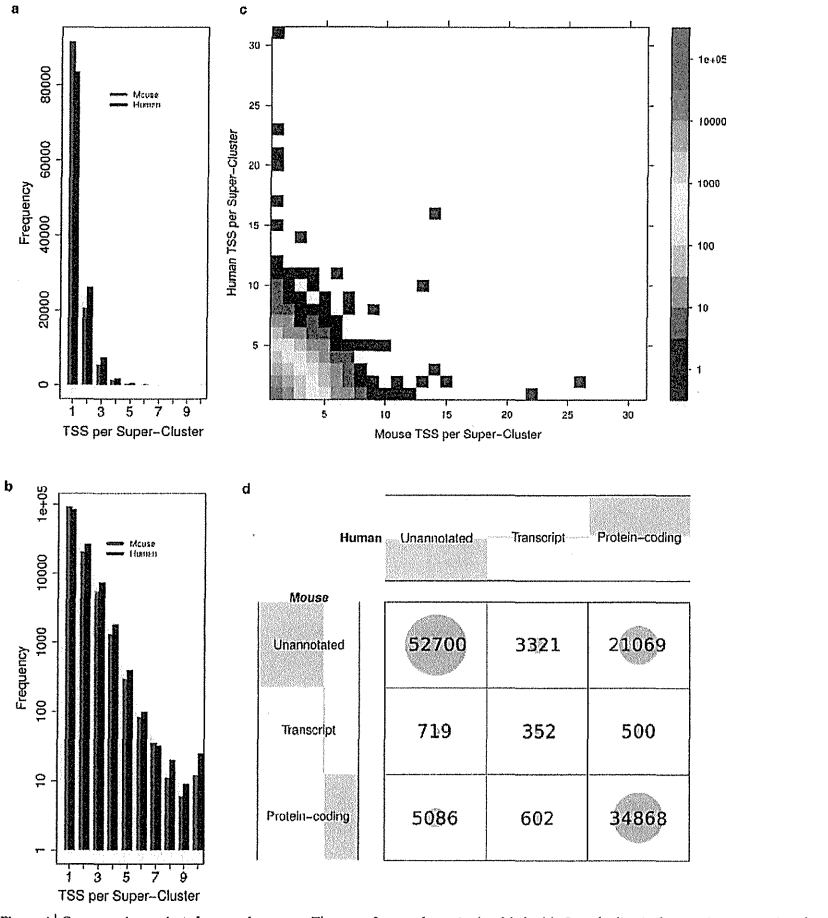
Extended Data Figure 1 | Decomposition-based peak identification (DPI). **a**, Schematic representation of each step in the peak identification. This starts from CAGE profiles at individual biological states (I), subsequently defines tag clusters (consecutive genomic region producing CAGE signals) over the accumulated CAGE profiles across all the states (II). Within each of the tag cluster, it infers up to five underlying signals (independent components) by using ICA independent component analysis (ICA) (III). It smoothens each of the independent components and finds peaks where signal is higher than the median (IV). The peaks along the individual components are finally merged if they are overlapping each other (V). **b**, **c**, Genomic view of actual examples (*B4GALT1* locus) for human and mouse. CAGE profiles across the biological states (I) are shown as a greyscale plot, in which the *x* axis represents the genomic coordinates and individual rows represent individual biological states. Dark (or black) dots indicate frequent observation of transcription initiation (that is, larger number of CAGE read counts) and light dots (white) indicate less frequency. The blue histogram on the top indicates the accumulated CAGE read counts, and the entire region shown represents a single tag cluster (II). The histograms below the greyscale plot indicate the independent components of the CAGE signals inferred by ICA (III), and the resulting CAGE peaks are shown at the blue bars closest to the bottom (V). The bottom track indicates a gene model in RefSeq. The figures overall indicate that only one TSS is defined by RefSeq gene models in this locus, however, transcription starts from slightly different regions depending on the context, and the DPI method successfully captured the different initiation events. **d**, Breakdown of singleton and composite transcription initiation regions with homogenous or heterogeneous expression patterns according to likelihood ratio test (see Supplementary Methods).



