ENCODE Phase III: Building an Encyclopedia of candidate cis-Regulatory Elements for Human and Mouse

Jill Moore[1*], Michael J. Purcaro[1*], Henry E. Pratt[1*], Charles B. Epstein[2*], Noam Shoresh[2*], Jessika Adrian[3*], Trupti Kawli[3*], Carrie A. Davis[4*], Alexander Dobin[4*], Rajinder Kaul[5*], Jessica Halow[5*], Eric L. Van Nostrand[6*], Peter Freese[7*], David U. Gorkin[8*], Yupeng He[9*], Mark Mackiewicz[10*], Jack Huey, Xiao-Ou Zhang, Diane E Dickel, Xintao Wei, Xiaofeng Wang, Juan Carlos Rivera-Mulia, Joel Rozowsky, Axel Visel, Gene Yeo, Chris Berge, Eric Lecuyer, David Gilbert, Job Dekker, Ali Mortazavi, John Rinn, Eric Mendenhall, Joe Ecker, Manolis Kellis, Robert Klein, William Noble, Job Dekker, Anshul Kundaje, Roderic Guigo, The ENCODE Consortium[#], J. Michael Cherry[11†], Richard M. Myers[10†], Bing Ren[12†], Brenton R. Graveley[13†], John A. Stamatoyannopoulos[5,14†], Mark B. Gerstein[15†], Len A. Pennacchio[16,17,18†], Thomas Gingeras[4†], Michael P. Snyder[3†], Bradley E. Bernstein[19†], Barbara Wold[20†], Ross C. Hardison[21†], Zhiping Weng[1,22†]

[*] Co-first authors
[#] Complete lists of authors and their affiliations appear at the end of the paper.
[†] Co-corresponding authors

SUMMARY

Many human genomes have been sequenced, yet we still lack comprehensive maps of genomic functional elements and do not fully understand how they specify cell and tissue types. Such information is critical for assessing how genomic variants affect development, ageing, and disease susceptibility. The goal of the Encyclopedia of DNA Elements (ENCODE) project is to discover and characterise the full repertoire of functional elements (http://www.encodeproject.org). Here, we summarise the data generated in Phase III of the project and introduce the ENCODE Encyclopedia, an evolving collection of annotations derived from assay-specific and integrative analyses. At the heart of the Encyclopedia is a new Registry of candidate cis-Regulatory Elements (ccREs), defined by a biochemical signature that uses chromatin accessibility, histone modification and transcription factor occupancy data. The Registry currently contains 1.31 M human and 0.43 M mouse ccREs, covering hundreds of biosample types. The ccRE landscape recapitulates the current understanding of cellular identity, tissue composition, developmental progression, and disease-associated genetic variants. Aided by a dedicated visualisation engine called SCREEN (http://screen.encodeproject.org), the Registry is a resource for exploring noncoding DNA elements and their variants.

INTRODUCTION

The genome contains the blueprint for organismal development and function. Deciphering genomes, particularly the vast noncoding regions, is an ongoing challenge that motivates many individual research labs and organised consortium efforts. Among these efforts is the Encyclopedia of DNA Elements (ENCODE) Project, with an overarching goal of providing an integrated resource to aid the scientific community in studying mammalian biology, development, and human diseases{ENCODEProjectConsortium:2007fu, ENCODEProjectConsortium:2012gc, Yue:2014gc}. ENCODE complements the NIH Epigenomics Project{RoadmapEpigenomicsConsortium:2015gq} and takes part in the International Human Epigenome Consortium{Stunnenberg:2016js}. Despite many efforts, the human, mouse, and other mammalian genomes remain incompletely annotated, and our understanding of the diversity of transcripts and their regulatory elements in each cellular context remains limited. To address these limitations, ENCODE Phase III (2012–2017) expanded data collection to include additional chromatin features, regulatory factors and RNA types with an emphasis on primary cells and tissue samples. During Phase III, ENCODE implemented the policy of immediately releasing all data that pass quality control via the freely accessible ENCODE portal (http://www.encodeproject.org).

We have assembled the ENCODE Encyclopedia of predicted and confirmed functional elements in hundreds of cell and tissue types, based on all ENCODE data collected during Phases II and III, supplemented by data from the NIH Epigenomics Project. This paper describes the ENCODE Encyclopedia and presents illustrative examples of its usage. A new focus in ENCODE Phase III has been to build a Registry of candidate cis-Regulatory Elements (ccREs). This effort is guided by the current understanding that robust biochemical signatures, including chromatin accessibility, particular histone modifications, and the binding of certain transcription factors, are preferentially associated with major classes of noncoding regulatory DNA elements—transcriptional promoters, enhancers, insulators, and silencers. While the biochemical signatures are neither causal nor perfect predictors of element activity, they enable the selection of an enriched set of ccREs, together with other genome annotations and their underlying experimental data, for exploration by users via a specifically designed visualisation tool called SCREEN (http://screen.encodeproject.org). The analyses described in this and companion articles have deepened our understanding of the human and mouse genomes, summarized below.

- This paper
- RBP paper: For the 18 RBPs with eCLIP, RBNS and, RBP-knockdown RNA-seq data, we identified 26 variants from the Exome Aggregation Consortium (ExAC){Lek:2016bi} that overlapped an eCLIP peak, disrupted an RBNS motif, and produced a splicing change upon knockdown of the corresponding RBP (van Nostrand et al., in preparation).
- RNA paper: Analysis of long and short strand-specific RNA sequencing and RAMPAGE produced on a set of 53 primary cells from 10 different body locations reveals that many human cells belong to only a few major cell types, corresponding to a relatively small number of transcriptional programs that re-define the classical histological tissue types. These cell types are broadly distributed anatomically, contributing substantially, to the normal cellular composition of many tissues and organs. Through the analysis of histopathological images, we found that significant alterations of the normal cellular composition of tissues correlated with histological phenotypes implicated in diseases, including different cancer types. (Breschi et al., in revision)
- TF paper
- Cancer paper
- ChIA-PET paper
- The 3D nucleome subgroup paper ?

Comment [1]: Brent: this is only a place holder.

SUMMARY OF ENCODE PHASE III DATA PRODUCTION

The ENCODE Consortium has produced data on three main aspects of genome activity—transcriptomes, RNA-based elements for post-transcriptional regulation, and DNA-based elements for regulating transcription and replication. Phase III greatly expanded the number of experiments in each category, having released 5,895 experiments as of 1 Feb 2018 (4,754 on human and 1,141 on mouse; **Fig. 1** and **Extended Data Fig. 1**). With ENCODE Phase II included, a total of 7,385 human experiments are available at the ENCODE Portal as of February 1, 2018 (**Extended Data Fig. 1**). We define an experiment as the application of a genomic assay (such as ChIP-seq, RNA-seq, DNase-seq, or ATAC-seq) to a particular biosample (such as a tissue, a cell line, primary cells, or stem cells) in replicate, summarized in **Table S1a** by category. In this section, we highlight the new assays and results of Phase III data production.

The production of polyA, total and short RNA transcriptome data has focused on primary cells from different body locations and various embryological origins. A new 5´-complete cDNA sequencing assay called RAMPAGE quantifies gene expression, identifies transcription start sites (TSS) at base pair (bp) resolution, and assigns 5´ capped termini to their corresponding RNA isoforms{Batut:2013kc}. This has allowed for richer quantification of tissue-specific TSS usage and identification of novel, tissue-specific TSSs; two examples are shown in **Extended Data Fig. 2**. Single-cell long-RNA-seq was further developed for laser-capture microdissection of human and mouse brain tissues. To better define full-length transcripts, we analysed captured RNAs using long-read sequencing; this effort, in collaboration with the GENCODE project{Harrow:2006ee}, improved the long noncoding RNA (lncRNA) gene and transcript annotation for 14,667 human and 8,708 mouse regions{Lagarde:2017bj}.

Experimental coverage for noncoding, biochemically marked DNA elements, many of which have potential regulatory functions, has been greatly expanded during ENCODE Phase III. We completed 189 new DNase accessibility maps in human, including deep sequencing DNase-seq datasets on hundreds of cell and tissue samples, thus facilitating the prediction of regulatory protein occupancy by footprinting{Hesselberth:2009ci}. The ATAC-seq assay{Buenrostro:2013bc}, which assesses chromatin accessibility via insertion by the Tn5 transposome, was conducted on 48 human tissues. We expanded the application of ChIP-seq to map the locations of modified histones, histone variants, and up to 33 chromatin regulators and modifiers in five human cell lines—K562, H1, GM12878, HepG2, and A549. Additionally, a more limited set of histone modifications were mapped in 21 cell and tissue types. In total, 953 ChIP-seq experiments were completed in Phase III for 410 different transcription factors (TFs), with some TFs assayed in multiple biosamples (1,789 experiments on 532 different TFs in Phases II and III combined). These ChIP-seq experiments used either TF-specific antibodies or epitope-tagged TFs created by BAC transfections or CRISPR/Cas9 genome editing{Savic:2015ea}; detailed antibody information is available at the ENCODE Portal (https://www.encodeproject.org/antibodies/). ChIA-PET of Rad21 and CTCF, which are involved in nuclear organisation, along with Hi-C and ChIA-PET experiments, provide 3D linkage data that include many regulatory regions and cognate target genes {ChIA-PET companion paper, 3D nucleome paper}.

DNA replication timing provides insights into both gene regulation and spatiotemporal genome compartmentalization{RiveraMulia:2016dp}. We measured replication timing during fate commitment of human embryonic stem cells, yielding 84 datasets for 26 cell types representing the embryonic layers endoderm, mesoderm, ectoderm, and neural crest{RiveraMulia:2015er} (**Extended Data Fig. 3a**). Replication timing differs among cell types, and these datasets recapitulate their developmental lineages (**Extended Data Fig. 3b**).

The mouse component of ENCODE Phase III focused on embryonic development at daily intervals between embryonic day 10.5 (e10.5) and postnatal day 0 (P0), with 6-12 tissues sampled per day. Sequencing of polyA-RNA and miRNAs, ChIP-seq for eight histone modifications, ATAC-seq, and whole-genome bisulfite sequencing were performed on all the samples of the mouse embryonic developmental series, augmented by DNase-seq and ChIP-seq of three TFs in selected samples {mouse companion paper}.

A new effort in ENCODE Phase III was to identify and characterise functional RNA elements bound by RNA-binding proteins (RBPs) (Van Nostrand et al., in revision). Four distinct types of experiments were performed: RIP-seq and enhanced UV crosslinking and immunoprecipitation of RBPs followed by sequencing (eCLIP-seq){vanNostrand:2016km} to identify bound RNAs in living cells and quantify the portions of these RNAs involved in binding interactions; RNA-seq on cells depleted of specific RBPs by shRNA or CRISPR; RNA Bind-N-Seq (RBNS){Lambert:2014jm} to determine the relative binding affinity of RBPs in vitro for all possible RNA sequences; and subcellular localization of RBPs by immunostaining. The breadth of our RBP data enables integrative analyses to relate genetic variation{Lek:2016bi} to the regulation of protein isoforms by RBPs. An example is shown in **Extended Data Fig. 4**.

The ENCODE portal (https://www.encodeproject.org) is the primary interface for retrieving all ENCODE data, metadata, experimental protocols, and data standards{Sloan:2015hy}. The Portal was completely redesigned during Phase III for better data access and metadata clarity (**Supplementary Information**). It also provides entry to the ground and integrative levels of the ENCODE Encyclopedia, described in the next section.

THE ENCODE ENCYCLOPEDIA

The raw data described above and their signal maps across the human and mouse genomes can be used for interrogating genome function in many ways, from browsing individual loci to large-scale data integration. To aid users in data mining and hypothesis building, we have derived summaries of key aspects of the raw data and organised them into the ENCODE Encyclopedia. The Encyclopedia presently has two levels of annotations (**Fig. 2**). The ground level includes peaks and quantifications produced by the uniform processing pipelines for individual data types, and the integrative level contains annotations derived from combined analyses across multiple data types and ground-level annotations.

Encyclopedia Ground Level

The ground level currently has nine components (**Fig. 2**). The chromatin accessibility component contains DNase hypersensitive sites (DHSs), genomic regions significantly enriched in DNase-seq reads, as well as ATAC-seq peaks. Locations of histone marks and histone variants are provided in the histone modification component as histone peaks. The transcription factor binding component contains TF peaks, or genomic regions significantly enriched in TF ChIP-seq reads. Peaks of each TF are further characterised by enriched sequence motifs, average histone mark ChIP signals, average nucleosome occupancy, and the ChIP-seq signals of other TFs in the same cell type, all viewable in the wiki-style web resource Factorbook{Wang:2012dk, Wang:2013fp} (http://factorbook.org). The gene expression and TSS activity components provide quantitative estimates of the abundance of the various types of RNA molecules in each of the assayed cell types, at gene level based on ENCODE RNA-seq and at TSS level based on RAMPAGE data, plus activity levels for novel TSSs identified by RAMPAGE. The RNA binding protein component provides RBP peaks, which are regions of the transcriptome enriched for binding by an RBP, as determined by the CLIPper pipeline for eCLIP-seq data{vanNostrand:2016km}. The DNA methylation component provides the methylation state for each cytosine in the genome based on whole-genome bisulfite sequencing data. The 3D chromatin interaction component provides interaction frequency estimates between genomic loci, such as between promoters and distal enhancers, as computed from ChIA-PET data. Finally, the component for chromatin domains and compartments provides topologically associated domains (TADs) and A/B compartments called using Hi-C data, which can be visualized at the 3D Genome Browser (http://3dgenome.org/).

New data are processed as soon as they are available and the resulting annotations added to the ground level of the Encyclopedia. More components will be added as additional analysis pipelines are developed and existing pipelines are improved. We will provide monthly updates to the ground level annotations of Encyclopedia and document their underlying raw data.

Encyclopedia Integrative Level

A longstanding goal of functional genomics is to discover and map the full regulatory element repertoire of the genome and then to delineate which elements are spatiotemporally activated or repressed in individual cell types. In pursuit of this goal, ENCODE and Roadmap Epigenomics Consortia have now broadly surveyed essential epigenetic signals (chromatin accessibility and key histone marks) in hundreds of human and mouse cell types and tissues. ENCODE has also examined a few cell types much more deeply for diverse transcription factor occupancy, genome-wide DNA methylation, and various RNA-binding protein occupancy. The breadth versus depth of assay coverage has motivated two complementary computational approaches to build catalogues of candidate transcriptional regulatory elements, and the integrative level of our Encyclopedia offers both.

The first approach, introduced in ENCODE Phase II, uses machine learning methods such as ChromHMM{Ernst:2010bh, Ernst:2011kw, Ernst:2012ii} and Segway{Hoffman:2012gn} to integrate many different types of epigenetic signals. ChromHMM and Segway are unsupervised probabilistic models that combine a specified number of epigenetic signals to define a dozen or so chromatin states, many of which correlate with known functional element types and activity levels, e.g., active promoters, enhancers, or heterochromatin domains. We applied ChromHMM to the mouse embryo development series—66 complete epigenomes each assayed by ChIP-seq of eight histone marks during ENCODE Phase III—and defined 15 chromatin states that showed coordinated changes with gene expression measured by RNA-seq for each of the 66 samples (Tsuji et al., in preparation). ChromHMM has been augmented to accommodate cell types with some missing assays and then applied to the contemporary Roadmap{Ernst:2015ep} and ENCODE Phase III cell types and tissues that achieved sufficient assay coverage. A strategy was developed in ENCODE Phase III to train separate Segway models on each cell type—allowing for different assay coverages in different cell types—and then interpret these results across all cell types using a Random Forests classifier. The chromatin states of 164 human cell types have been annotated using this strategy by integrating 1,615 genomics datasets, and the resulting annotations have been summarized into a single “functionality score” across the genome that aims to link biochemical activity to functional importance via evolutionary conservation (Libbrecht et al., bioRxiv 086025). The resulting ChromHMM and Segway chromatin state maps are all included in the integrative level of the Encyclopedia.

The second approach is motivated by the substantially increased number of experiments on primary cells and tissues during ENCODE Phase III. The limited quantities of human primary cells and tissues have led to incomplete assay coverage for many of these samples; this precludes the application of approaches such as ChromHMM and Segway, which require relatively dense assay coverage, to these cell types. In order to include these samples in the integrative level of the Encyclopedia, we have developed an approach that uses a highly parsimonious combination of results from just four assays, though at the expense of less detailed inferences about each element’s possible activity. The rest of this paper focuses on this second approach, which has led to the new Registry of candidate cis-Regulatory Elements.

THE REGISTRY OF CANDIDATE CIS-REGULATORY ELEMENTS

Given the breadth of biosamples in the union of ENCODE and Roadmap data, we aspired to build an initial Registry covering a majority of ccREs in the genome. The most direct approach to identifying ccREs would be to include all relevant epigenetic signals in a comprehensive statistical model and then train the model with experimentally validated regulatory elements. Indeed, such methods have been developed{Rajagopal:2013jg, Erwin:2014fo}. However, at this time, relatively few enhancers and insulators have been systematically tested across many cell environments with functional assays; without such a "gold standard," it is not possible to train a general statistical model that remains predictive in new cell types.

Therefore, we pursued a different approach based on four epigenetic signals that are characteristic of major classes of regulatory elements: chromatin accessibility (measured by DNase-seq), the histone modifications H3K4me3 and H3K27ac, and CTCF binding. A ccRE is defined as a region with high chromatin accessibility plus at least one high ChIP-seq signal among H3K4me3, H3K27ac, and CTCF. This definition was initially motivated by substantial prior work in the field{Felsenfeld:2003bw, Allis:2016cp}. DNase hypersensitive sites delineate all of the main classes of cis-regulatory elements in a cell-type-specific manner, including promoters, enhancers, insulators, and repressors{Boyle:2008hm, Thurman:2012fe}. H3K4me3 is highly enriched at active promoters{Barski:2007gh, Heintzman:2007ke, Mikkelsen:2007jg} and H3K27ac marks active enhancers{Creyghton:2010ek, RadaIglesias:2010hy}. CTCF is distinguished from other TFs by being the main insulator-binding protein in vertebrates, with the ability to interfere with enhancer-promoter communication{Kim:2007bm}. Moreover, CTCF functions as an architectural protein of the three-dimensional chromatin structure {Ong:2014ku}, by physically bringing together distant chromatin loci{Rao:2014eo}. Thus, the inclusion of CTCF is well justified for the insulator branch of our selection.

Comment [2]: Active (v. poised) chromatin, including promoters and enhancers; e.g. Ernst and Kellis PMID: 20657582, PMID: 22373907; modENCODE PMID: 21177974; Bernstein PMID: 21441907; Noble (Weng) PMID: 22426492;

We computationally tested the enhancer and promoter branches of our ccRE selection by comparing the effectiveness of ten different types of epigenetic signals in predicting enhancer activity and gene expression: DNase hypersensitivity, eight histone marks (H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K27ac, H3K36me3, H3K9me3, and H3K27me3), and DNA methylation (**Supplementary Methods**). These epigenetic signals and RNA-seq were all assayed with specific mouse e11.5 tissues during ENCODE Phase III; also available were hundreds of enhancers active in four e11.5 mouse tissues—midbrain, hindbrain, neural tube, or limb—by transgenic mouse assays from the VISTA database{Visel:2007jw}. Using VISTA enhancers as the gold standard (**Supplementary Table 2**), we found that DNase and H3K27ac were the best single features for predicting tissue-specific enhancers (**Extended Data Fig. 5a,b**; **Supplementary Table 3**; **Supplementary Information**). We then used RNA-seq to evaluate the effectiveness of these ten epigenetic signals in predicting gene expression levels and found H3K4me3 to the best single feature (**Extended Data Fig. 5c**, **Supplementary Table 4**; **Supplementary Information**).

Comment [3]: Can a simple summary of these results be included in a main figure? This was a strong recommendation from reviewer 1.