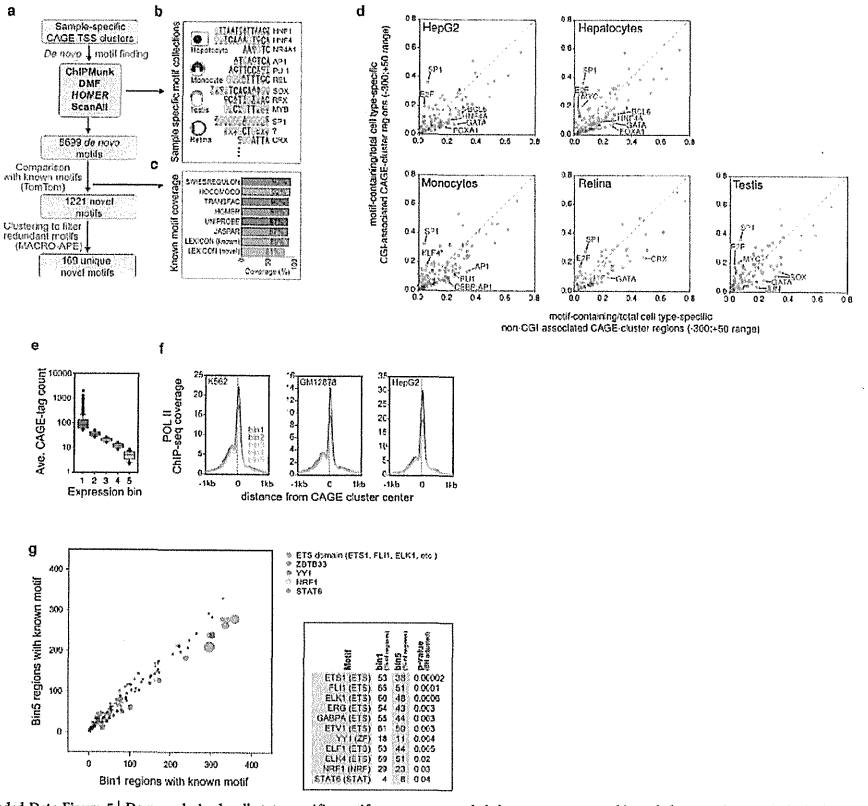
Extended Data Figure 2 | Broad and sharp promoters. DP1 peaks from the permissive set were aggregated by grouping neighbouring peaks less than 100 bp apart. Cumulative distribution of CAGE signal along each region was calculated and positions of 10th and 90th percentiles were determined. **a**, Schematic representation of CAGE signal within promoter region and calculation of interquantile width. Signal from CAGE transcription start sites (CTSS) is shown. Distance between these two positions (interquantile width) was used as a measure of promoter width. **b**, Distribution of promoter interquantile width across all 988 human samples. Individual grey lines show distribution in each sample and the average distribution is shown in yellow. For each sample only promoters with > = 5 TPM were selected. Distribution of obtained interquantile width was clearly bimodal and allowed us to set the empirical threshold at 10.5 bp that separates the best sharp from broad promoters. **c**, Distribution of expression specificity. The distribution of log ratios of expression in individual samples against the median expression across all samples is shown separately for sharp and broad promoters. Solid line shows the average distribution for all samples and the semi-transparent band denotes the 99% confidence interval. The dashed line corresponds to an expected log ratio if all samples contributed equally to the total expression. **d**, Average frequency of AA/AT/TA/TT (WW) dinucleotides around dominant TSS of sharp (red) and broad (blue) promoters across all human samples. Lines show the average signal and semi-transparent bands indicate the 99% confidence interval. Closer view of WW dinucleotide frequency displaying 10 bp periodicity is shown in the inset and indicates the likely position of the +1 nucleosome. For comparison, the signal aligned to randomly chosen TSS in broad promoters is shown in orange. **e**, As in **a** but for promoters in $CD14^+$ monocytes. H2A.Z signal (subtracted coverage = plus strand coverage – minus strand coverage) around sharp and broad promoters is shown in corresponding semi-transparent colours (data from ref. 51). Transition point in subtracted coverage from positive to negative values indicates the most likely position of the nucleosome (shown as semi-transparent blue circle) centre. **f**, As in **b** but for promoters in frontal lobe. H3K4me3 signal (subtracted coverage = plus strand coverage – minus strand coverage) around sharp and broad promoters is shown in corresponding semi-transparent colours (data from ref. 52).