Given the predictive power of these four epigenetic features, we developed a simple method that anchors ccREs on a representative set of DHSs (called rDHSs), and then evaluates ccRE types and activities based on H3K4me3, H3K27ac, and CTCF signals (illustrated in **Fig. 3a** for human and detailed in the next subsection). DNase offers high spatial precision in defining ccREs: rDHSs are ~350 bp wide and typically correspond to the cores of regulatory elements. In contrast, the H3K27ac and H3K4me3 signals are more diffuse: they tend to be low at the centre of a regulatory element—presumably because of the lack of a nucleosome —but are elevated at flanking nucleosome positions. CTCF binding also has a high spatial resolution; nevertheless, we used rDHSs as the starting point for defining CTCF-bound ccREs to accommodate the sites of other TFs that may co-bind with CTCF. Therefore, rDHSs specify the localisation of ccREs, while H3K4me3, H3K27ac, and CTCF suggest their activity types. The requirement for significant signal from at least two assay types (DNase plus at least one of H3K4me3, H3K27ac, and CTCF) increases the overall confidence in each ccRE.

We applied our method to all cell types interrogated by at least one of these four assays, covering 301 human cell types (620 when primary cells or tissues from different donors are counted separately) and 58 mouse cell types (or 138 when developmental time-points are counted separately) with all ENCODE and Roadmap data considered. The first release of the Registry presented here includes 1.31 M human ccREs and 0.43 M mouse ccREs; future versions will be released periodically. The smaller number of ccREs in the mouse Registry reflects smaller cell type coverage of our input epigenetic datasets for mouse. The selection, classification, and characterisation of ccREs are detailed in the remaining subsections.

Selection of ccREs for the Registry

We selected ccREs for the Registry as follows. We first condensed all DHSs from 449 DNase-seq experiments (587 individual replicates; **Supplementary Table 5a**) into a set of non-overlapping representative DHSs (rDHSs) as described in **Methods**. We then selected only the rDHSs with a maximal Z-score across all cell types (abbreviated henceforth as max-Z) above 1.64—a threshold corresponding to the 95th percentile of a one-tailed Z-test and used to define a *high* signal throughout this paper. Approximately 1.6 M human and 0.63 M mouse rDHSs met this threshold. We then promoted an rDHS to be a ccRE if it was supported by at least one additional type of epigenetic signal among H3K4me3, H3K27ac, and CTCF, i.e., had a high max-Z across the cell types with data available (**Supplementary Table 5b-e**). Among the primary cells and tissues with DNase-seq data, DNase signal levels at these ccREs recapitulated cell type lineages in human and mouse, and likewise for cell and tissue types with H3K27ac ChIP-seq data (**Supplementary Information**; **Extended Data Fig. 7, 8**), lending support to our ccRE definition.

In total, 1,310,152 human ccREs were in the first version of the Registry (**Fig. 3a**), with median and mean lengths of 343 and 420 bps, respectively, along with 431,202 mouse ccREs (**Extended Data Fig. 6a**), with median and mean lengths of 370 and 466 bps, respectively. A set of 724,590 human ccREs and 228,027 mouse ccREs with high DNase and high H3K4me3, H3K27ac, or CTCF signals in the same cell type, were recognised as having "concordant" support, i.e. the data supporting the prediction were from the same cell type. These concordant ccREs were labelled with a star next to their accessions in SCREEN. The remaining 585,562 human and 203,175 mouse ccREs had high DNase signal in one cell type but high H3K4me3, H3K27ac, or CTCF signals in a different cell type, due to two different reasons: missing data (i.e. not all four core assays had been performed in the biosamples that yielded high Z-scores) or non-concordant data (all four assays have been performed in at least one cell type that yielded a high Z-score, but the ccRE has not received concordant support in any of these fully assayed cell types). **Fig. 3b** shows the breakdown of human ccREs into concordant, missing data, and non-concordant. As more data become available, we will update the ccREs in the concordant class; meanwhile, we distinguish low signal from missing data while annotating ccREs in individual cell types and in the analyses throughout this paper.

Comment [4]: I don't understand what the high max-Z refers to. If the assay were done but the ccRE interval did not have a sufficiently high signal, then we should not call it as a specific type.

Comment [5]: it should have been simply Z-score, and not max-Z. Is this sentence better now?

Classifying ccREs in the Registry

Gene catalogues such as GENCODE define gene models irrespective of their varying expression levels and alternative transcripts across different cell types. By analogy, we provide a general, "cell-type agnostic" classification of ccREs based on the max-Z of each of the four epigenetic features across all cell types with ENCODE or Roadmap data. The goal is to provide a useful overview of the entire ccRE landscape by integrating all input cell types for the four epigenetic features. We then classify ccREs according to these four features at two levels of detail—the state classification and group classification—described below in turn.

As described above (**Fig. 3a**), all ccREs must have a high DNase max-Z (max-Z ≥ 1.64), and a ccRE must have a high max-Z for least one of three epigenetic signals—H3K4me3, H3K27ac, or CTCF. The state classification is simply a delineation of all possible combinations of high or low H3K4me3, H3K27ac, and CTCF signals, with each combination called a "state." This classification captures the fact that while some ccREs are marked by just one high signal among H3K4me3, H3K27ac, and CTCF (41% of human and 59% of mouse ccRE), many ccREs have two or three high signals (**Fig. 3c** for human). Because the all-low state is not allowed, a ccRE can adopt one of seven states for the cell-type-agonistic classification. Each ccRE is further designated TSS-proximal or TSS-distal, defined as within or outside the ± 2 kb window centered on the nearest GENCODE-annotated TSS. There are 242,739 TSS-proximal ccREs in human and 92,405 in mouse. The state classification simply indicates which of the seven states a ccRE is in and is displayed in SCREEN with a colour code alongside the information of TSS proximity (P/D) and whether the ccRE is supported by concordant data from the same cell type (star).

The group classification is an abbreviated abstraction that assigns each ccRE to a group according to its dominant biochemical signature. As reported below for transgenic mouse enhancer assays, the intensity of biochemical signals is positively but modestly predictive of functional enhancer activity. We define three broad, mutually exclusive groups of elements with the expectation that they will be enriched in the respective promoter-like, enhancer-like, or CTCF-mediated functions (**Fig. 3d**; **Extended Data Fig. 6a**):

1. ccREs with promoter-like signatures (*ccRE-PLS*; N=254,880) must have high H3K4me3 max-Zs. If they are TSS-distal, they must also have low H3K27ac max-Zs.
2. ccREs with enhancer-like signatures (*ccRE-ELS*; N=991,173) must have high H3K27ac max-Zs. If they are TSS-proximal, they must also have low H3K4me3 max-Zs.
3. *CTCF-only ccREs* (N=64,099) are the remaining ccREs. They do not fall into either of the first two groups and thus by definition must have high CTCF max-Zs to qualify as ccREs.

Classifications are assigned in the above order; thus, a ccRE possessing high histone mark signals and a high CTCF signal will be classified as either PLS or ELS. In total, 17.6% of the human genome is covered by ccREs (3.6% by ccREs-PLS, 13.4% by ccREs-ELS, and 0.6% by CTCF-only ccREs), and 7.7% of the mouse genome is covered by ccREs (**Fig. 3e, Extended Data Fig. 6b**). The lower coverage for mouse is due to the smaller number of cell types with data to define ccREs.

Comment [6]: Is it 17.6% or 20% of the human genome that is covered by ccREs?

Comment [7]: We updated this figure to simply use the non-blacklisted genome bases. It should be 17.6%. I will make the rebuttal letter consistent.

Previously we used DNase-mappable regions of the genome, but recently we found problems with the DNase-mappable track (this was generated by the Stam lab), so we simply went back to the whole genome.

The state and group classification schemes extend naturally to a specific cell type, but two additional states and their corresponding groups are needed: an *inactive* state/group, containing all ccREs with low DNase Z-scores, and a *DNase-only* state/group, containing ccREs with high DNase Z-scores but low H3K4me3, low H3K27ac, and low CTCF Z-scores within that cell type. Therefore, there are nine possible states and five possible groups in a particular cell type, as shown for human ccREs in GM12878 cells (**Supplementary Information**; **Extended Data Fig. 6c, d**). We used ChIP-seq data of RNA Pol II, EP300, and RAD21 in GM12878 to evaluate the group classification of ccREs in this cell type (**Supplementary Information**; **Extended Data Fig. 9**). The states and groups of human ccREs in three additional cell types (B cells, bipolar neurons, and hepatocytes) and their underlying data are shown in **Extended Data Fig. 10**.

Twenty-one cell types are covered by all four assays, and incomplete assay coverage complicates state and group assignments; nevertheless, we make partial assignments using all available data (**Extended Data Fig. 6e** for human). We analyzed the relative abundances of ccREs-PLS and ccREs-ELS in the 21 human cell types covered by all four assays (**Supplementary Information**; **Supplementary Table 6**).

The simple group classification scheme is designed to give users a first-cut idea of the most likely function for each ccRE, although we are acutely aware that regulatory elements are known to play multiple roles. For example, the *IFITM3* promoter is bound by CTCF, and a single-nucleotide polymorphism (SNP) that interrupts CTCF binding is associated with severe influenza risk in humans{Allen:2017ct}. Many instances of promoters also having enhancer activities and some enhancers also having promoter activities are in the literatures, and recent massively parallel reporter assays reinforce these observations {Nguyen:2016kt}. A tiling-deletion-based CRISPR screen of the 2-Mb *POU5F1* locus identified 45 enhancers that regulate this gene, among which 17 are promoters of functionally unrelated genes{Diao:2017bt}. CapStarr-seq data revealed that 2-3% of the coding-gene promoters display enhancer activity in a given mammalian cell line{Dao:2017el}. Thus, the group classification is intended to simplify analysis and discussion, and we emphasise that some ccREs may belong to multiple groups.

As currently formulated, the Registry does not explicitly define elements that repress gene expression, but we aim to include them in the next version of the Registry. We note that some of the ccREs in the Registry may be repressive in the appropriate cellular contexts. Repression can be achieved through diverse mechanisms: binding a sequence-specific repressor, replacing the binding of a strongly activating transcription factor by a weakly activating one, competing for transcription factors with low abundance, TFs forming complexes with co-repressors rather than co-activators, attracting repressive epigenetic regulators such as Polycomb group proteins, or increasing DNA methylation. Indeed, depending on the cellular context, 25% of *Drosophila* developmental enhancers can also function as Polycomb response elements, silencing transcription in a Polycomb-dependent manner{Erceg:2017kw}. Such findings underscore the notion that ccREs can belong to multiple groups.

We asked whether TSS-distal (> 2 kb) ccREs-ELS still colocalise with genes. Indeed, both human and mouse ccREs-ELS were significantly closer to TSSs than equal-sized genomic regions randomly sampled from the genomes (Wilcoxon rank-sum *p*-values < 2.2E-16 for both genomes). We then asked whether TSS-distal ccREs-ELS were enriched around housekeeping genes or tissue-specific genes. We used the 66 datasets in the mouse embryonic developmental series to answer this question (**Methods**). Indeed, the top 1,000 tissue-specific genes are significantly *enriched* in active ccREs-ELS within 10 kb of their TSSs (*p*-value < 0.005 by random sampling) in 47 out of the 66 tissue–time-point combinations. In contrast, the bottom 1000 tissue-specific genes, i.e., housekeeping genes, are significantly *depleted* of active ccREs-ELS within 10 kb of their TSSs (*p*-value < 0.005) in 50 out of the 66 tissue-time-point combinations (**Supplementary Table 7**). In summary, ccREs-ELS are near genes, especially cell-type-specific genes, that are active in the same cell type.

Evolutionary conservation and repeat content of ccREs

Comment [8]: We need to discuss this.

We analyzed the evolutionary conservation of ccREs in two ways. First, we asked whether ccREs in aggregate are more conserved than randomly selected genomic regions. Indeed, we found all three groups of ccREs are more conserved than random as measured by PhyloP scores in 46 vertebrates{Pollard:2010fj}, with promoter-like elements more conserved than enhancer-like elements, which are in turn more conserved than CTCF-only elements (**Extended Data Fig. 11a**). Additionally, within each group, concordant ccREs are more conserved than the remaining ccREs (**Extended Data Fig. 11a**). Second, we asked whether the degree of evolutionary conservation was correlated with the epigenetic signal, as quantified by DNase max-Z. Following an analysis we performed previously{Kellis:2014gy}, we determined the percentages of nucleotides in ccREs that overlapped with the GERP++ set of evolutionarily conserved regions{Davydov:2010dg}, which cover 8% of the human genome. We observed that ccREs with higher DNase max-Z scores have higher percentages of conserved nucleotides, and the correlation between these two quantities are particularly strong for promoter-like and enhancer-like ccREs (**Extended Data Fig. 11b**). With the exception of the CTCF-only ccREs with the lowest DNase max-Z scores, ccREs have higher percentages of conserved nucleotides than the genome-wide average of 8%.

Among the 0.43 M mouse ccREs, 0.29 M (68%) are orthologous to regions in the human genome, and 0.22 M (77%) of these orthologous regions overlap a human ccRE. Of the 1.3 M human ccREs, 0.69 M (53%) have orthologous mouse regions, but a smaller percentage of these regions overlap mouse ccREs (0.24 M, 36%), which reflects the incompleteness of the mouse Registry.

The entire sets of ccREs is are significantly depleted of transposons and non-transposon repeats in both human and mouse (**Supplementary Table 8**). Particular groups of ccREs are enriched in specific transposon families. The human CTCF-only ccREs are enriched in the long terminal repeat (LTR) class of retrotransposons (1.85 fold; Chi-square p-value = 7.80E-213) and mouse CTCF-only ccREs are enriched in SINE elements (2.05 fold; p-value = 2.02E-171). These results are consistent with early reports on CTCF ChIP-seq peaks{Schmidt:2012dt, Sundaram:2014ku}."

Comprehensiveness of the current Registry of ccREs

Our working hypothesis in defining the Registry of ccREs based on rDHSs is that a collection of rDHSs derived from hundreds of DNase-seq experiments will represent a large fraction of all ccREs in the genome and that a new cell type is likely to use as its ccRE repertoire a subset of the ccREs already in the Registry. To test this hypothesis, we set out to analyse how comprehensive the Registry is in three ways.

First, we examined how many of the GENCODE-annotated TSSs (V19 for human and M4 for mouse) were covered by the Registry, with the coverage defined as the 5´-end of a TSS being contained in a ccRE. GENCODE has released a mature repository of expressed RNAs across all cell types and stages in the human and mouse life cycle, and this test indicates that our Registry includes about two-thirds of the expected promoter-proximal ccREs. More specifically, In human, 67% (121,692/181,177) of all annotated TSSs and 72% (105,196/145,671) of the TSSs of protein-coding genes overlap a ccRE in the Registry; in mouse, 61% (57,459/93,719) of all annotated TSSs and 66% (52,066/78,782) of the TSSs of protein-coding genes overlap a ccRE in the Registry. The ccREs that overlap a GENCODE-annotated TSS are significantly longer than the rest of the TSS-proximal ccREs and TSS-distal ccREs (median length = 548 vs. 317 and 342 for human and 589 vs. 320 and 339 for mouse; Wilcoxon test *p*-values < 2.2E-16 for all four pairwise comparisons). We performed a similar coverage analysis on FANTOM CAT, a collection of human TSSs defined using CAGE peaks{Hon:2017ea}. These TSSs are provided at *stringent* and *robust* levels with false discovery rate (FDR) thresholds of 0.026 and 0.077, respectively. For stringent FANTOM CAT TSSs, 79.8% (83,408/104,479) of them and 79.9% (67,534/84,631) of the protein-coding subset overlap a ccRE. At the robust level, 64.0% (148,284/231,885) of all FANTOM CAT TSSs and 62.5% of the protein-coding subset (106,864/171,195) overlap a ccRE. Note that the ccRE coverage of GENCODE v19 TSSs falls between the coverages of stringent and robust FANTOM CAT TSSs.

Second, we analysed how rapidly the total number of unique human rDHSs saturated as more and more cell types were added. In ENCODE Phase II, we modelled DHS saturation using a Weibull distribution and estimated that we had discovered around half of the total DHSs. We performed this analysis again using all human DNase-seq data generated by ENCODE and Roadmap projects until Feb. 1, 2017. The saturation curves of rDHSs continue to follow Weibull distributions, plateauing at 1.76 M rDHSs with FDR < 0.1% and Z-score > 1.64 (**Fig. 4a**). Because only a subset of such rDHSs can be ccREs—those with a high H3K4me3, H3K27ac, or CTCF Z-score in at least one cell type—this analysis suggests that we have identified roughly three-quarters of human ccREs (1.31 M ccREs among 1.76 M rDHSs is 74.4%).

Comment [9]: We found 1.6M rDHS. How did 1.6M/1.76M end up being 3/4? Why not just compare expected vs observed rDHSs?

Comment [10]: 1.31/1.76 = 74%. I added the red text. Does it help?

Third, we found that the Registry's coverage of H3K27ac, H3K4me3, and CTCF peaks (FDR < 0.01) was very high (often over 90%) in those cell types with the corresponding ChIP-seq data but without DNase-seq data. The Registry covered 90 ± 8% of H3K4me3 peaks (74 cell types), 87 ± 5% of H3K27ac peaks (54 cell types), and 99 ± 1% of CTCF peaks (31 cell types) (**Extended Data Fig. 12a-c**). The coverage was equally high for mouse, despite a smaller number of DNase-seq experiments for building the mouse Registry: 88 ± 5% of H3K27ac peaks (69 tissue–time-points) and 96 ± 8% of H3K4me3 peaks (74 tissue–time-points) were accounted for (**Extended Data Fig. 12d, e**). (No mouse cell or tissue types had CTCF but no DNase data.) The coverages for H3K4me3 peaks were low for several human and mouse cell types. In these cell types, the average –log(FDR) of the H3K4me3 peaks were low (**Extended Data Fig. 12f, g**). We visually inspected the two datasets with the lowest coverage (CD-1 megakaryocyte and GR1-ER4 in mouse) and confirmed that the peaks that were not covered by the Registry had low signals and were likely false positive peaks by the peak calling algorithm.

Comment [11]: I pointed out issues with this analysis in the previous version. There are no DNase-seq data for the megakaryocytes, and I think you did not include the DNase-seq data for G1E-ER4 because it was using the Crawford protocol. If I'm correct that DNase-seq data were not available or used for these cell types, then that could also contribute to the lack of overlap between the H3K4me3 data and the ccRE sets (built on DHSs). We have not found the H3K4me3 data to be dominated by false positives.

Comment [12]: We did not use the DHSs from Crawford data to call rDHSs, but we used all of the Crawford datasets to compute Z-scores. The reason is that although peak calling does not work for these datasets, their signal files are fine to use.

Note that we are not saying your H3K4me3 data are dominated by false positives; we are saying the ChIP-seq peaks of this dataset has positive peaks.

In conclusion, the human Registry appears to be comprehensive: by the above criteria, it covers two-thirds of all ccREs and over 85% of elements marked by H3K4me3 or H3K27ac or bound by CTCF (FDR < 0.01) in any cell type. A cautionary note is that we do not yet know the extent of our coverage of highly cell-type-specific ccREs that are active in rare cell types (numerically minor in their tissues of origin) that have not yet been sensitively assayed. The mouse Registry is less comprehensive than the human Registry, but we expect that it will continue to grow with experiments performed on additional cell types.

Comment [13]: It is not obvious how that number was determined.

Comment [14]: This comes from the red sentence in the previous page. We have a bunch of numbers, how should we give the best estimate?

Comment [15]: I know mouse has fewer ccREs, but the results in this section indicate that the coverage of elements deduced from histone mods and CTCF are as high as observed for human. Why do we think it is less comprehensive?

Comment [16]: A limited set of mouse samples were used for generating all these DNase, DNase, and CTCF data. The fact that these data are self-consistent does not contract the notion that the mouse Registry is incomplete.