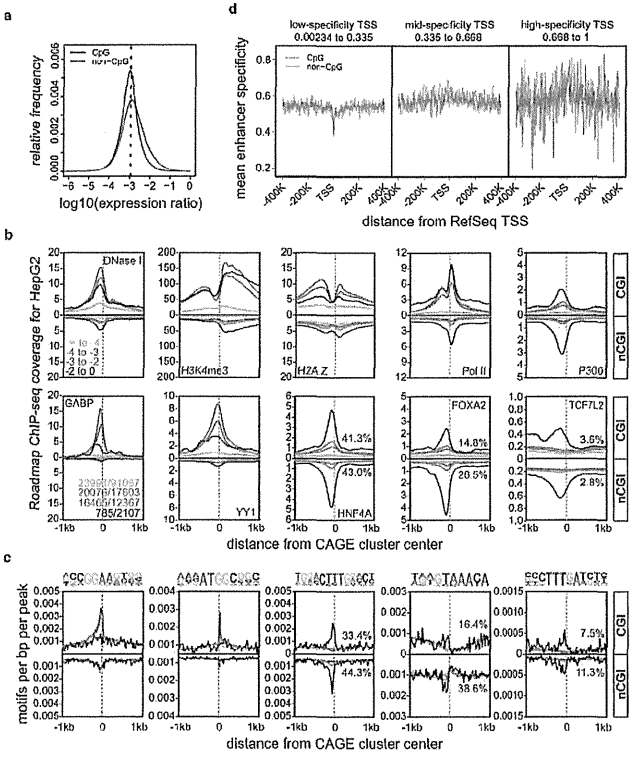
Extended Data Figure 3 | Density plots of DPI peaks maximum and median expression. **a**, Distribution for all human robust peaks. **b**, Distribution for all mouse robust peaks. Fraction on left of vertical dashed line corresponds to peaks with non-ubiquitous (cell-type-restricted) expression patterns (median <0.2 TPM). Fraction below the diagonal dashed line corresponds to ubiquitous-uniform (housekeeping) expression profiles (less than tenfold difference between maximum and median). Fraction in top-middle corresponds to ubiquitous-non-uniform expression profiles (maximum >tenfold median). **c–e** Show distibutions based on cell line, primary cell and tissue data, respectively. The mixture of cells in tissues may overestimate the fraction of ubiquitously expressed genes. **f**, Boxplot showing the number of peaks and detected > = 10 TPM in primary cells, cell lines or tissues. **g**, As in **a** but showing transcription factor p1 peaks only. **h**, Boxplot showing maximum expression of the main promoter for transcription factors or all coding genes. **i**, Density plots of human robust DPI peaks maximum and median expression for the main promoter of coding genes. **j**, As in **d** but showing the main promoter of transcription factors. Fraction on the left of the vertical dashed line corresponds to peaks with non-ubiquitous (cell-type-restricted) expression patterns (median <0.2 TPM). Fraction below the diagonal dashed line corresponds to ubiquitous-uniform (housekeeping) expression profiles (less than tenfold difference between max and median). Fraction above diagonal and to right of vertical dashed lines corresponds to ubiquitous-non-uniform expression profiles (maximum > tenfold median). **k**, Distribution for peaks with CpG island only (n = 55,897). **l**, Distribution for peaks with only a TATA motif (n = 3,933). **m**, Distribution for peaks with both CpG islands and TATA box motifs (n = 834). **n**, Distribution for DPI peaks with neither a TATA motif nor CpG island (n = 124,152). Fraction on the left of the vertical dashed line corresponds to peaks with non-ubiquitous (cell-type-restricted) expression patterns (median <0.2 TPM). Fraction below the diagonal dashed line corresponds to ubiquitous-uniform (housekeeping) expression profiles (less than tenfold difference between max and median). Fraction above diagonal and to right of vertical dashed lines corresponds to ubiquitous-non-uniform expression profiles (maximum > tenfold median).