Contribution of ENCODE Phase III data to the Registry of ccREs

One way of assessing the impact of ENCODE Phase III data is to simulate the growth of the Registry since the end of Phase II. We first mapped on to the human rDHS saturation curve (**Fig. 4a**) the rDHSs that would have resulted from ENCODE Phase II data only (0.75 M), from the Roadmap Epigenomics data only (0.78 M), or from these two sets of data combined (0.98 M). It is evident that ENCODE Phase III has substantially boosted the repertoire of human rDHSs (1.60 M with all ENCODE and Roadmap datasets). A corresponding growth is seen for human ccREs (**Fig. 4b**): ENCODE Phase III data increased ccREs-PLS from 0.14 to 0.25 M, ccREs-ELS from 0.51 to 0.99 M, and CTCF-ccREs from 21 to 64 k.

As mentioned above, ENCODE Phase III shifted its data production focus from cell lines to primary cells and tissues. We defined human ccREs using just cell line data, just primary cell data, or just tissue data (**Fig. 4c-e**). Tissue data substantially augmented the counts of ccREs-PLS and ccREs-ELS, especially the latter, and primary cell data made further contributions, albeit to a lesser extent than tissue data. CTCF-only ccREs did not increase substantially with the addition of tissue and primary cell datasets, partially because there is a modest number of CTCF ChIP-seq datasets in tissues and primary cells, and partially because CTCF tends to bind overlapping sets of sites between different cell types.

In summary, inclusion of ENCODE Phase III data has roughly doubled the Registry of ccREs. Furthermore, with the additional biosample types, especially tissue and primary cells, we have gained knowledge about which combinations of ccREs are chosen to carry these epigenetic marks in each cell type, and such knowledge will aid the investigations of transcriptional regulation in specific cell types.

Comment [17]: This is not clear to me.

Comment [18]: The Registry is like the dictionary while each biosample is like an assay that uses a specific combination of the words in the dictionary. Do you agree with this analogy? How can I express it more clearly?

Estimating the false discovery rate of ccREs using massively parallel reporter assay data

We wanted to assess the false discovery rate (FDR) of ccREs. This would require functional data in specific cell types by high-throughput assays such as massively parallel reporter assays (MPRA), STARR-seq, or CRISPR{Shlyueva:2014ey}. ENCODE Phase IV has a new set of Functional Characterization Centers that will produce such validation data. Meanwhile, we used MPRA data in lymphoblastoid cell lines (LCLs){Tewhey:2016} to provide an FDR estimate. The authors tested 20,065 regions surrounding 32,373 variants from 3,642 cis-QTLs and control regions, and 3,432 variants (2,098 regions) were active with either the reference or the alternative allele. We overlapped the Registry of human ccREs with these regions and observed that MPRA-positive regions overlapped significantly more ccRE than MPRA-negative regions: 60.8% vs 38.5% (Chi-square test *p*-value = 1.14E-79; **Extended Data Fig. 13a**). We then analyzed the ccREs defined for each of the 462 biosamples that had DNase-seq data. Two of the three LCLs among the 462 biosamples, GM12878 and GM12864, intersected the highest percentages of MPRA-positive regions. Additionally, significantly higher percentages of MPRA-positive regions overlapped ccREs active in other immune cell types than ccREs active in non-immune cell types (Wilcoxon rank sum test *p*-value = 1.3E-17; **Extended Data Fig. 13a**).

We then ranked ccREs in each biosample by their DNase Z-scores and asked how well they could predict MPRA-active regions, plotting precision (fraction of ccREs that are MPRA-active) against recall (fraction of MPRA-active regions that overlap ccREs) in **Extended Data Fig. 13b**. Again, LCL ccREs were the most predictive: precision ranged from 0.8 to 0.65 for the top 1,000 GM12878 ccREs and remained above 0.50 for the top 6,000 GM12878 ccREs, with an overall precision of 0.31 across the full set of 91 k GM12878 ccREs. Some of the ccREs that MPRA deemed inactive are false-positive ccREs, while others could be due to inaccuracy of the MPRA, one commonly known limitation being that the regions tested are only 180-bp long and may not capture the entirety of enhancers. Cell types distant from lymphoblastoid cell lines, particularly non-immune cell types, had lower precisions, with the lowest being trophoblast. Cell-type-agnostic ccREs performed in between GM12878 and trophoblast (**Extended Data Fig. 13b**). The entire set of GM12878 (N=91,961) achieved a recall of 0.32. The low recall might be due to a number of reasons, the first being false-negative ccREs. Alternatively, as the MPRA regions were cloned into plasmids and were not tested in their endogenous chromosomal state, the endogenous loci corresponding to some of the MPRA-positive regions may not be active.

We repeated the analysis for ccREs-PLS and ccREs-ELS separately, computing the enrichment of overlapping MPRA active vs. MPRA inactive regions. Higher enrichments were observed for ccREs-PLS than for ccREs-ELS, but cell-type-specific enrichment was only overserved for ccREs-ELS (**Extended Data Fig. 13c**). LCL ccREs-ELSs showed stronger enrichments than other immune cell types, which in turn showed stronger enrichments than non-immune cell types (mean enrichments 2.16, 1.65, and 1.42 respectively, *p*-values = 6.1e-3 and 1.2e-14 for the two pairwise comparisons). **Extended Data Fig. 13d** shows all 462 biosamples, sorted by the enrichment of ccREs-ELS, illustrating that LCLs and other immune cell types were enriched in the top ranks (Kolmogorov–Smirnov test *p*-value = 2.2e-12).

Overall, based on these MPRA results we estimate that the top ccREs active per cell type validate at high rate, with the top 1,000 validating at approximately 65% and the top 6,000 validating at approximately 50%. Importantly, these results suggest that validation rate is cell type specific, as immune ccREs were more likely to validate in this MPRA study on immune cell types than non-immune ccREs. Additionally, the results lend further support to our observation that the enhancer landscape is far more specific to particular cell types and lineages than the promoter landscapes, given that this enrichment for immune validation was primarily observed for ccREs-ELS.

Experimental testing of 151 ccREs-ELS by transgenic mouse assays

To experimentally test the enhancer branch of our selection, we used the average rank of the DNase and H3K27ac signals to identify previously untested, TSS-distal (> 2 kb from the nearest GENCODE-annotated TSS), ccREs-ELS in the mouse e11.5 midbrain, hindbrain, and limb. The tested regions were centered on ccREs-ELS and their boundaries defined using the overlapping H3K27ac ChIP-seq peaks (**Methods**). For each tissue, we tested 20, 15 and 15 new regions around the ranks of 1-20, 1500-1520, and 3000-3020, respectively. In total, we tested 151 regions, and the results are in **Supplementary Table 9** and summarized as three precision-recall curves (**Fig. 5a)**. Representative e11.5 transgenic mouse embryos for the enhancers that validated in the expected tissues are shown in **Extended data Fig. 14**. Consistently, higher ranking regions were more likely than lower ranking regions to show enhancer activity in their predicted tissue (**Fig. 5b**; e.g., 60%, 40%, and 27% for the midbrain). When enhancers were active in multiple tissues, these tissues also had high H3K27ac signals across the predicted enhancer regions (**Fig. 5c-e**). For example, a predicted enhancer in the midbrain was also active in the forebrain, hindbrain, neural tube, and eye; accordingly, high H3K27ac signals were observed in the first three tissues (**Fig. 5c**; eye H3K27ac data were not available). A predicted enhancer in hindbrain is also active in midbrain and neural tube, consistent with its H3K27ac signal in these two tissues, but it is not active in forebrain despite a low level of H3K27ac signal in that tissue (**Fig. 5d**). In contrast, an enhancer active almost exclusively in the limb (**Fig. 5e**) did not show high H3K27ac signals in other tissues assayed. These results suggest that combining DNase and H3K27ac can identify active enhancers in a particular tissue and quantify their tissue selectivity patterns.

The overall validation rates for these predictions (43-46% across the three tissues) are lower than those of one earlier study{Visel:2009jp} (78-82% in forebrain, limb, and midbrain) but higher than those of two later studies—32% in forebrain{Visel:2013it} and 38% in heart{Dickel:2016jg}. The higher validation rate by Visel et al. may be partly due to the requirement of evolutionary conservation {Visel:2009jp}, which was not imposed on our enhancer predictions. The enhancer predictions that yield negative results in mouse transgenic assays can be due to a number of reasons: false positive predictions, fragmented enhancers, enhancers active at other time points, or low-activity enhancers below the detection limit of the assay. Thus, it is important to experimentally test the predicted functions of large numbers of ccREs. Indeed, during ENCODE Phase IV, a group of Functional Characterization Centers will produce validation data using massively parallel reporter assays, STARR-seq, and various flavors of CRISPR assays{Shlyueva:2014ey}.

Comparison of ccREs with ChromHMM states

As described above, there are two approaches to building catalogues of regulatory elements, with the Registry of ccREs representing one and automated machine learning methods such as ChromHMM and Segway representing the other. We asked how the simple, rDHS-anchored, one-additional-support approach of defining ccREs compared with the more sophisticated,

Comment [19]: So if the top 1000 validate 65% or the time, and the top 6000 validate 50%, what do we think the overall rate is for 1,310,152 human ccREs?

Comment [20]: I would be happy if 50% of 1.31 M ccREs validate by this particular assay (MPRA) given MPRA's own limitations.

Comment [21]: surely that is expected.

Comment [22]: This section needs substantial revisions.

Comment [23]: This is an important section, and one emphasized by the reviewers. I agree that we have to make it clearer, and we have to reach conclusions along the lines of those requested by the reviewers.

hidden-Markov-model-based approach of ChromHMM, which also incorporates more histone marks.

The ccREs-PLS and ccREs-ELS in GM12878 are consistent with the respective chromatin states called by ChromHMM using eight histone marks and CTCF in this cell type{ENCODEProjectConsortium:2012gc}. Roughly 90% of top ccREs-PLS (ranked by the H3K4me3 Z-score) overlap with ChromHMM promoters, and over 85% of the top ccREs-ELS (ranked by the H3K27ac Z-score) overlap with ChromHMM high-signal enhancers (**Extended Data Fig. 15a, b**). With decreasing ranks of ccREs-ELS, the overlap with ChromHMM high-signal enhancers decreases while the overlap with ChromHMM low-signal enhancers increases—82% of the ccREs-ELS ranked above 20,000 overlap with either type of ChromHMM enhancers. Reciprocally, the TSS, enhancer, and insulator ChromHMM states occupy much larger genomic space than do ccREs; nevertheless, these states are enriched in the corresponding groups of ccREs (**Supplementary Table 10)**. We performed the comparison in two ways—by regions and by genomic positions—considering that ChromHMM regions are much larger than and can overlap multiple ccREs. The ChromHMM TSS state overlaps many more ccREs-PLS than it does ccREs-ELS and CTCF-only ccREs (45.22% vs. 8.33% and 1.63% by region; 39.35% vs. 3.85% and 0.53% by position; **Supplementary Table 10a, b**). Likewise, the high-signal enhancer state overlaps the most with ccREs-ELS, and the insulator state overlaps the most with CTCF-only ccREs (**Supplementary Table 10a, b**).

We also compared the ccREs for five e11.5 and six e14.5 mouse tissues with the ChromHMM states called using eight histone marks in the corresponding tissues, as described in a companion paper (Tsuji et al., in preparation). We observed that 95 ± 2% of ccREs-PLS overlapped ChromHMM-annotated promoters and 78 ± 3% of ccREs-ELS overlapped ChromHMM-annotated enhancers in the corresponding tissue and time point (**Extended Data Fig. 15c, d**). Reciprocally, the ChromHMM TSS and bivalent TSS states overlap most extensively with ccREs-PLS (e.g., on average 71.23% of TSS regions and 56.72% of bivalent TSS regions at e11.5**)**, while the high-signal enhancer state overlaps most extensively with ccREs-ELS (e.g., on average 35.52% by region and 23.76% by position; **Supplementary Table 10c, d**).

We asked whether a genomic position assigned to a ChromHMM state is more likely to be evolutionarily conserved if it overlaps a ccRE than otherwise. Indeed, for both GM12878 cells and mouse tissues, the positions in a ChromHMM region that overlap ccREs are significantly more conserved than the rest of the positions in the same ChromHMM region (**Supplementary Table 10e)**. The difference is most pronounced for the high-signal enhancer state, showing a 1.45 and 2.88 fold higher conservation with ccRE overlap in GM12878 cells and averaged across mouse tissues (Wilcoxon rank-sum test *p*-value = 1.69E-117 and < 1E-250 respectively, **Supplementary Table 10e**).

In summary, most ccREs-PLS are contained in ChromHMM TSS regions, ccREs-ELS in ChromHMM enhancer regions, and CTCF-only ccREs in ChromHMM insulator regions. Moreover, ccREs occupy much smaller genomic footprints than ChromHMM regions, and the positions of ChromHMM regions that overlap ccREs are more evolutionarily conserved than the remaining positions, suggesting that ccREs can pinpoint the more likely functional positions in the genome.

Comparison of ccREs with FANTOM enhancers

The FANTOM5 Consortium has performed cap analysis of gene expression{Kodzius:2006gy} (CAGE) on a large number of human and mouse cell types. Andersson et al. detected balanced, bidirectional transcription at many enhancers and predicted 43,011 human enhancers using the FANTOM5 CAGE data. This effort was augmented with additional CAGE data of 19 time courses upon various stimuli, resulting in an expanded set of 65,423 human enhancers{Arner:2015jb}. We compared ENCODE ccREs with both sets of FANTOM enhancers and found the majority of them covered by ccREs (81.4% of the earlier set and 77.5% of the latter set). The lower coverage of ccREs on the second set of FANTOM enhancers may reflect the lack of time-course ENCODE data for defining activated ccREs.

We also compared ccREs with FANTOM CAT RNAs, which was defined accounting for Roadmap DNase-seq data{Hon:2017}. The majority of the 59,011 FANTOM CAT RNAs overlap ccREs at their TSSs, with eRNA-like FANTOM CAT RNAs predominantly overlapping ccREs-ELS (**Supplementary Information**; **Extended Data Fig. 16f**). Notably, ccREs intersecting FANTOM enhancers are significantly closer to TSSs than those that do not intersect FANTOM enhancers; the median FANTOM-intersecting ccRE is 9,170bp from the nearest TSS compared to 11,845bp for non-FANTOM ccRE (Wilcoxon rank sum p=8.2E-265). This observation also holds for ccREs-ELS (median distances 13,196bp and 14,927bp, respectively; p=1.2E-59) and ccREs-PLS (median distances 957bp and 1,038bp respectively, p=8.1E-55) when considered individually. Furthermore, FANTOM-intersecting ccREs have highest epigenetic signals. Among the FANTOM enhancers that overlap a ccRE, nearly half (24,815 of 50,731; 48.9%) of them overlap a ccRE-ELS ranked in the top 15% by DNase and H3K27ac (the average rank of their max-Zs). ccREs that overlap FANTOM enhancers have significantly higher DNase, H3K27ac, H3K4me3, H3K4me1, and Pol II signals than ccREs that do not (p < 1.0E-300 for all comparisons, Wilcoxon sum rank test, **Extended Data Fig. 16a-e**).

Furthermore, the ccREs that overlap FANTOM enhancers are more evolutionarily conserved than the ccREs that do not, and this difference is observed for all three groups of ccREs (18.65% of ccREs-PLS, 3.48% of ccREs-ELS, and 1.17% of CTCF-only ccREs overlap FANTOM enhancers; **Extended Data Fig. 11e**). Notably, those 751 CTCF-only ccREs that overlap FANTOM enhancers are nearly twice as conserved as the other two groups of ccREs that overlap FANTOM enhancers. In summary, FANTOM enhancers overlap the strongest, albeit a small subset of ccREs.

ccREs exhibit bidirectional, cell-type specific transcription patterns

Motivated by the above comparison with FANTOM enhancers, we asked whether ccREs show evidence of bidirectional transcription. We explored GRO-seq data for GM12878 cells{Core:2014ez} and PRO-seq data for CD4+ T cells{Danko:2015co} at ccREs-ELS active in the corresponding cell types. On average, ccREs-PLS and ccREs-ELS deemed active in a cell type by epigenetic signals also show high GRO-seq or PRO-seq signals in the corresponding cell type (**Extended Data Fig. 17a**). ccREs-PLS show a burst of transcription in both sense and antisense directions, peaking at 150 bp to either side of the DNase-seq signal summit, with the sense strand signal rising approximately 2.5 folds higher above background than the antisense strand. Transcription on both strands, but particularly the sense strand, remains above local background past 2 kb in each direction. In contrast, ccREs-ELS display a symmetric burst of transcription on the Watson and Crick strands, peaking at the same height approximately 150 bp on either side of the DNase summit. The signal on both strands returns to local background at 1–1.5 kb from the DNase summit. Thus, ccREs show bidirectional transcription, at higher levels for ccREs-PLS than for ccREs-ELS.

We further divided ccREs-ELS into three groups according to their epigenetic activities in the two cell types. Among the ccREs-ELS that are active in GM12878 but not in CD4+ T cells (N = 20,533), 42.8% show above-background transcription on both strands, in sharp contrast with only 9.0% with transcription on both strands in CD4+ T cells (**Extended Data Fig. 17b**). Cell-type-specific bidirectional transcription was similarly observed for the ccREs-ELS active in CD4+ T cells but not in GM12878. Meanwhile, the majority of the ccREs that are active in both cell types (N = 3,148) show bidirectional transcription in both GM12878 (67.8%) and CD4+ T cells (58.8%).

We then asked whether the ccREs-ELS with transcription were more evolutionarily conserved than those without. Indeed GM12878 ccREs-ELS with transcription were significantly more conserved than the remaining GM12878 ccREs (average PhyloP scores at center of ccREs of 0.15 and 0.10 respectively, t-test *p*-value= 9.4E-5; passing normality test; **Extended Data Fig 11f**). We did not observe a significant difference in conservation between T-cell ccREs with transcription or without transcription. We hypothesize that this is because we call far fewer ccREs in T cells (11 k) than for GM12878 (23 k) and the ccREs we do call in T cells are of higher confidence. For example, the T-cell ccREs-ELS have high DNase Z-scores in a median of 67 biosamples, compared with a median of 34 biosamples for GM12878 ccREs-ELS (Wilcoxon rank sum test p-value<1E-300). In conclusion, ccREs-ELS show balanced, bidirectional transcription in a cell-type specific manner. The ccREs-ELS with transcription are more conserved than those without, indicating transcription is a useful feature for identifying functional ccREs.

SCREEN: A WEB ENGINE FOR SEARCHING AND VISUALIZING ccREs

We have built a web-based tool called SCREEN (Search Candidate cis-Regulatory Elements by ENCODE; http://screen.encodeproject.org) for searching and visualising each of the 1.31 M human ccREs and 0.43 M mouse ccREs in the Registry and their annotations, totalling 1.1 terabytes of processed data and metadata stored in two databases. Orthologous human and mouse ccREs are linked to each other. At multiple entry points, the user can select a subset of biosamples for which to display ccREs and the underlying epigenetic and RNA-seq signal profiles on the UCSC genome browser (**Supplementary Information**).

The first version of SCREEN is divided into three “apps”, each providing a different perspective on the ccREs (**Fig. 6**). The core app is the ccRE-centric search, where users can retrieve a subset of ccREs according to various characteristics, including genomic coordinates and signal profile within a selected cell type (**Fig. 6a**). The search app also provides a built-in details view, which lists other annotations at a ccRE, such as transcription factor binding sites, SNPs, topologically associated domains, genes, and TSSs, as well as showing clips of the DNase, H3K4me3, H3K27ac, and CTCF signals at the ccRE’s locus (mini-peaks) across all cell types with available ENCODE data. The gene-centric app plots RNA-seq and RAMPAGE expression data for genes and TSSs (**Fig. 6b**), and, for the mouse, plots differentially expressed genes and differential ccRE activities across cell types and developmental time points (see the first use case below). Finally, the SNP-centric GWAS app intersects ccREs with SNPs from 2,927 published GWAS, provides a list of cell types with the most enriched ccREs for each study, and allows the user to browse the ccREs by cell type (**Fig. 6c**). All three apps link to the UCSC genome browser for visualising the epigenetic signals in user-specified samples at a ccRE’s or a gene’s locus (**Fig. 6a**).