Extended Data Figure 4 | Cross-species projected super-clusters. **a**, The number of mouse and human TSSs (both permissive and robust) per projected super-cluster. **b**, Same data as presented in panel **a**, with the y axis on a log scale. There is a slight tendency for more human TSSs per super-cluster than mouse TSSs. **c**, The number of human and mouse TSSs per projected super-cluster, density of data points indicated by log-scaled colour gradient shown on the right. Most super-clusters contain < = 4 DPI defined TSSs in both species. **d**, Evaluating the conservation of TSS annotation between species. Projected super-clusters are annotated by the most functional contributing TSS from each species (see Methods). Grey shading in the margins summarizes the proportion of super-clusters with each category of annotation in both mouse (y axis) and human (x axis). Numbers and volumes of circles represent counts of projected super-clusters, for example there are 34,868 super-clusters in which > = 1 human and > = 1 mouse component TSS are annotated as protein coding and 719 super-clusters in which the human TSSs are unannotated and at least one of the mouse TSSs are annotated as the 5′ end of a non-coding transcript.



Extended Data Figure 5 | *De novo* derived, cell-state-specific motif signatures. **a–c**, The *de novo* motif discovery tools DMF, HOMER, ChIPMunk and ScanAll were applied to detect sequence motifs enriched in the vicinity of sample-specific peaks (**a**), yielding 8,699 *de novo* motifs (**b**). The coverage of known motif space by the *de novo* motifs was evaluated by comparing them to the SWISSREGULON, HOCOMOCO, TRANSFAC, HOMER, JASPAR, and ENCODE LEXICON motif collections. **c**, The remaining 1,221 *de novo* motifs that were not similar to known motifs were then clustered using MACRO-APE, resulting in 169 unique novel motifs. **d**, Known motifs from the HOMER database were annotated and counted in around cell-type-specific TSSs (−300 to +50 bp) associated with CpG islands (CGI) or non-CGI regions. **e–g**, RNA Pol II ChIP-seq signal and motif finding in 'housekeeping gene' promoters with different absolute expression levels. Human housekeeping gene promoters were defined as ($\log_{10}(\max + 0.1) - \log_{10}(\text{median} + 0.1) < = 1$). The resulting clusters were then extended by −300 and +50. Overlapping extended clusters were removed by only keeping those with the highest expression. **e**, Extended clusters were then split into 5 equal sized bins with decreasing absolute expression. **f**, RNA Pol II occupancy at binned clusters in ENCODE cell lines (highly expressed genes show the highest occupancy, but even bin5 clusters showing very low tag counts are still highly occupied). **g**, Bubble plot representation comparing known motif enrichments in bin1 (high expression) and bin5 (low expression) extended CAGE clusters. The bubble plots encode two quantitative parameters per motif: difference in motif occurrence between bin1 (*x* axis) and bin5 (*y* axis) as well as the adjusted *P* values for enrichment (bubble diameter). Colouring indicates significantly differentially distributed motifs (5% FDR). The right panel additionally summarizes the fraction of clusters in each bin that contain the indicated motifs along with the Benjamini Hochberg adjusted hypergeometric *P* value for differential enrichment.



Extended Data Figure 6 | Features of cell-type-specific promoters. **a**, The distribution of expression log ratios of all individual samples against the median of all samples is shown separately for CGI-associated and non-CGI-associated CAGE clusters. The dashed line corresponds to an expected log ratio if all samples contribute equally to the total expression. **b**, Histograms for genomic distance distributions of HepG2 DNase I hypersensitivity, H3K4me3, H2A.Z, POL2, P300, GABP, YY1, HNF4A, FOXA1 and FOXA2 ChIP-seq tag counts centred across CGI-associated and non-CGI-associated CAGE clusters (separated according to expression specificities) across a 2 kilobase (kb) genomic region. Expression specificity bins are colour-coded (as indicated in the DNase I panel) with blue representing the highest degree of specificity. Numbers of regions in bins are given in the GABP panel (CGI no. / nCGI no., colour coding as above). **c**, Histograms for genomic distance distributions of ChIP-seq-derived sequence motifs for GABP, YY1, HNF4A, FOXA1 and FOXA2 (corresponding to the samples in the lower panel of **c**) centred across CGI-associated and non-CGI-associated CAGE clusters (separated according to expression specificities) across a 2 kb genomic region. Motifs are shown on top. The percentage of promoters overlapping with ChIP-seq peaks (**b**) or consensus sequences (**c**) for transcription factors binding the highest specificity clusters (HNF4A, FOXA2, TCF7L2) is also given in blue. **d**, Plots showing mean expression specificity (high values indicate more constrained expression over cells, see the accompanying manuscript[4]) in enhancers close to RefSeq promoters as a function of promoter CpG content and three classes of promoter expression specificity.