USE CASES OF THE ENCODE ENCYCLOPEDIA

The ENCODE Encyclopedia can be used in two ways. The various annotations in the Encyclopedia can be downloaded from the ENCODE Portal and further analysed along with the user’s own data. SCREEN allows users to directly search for ccREs in the Registry and explore all associated annotations. Here, we provide several use cases for the Registry of ccREs through SCREEN. The first use case explores mouse data as a panel of tissue types over a series of developmental time-points. We use SCREEN to present differentially expressed genes in a locus between pairs of time-points or tissues, along with differential H3K4me3 and H3K27ac signal levels of nearby ccREs-PLS and ccREs-ELS respectively. One major application of the Encyclopedia is to interpret the genetic variants uncovered by genome-wide association studies (GWAS); the other use cases illustrate how to characterise GWAS SNPs using ccREs.

Comparing ccRE activities in mouse tissues during development

As previously mentioned, the mouse component of ENCODE Phase III generated ChIP-seq data of eight histone modifications and RNA-seq data for 12 tissues across embryonic development. By integrating these two types of data at ccREs and their neighbouring genes, we can gain a better understanding of how regulatory element activity impacts gene expression across development. Specifically, we performed differential gene expression analysis between all available pairs of tissues and time-points (**Methods**) and configured SCREEN to display any differentially expressed gene and the differential biochemical activities of ccREs within 500 kb of it—activity here is defined as the H3K4me3 Z-score for ccREs-PLS and the H3K27ac Z-score for ccREs-ELS.

As an example, *Apoe* encodes apolipoprotein E, which is primarily produced by the liver and essential for cholesterol metabolism{Mahley:1988vs}. SCREEN illustrates that the expression of mouse *Apoe* is 5.1 fold higher at P0 than e11.5, supporting previous findings of increased *Apoe* expression at birth in rats {Mangeney:1989tl}. Nearby apolipoprotein C genes *Apoc1*, *Apoc2,* and *Apoc4* are likewise overexpressed at P0 (green boxes in **Fig. 6a**). Accordingly, ccREs-PLS and ccREs-ELS also show higher H3K4me3 and H3K27ac signals at P0 than at e11.5 (red and yellow dots in **Fig. 6a**).

In humans, *APOE* liver expression is regulated by two enhancers, hepatic control regions 1 and 2 (HCR.1 and HCR.2), which are 21 kb downstream of the *APOE* TSS{Allan:1997th}. These regions share high sequence similarity (85%) as a result of a 10 kb duplication of the region around *APOC1*. Human HCR.1 and HCR.2 map to a single HCR in mice. This mouse HCR overlaps ccREs-ELS EM10E0289438, which has higher H3K27ac signal at P0 than e11.5. EM10E0289438 has high H3K27ac signal primarily in the liver (**Extended Data Fig. 18a**) which is highly correlated with *Apoe* expression, increasing during development (correlation coefficient *r* = 0.98, **Fig. 6b,c**).

EM10E0289438 has two orthologous human ccREs-ELS: EH37E0492388 in HCR.1 and EH37E0492390 in HCR.2; a nearby mouse ccRE-ELS EM10E0289437 maps to EH37E0492391, a human ccRE-ELS that also overlaps HCR.2 (**Extended Data Fig. 18b**). Previous studies demonstrated that HCR.1 and HCR.2 differed in tissue specificity{Allan:1997th}: when transfected into mouse, HCR.1 led to a broad *Apoe* expression including hepatic and non-hepatic tissues while HCR.2 only resulted in liver-specific *Apoe* expression. Consistent with these previous results, SCREEN illustrates that the ccRE that overlap HCR.1 (EH37E0492388) has high H3K27ac signals in liver and many other tissue types such as the gastrointestinal system and brain whereas the two ccREs that overlap HCR.2 (EH37E0492390 and EH37E0492391) have liver-specific H3K27ac signals (**Extended Data Fig. 18c**).

Using the Registry of ccREs to annotate GWAS SNPs

Previous studies have repeatedly demonstrated that most GWAS variants reside outside coding exons and annotations of noncoding regions can guide the interpretation of GWAS variants by predicting disease-relevant cell types and regulatory factors{Ernst:2011kw, Maurano:2012hk, ENCODEProjectConsortium:2012gc, Andersson:2014bn, Farh:2014ka, Dickel:2016jg}. With the broad coverage of cell types and rich epigenetic and transcription factor binding data associated with the ccREs, the Registry can be particularly useful for annotating GWAS SNPs. We have preloaded SCREEN with 2,927 studies from the NHGRI-EBI GWAS catalogue{Hindorff:2009cc, MacArthur:2017kz} surveying ~1,800 phenotypes (**Supplementary Table 11a**). For GWAS that report 25 or more SNPs on a human population with linkage disequilibrium (LD) data, we tested each cell type for whether its ccREs-ELS were significantly enriched in the GWAS SNPs after accounting for SNPs in LD ($r^2 \geq 0.7$; **Methods; Extended Data Fig. 19a-c**). Once a cell type is selected, the biochemical activities of the ccREs that overlap the SNPs can aid further studies of their functions (**Extended Data Fig. 19d**). In this section, we delve into several SNPs in four studies, with the SNPs in first three studies already validated by functional assays.

The first GWAS example contains 75 SNPs significantly associated with red blood cell phenotypes{vanderHarst:2012cb}. These SNPs fall into 45 LD blocks ($r^2 \geq 0.7$) and 91% of the blocks contain at least one SNP that overlap a ccRE. We found these ccREs to be enriched for high H3K27ac signals in blood cells, particularly K562, an erythroleukemia cell line (**Fig. 8a**), reproducing previous reports on the K562 enrichment{RoadmapEpigenomicsConsortium:2015gq, Ulirsch:2016gm}.

Ulirsch *et al.* performed massively parallel reporter assays (MPRA) to functionally characterise 2,756 SNPs in LD with these 75 SNPs and found 32 to show significant differences in reporter expression between the two variants in K562 cells. They reported that 28% of their MPRA+ SNPs overlapped K562 DHSs{Ulirsch:2016gm}. We observed that 50% of the MPRA+ SNPs overlapped ccREs, but only 30% of the MPRA– SNPs overlapped ccREs (1.7 fold enrichment; *p* = 1.8E-2; **Fig. 8b**). This enrichment was even greater when we examined ccREs active in K562—25% of the MPRA+ SNPs overlapped a K562 ccRE compared with less than 10% of the MPRA– SNPs (2.6 fold enrichment; *p* = 9.7E-3). Thus, ccREs are predictive of functional SNPs.

Ulirsch *et al.* further tested three of the 32 MPRA+ SNPs using CRISPR/Cas9 in K562 cells and found all three SNPs to be functional{Ulirsch:2016gm}. Two of these SNPs, rs1175550 and rs1546723, overlap ccREs-PLS with high H3K4me3 and DNase signals in blood cell types, while the other SNP, rs737092, overlaps a ccRE-ELS (EH37E0606160). Ulirsch *et al.* found that rs737092 affected the expression of a gene 22.7 kb away, *RBM38*, which codes a RNA-binding protein. Consistent with this result, EH37E0606160 shows high H3K27ac and DNase signals across blood cell types including K562 (**Fig. 8c**). The orthologous mouse ccRE of EH37E0606160, EM10E0214140, also has high H3K27ac, H3K4me3, and DNase signals in blood cell lines and embryonic liver, a major contributor of hematopoiesis during development{Swain:2014jr}. Promoter-capture Hi-C (CHi-C) data linked rs737092 with *RBM38* in CD34+ hematopoietic progenitor cells{Mifsud:2015en}, as noted by Ulirsch *et al*. We capture this ccRE-gene link for EH37E0606160 in SCREEN.

The second example involves a SNP (rs2742624) that was identified by integrative analysis of genomic data (DNase-seq, ChIP-seq of H3K27ac and transcription factors, and expression quantitative trait loci or eQTL) on the prostate cancer cell line LNCaP and subsequently validated by transient transfection and CRISPR/Cas9 assays in the same cell line{Jin:2016ch}. The CRISPR/Cas9 data revealed that rs2742624 modulated the expression of *UPK3A*, a gene

3,984 bp away{Jin:2016ch}. SCREEN shows that rs2742624 overlaps EH37E0636504, a ccRE-ELS with the highest DNase Z-score in LNCaP cells and is linked to *UPK3A* by eQTLs in multiple tissues including prostate (**Extended Data Fig. 20a-d**). There are no ENCODE H3K27ac ChIP-seq data on LNCaP, but this ccRE-ELS shows high H3K27ac signal in PC-3, another prostate cancer cell line. Thus, SCREEN can facilitate integrative analyses that prioritise SNPs, such as those performed by Jin et al.

The third example is about a SNP (rs12740374) associated with 18 phenotypes (3 as a lead SNP and 15 in LD with a lead SNP), including cholesterol levels, coronary diseases, and metabolite levels (**Supplementary Table 11b**). The SNP overlaps EH37E0107819, a ccRE-ELS with high activities across many biosamples including hepatocytes, brain tissue, intestinal tissue, and stem cells (63/136 H3K27ac and 209/462 DNase experiments, **Extended Data Fig. 20e-h**). Warren et al. demonstrated through iPSC differentiation and CRISPR/cas9 assays that this region controls different genes depending on tissue context{Warren:2017en}. SCREEN shows ChIA-PET interactions and eQTLs between this ccRE-ELS and multiple genes including *CELSR2, PSRC1, and SORT1*, corroborating previous results of functional testing. We can therefore use SCREEN to prioritise new SNPs and identify novel disease-linked regions for further experimentation.

The fourth example involves rs1250568, which is in LD (r^2=0.7) with two SNPs associated with multiple sclerosis (MS), rs1250542{Patsopoulos:2011bq} and rs1250540{DeJager:2009bn} and was predicted by the deltaSVM algorithm to be a causal SNP {Lee:2015hi}. Overall, MS SNPs are enriched for ccREs active in T cells and B cells including lymphoblastoid cell lines (FDR = 1.3E-6), supporting previous findings {Maurano:2015ef}. rs1250568 lies in ccRE-ELS EH37E0182314, which has a high H3K27ac Z-score in GM12878 (**Extended Data Fig. 19d**). It overlaps a ChIP-seq peak for the transcription factor ELF1 and disrupts an ELF1 motif site (**Extended Data Fig. 19e**). *ELF1* is primarily expressed in lymphoid cells and is involved in the IL-2 and IL-23 immune response pathways, both of which have been implicated in multiple sclerosis{Gallo:1992wr, VakninDembinsky:2006um}. RNA Pol II ChIA-PET data links EH37E0182314 with both *ZMIZ1*, the gene containing rs1250568 within an intron, and *PPIF*, a downstream gene also known as Cyclophilin D (**Extended Data Fig. 19f**). *ZMIZ1* is in the androgen receptor signalling pathway and is expressed at lower levels in patients with multiple sclerosis than in controls{Fewings:2017jw}. *ZMIZ1* has been reported in the GWAS but *PPIF* has not{Patsopoulos:2011bq, DeJager:2009bn}. *PPIF* encodes a mitochondrial permeability transition pore protein. Knockdown or knockout of *Ppif* led to neuroprotective effects in mouse disease models of multiple sclerosis{Forte:2007df, Warne:2016de}. We predict that *PPIF* functions in lymphocytes which are associated with the demyelination of neighbouring neurons. In summary, SCREEN enables users both to identify the cell types that are likely implicated in a disease and to explore possible mechanisms by which ccREs and SNPs may cause the disease.

Combining human and mouse orthologous ccREs to interpret GWAS SNPs

One particular strength of the Registry is its inclusion of both human and mouse ccREs and the definition of orthologous ccREs in these two species. ENCODE has extensive data on mouse tissues during embryonic development; while human developmental tissues are impractical to obtain. Thus the orthologous mouse ccREs can complement the human ccREs in applications such as interpreting GWAS variants associated with developmental diseases, especially those that affect the brain.

For example, rs13025591 has been reported by two studies to be associated with schizophrenia (p-values 8E-8 and 6E-

6){SchizophreniaPsychiatricGenomeWideAssociationStudyGWASConsortium:2011bm, Bergen:2012il}. rs13025591 lies in the intron of the *AGAP1* gene, which is most highly expressed in bipolar spindle neurons and the eye in human and all assayed embryonic brain regions in mouse according to results contained in the Encyclopedia (**Extended Data Fig. 21a,b**). rs13025591 does not lie within a ccRE. Therefore, we hypothesized that the signal driving this genetic association arises from SNPs in high LD with rs13025591. There are five ccREs that overlap such LD SNPs (**Fig. 9a**), four of these ccREs show enhancer-like signatures and one shows a promoter-like signature. None of the five ccREs show a high H3K27ac or H3K4me3 signal in the surveyed adult human brain tissues associated with schizophrenia, such as the frontal temporal cortex or the angular gyrus{Niznikiewicz:2000ka, Nierenberg:2005jf}; nevertheless, EH37E0579839 has high H3K27ac signals in neural cells and bipolar spindle neurons.

SCREEN's Signal Profile tool, which displays DNase or histone modification signals at ccREs as "mini-peaks" across cell types, reveals that EH37E0579839 has high DNase signals in human foetal brain and eye tissues, but the signals disappear in older foetal brain and adult brain tissues (**Fig. 9b**). EH37E0579839 is orthologous to the mouse ccRE EM10E0042440, which shows enhancer-like signatures in brain tissues. Consistently with its human ortholog, EM10E0042440 has high DNase signals in embryonic brain and retina. Across twelve tissues at eight time-points of embryonic development, EM10E0042440 has the highest H3K27ac signals in brain regions (**Extended Data Fig. 21c**). In the forebrain, midbrain, and hindbrain, H3K27ac signals increase over time, reaching a maximum at e13.5. Then, similar to the orthologous human ccRE, H3K27ac signals at the mouse ccRE decrease after this time-point through birth (**Extended Data Fig. 21c**). These results indicate that this ccRE is active only during a narrow window of brain development. We tested this ccRE-ELS using transgenic mouse assays and found that this ccRE has enhancer activity exclusively in brain tissues at e11.5 (**Fig. 9d**)

The region harbouring these two orthologous ccREs is conserved across mammals (**Fig. 9d**). Although we do not have TF ChIP-seq data in human foetal brain or mouse embryonic brain tissues, motif analysis using both HaploReg and RegulomeDB{Ward:2011gh, Boyle:2012em} reveals that the LD SNP rs13031349 overlaps an SP3 motif and improves the motif match from a log-odds score of 8.1 to 19. Additional experiments are needed to test whether the SNP improves SP3 binding; nevertheless, using SCREEN and Encyclopedia, we were able to narrow down a region to suggest experimental testing for biological function.

CONCLUSION

In conclusion, we have integrated ENCODE and Roadmap Epigenomics data to produce an "Encyclopedia," a compendium of candidate functional elements in human and mouse. Because the precise biological functions of only a small fraction of ccREs have been experimentally tested to-date, the Encyclopedia enables the exploration of the regulatory roles of these elements in diverse cell types. The Encyclopedia serves as a continually improving resource to enable the broader scientific community to seek new information regarding biological processes and diseases.

METHODS

The ENCODE uniform processing pipelines are briefly described in **Supplementary Methods**. The scripts for generating the Registry of ccREs and performing the analysis on them as described above are available in a GitHub repository (https://www.github.com/weng-lab/encode_encyclopedia_v4), with a summary provided in **Supplementary Methods**.

Identifying rDHSs

We used all DNase-seq datasets as of February 1, 2017 with HOTSPOT2 calls on the hg19 or mm10 genomes (**Supplementary Table 5a**). For each dataset, we calculated the Z-score of the log of the DNase signals across the DHSs—see below for an explanation of Z-score of log(signal). We then selected a representative set of DHSs (rDHS) in the following steps. All DHSs passing an FDR threshold of < 0.1% were clustered across all DNase-seq experiments, and we selected the DHS with the highest signal (normalized as a Z-score to enable the comparison of signal levels across samples) as the representative DHS for each cluster. All the DHSs that overlapped with this rDHS by at least one bp were removed. We updated the clusters, identified the next rDHS with the highest signal, and removed all the DHSs that it represented. This process was repeated until it finally resulted in a list of non-overlapping rDHSs representing all DHSs. Using a modified version of a script from John Stamatoyannopoulos's laboratory, we iteratively cluster rDHSs and report those with the highest Z-score.

Normalizing epigenomic signals

For each rDHS, we computed the Z-scores of the $\log_{10}$ of DNase, H3K4me3, H3K27ac, and CTCF signals. Z-score computation is necessary for the signals to be comparable across biosamples because the uniform processing pipelines for DNase-seq and ChIP-seq data produce different signals. The DNase-seq signal is in raw read counts, whereas the ChIP-seq signal is the fold change of ChIP over input. We converted the DNase raw read counts into Z-scores to remove the effect of different sequencing depths.

Even for the ChIP-seq signal, which is normalized using a control experiment, substantial variation remains in the ranges of signals between cell types. To illustrate this effect, we examined the distributions of H3K27ac signals for 100 k randomly selected rDHSs across five different cell-types (**Extended Data Fig. 22a**). Even though these datasets were processed uniformly by the same pipeline, the ranges and distributions of signals differ among the datasets. After taking the $\log_{10}$ of the signals (**Extended Data Fig. 22b**), we observed that the distribution for each biosample roughly follows a normal distribution. The Z-scores of $\log_{10}$(signal) values have the same distribution across biosamples (**Extended Data Fig. 22c**).

To implement this normalization, we used the UCSC tool *bigWigAverageOverBed* to compute the signal for each rDHS (averaged across the entire rDHS for DNase and CTCF and across the entire rDHS plus 500 base-pairs on each end for H3K4me3 and H3K27ac). Using a custom Python script, we took the log of these signals and computed a Z-score for each rDHS compared with all other rDHSs within a cell type. rDHSs with a raw signal of 0 were assigned a Z-score of -10.
(*Extended Data Figure 22*)

Classifying ccREs

For cell type agnostic classification of ccREs with promoter-like, enhancer-like, or CTCF-only signatures, we first calculate the maximal DNase, H3K4me3, H3K27ac, or CTCF Z-scores across all cell and tissue types (called max-Z). Then, using these max-Zs and distance from the nearest TSS (GENCODE V19), we classify rDHSs into seven states according to the high-low combinations of their H3K4me3, H3K27ac, or CTCF max-Zs. These seven states are grouped into three general, mutually exclusive groups using the classification trees in **Fig. 3d** for human and **Extended Data Fig. 6a** for mouse. The rDHSs that were classified as having promoter-like, enhancer-like, or CTCF-only signatures are deemed ccREs and assigned an accession; the rDHS that are not classified in any of these categories are discarded.

To classify ccREs in a particular cell type, we use DNase, H3K4me3, H3K27ac, or CTCF Z-scores in that cell type. We have all four types of data for 21 cell types. The ccREs in each of these cell types are assigned to one of eight states—one inactive state (low DNase Z-scores) regardless of H3K4me3, H3K27ac, and CTCF Z-scores and seven active states (high DNase Z-scores) depending on the high-low combinations of their H3K4me3, H3K27ac, and CTCF Z-scores. The seven active states are further classified into five general, mutually exclusive groups: ccRE-PLS, ccRE-ELS, and CTCF-only are assigned according to the classification trees (**Fig. 3d** for human and **Extended Data Fig. 6a** for mouse), the DNase-only group contains ccREs with high DNase Z-scores but low H3K4me3, H3K27ac, and CTCF Z-scores, and the inactive group coincides with the inactive state, containing ccREs with low DNase Z-scores regardless of their H3K4me3, H3K27ac, and CTCF Z-scores.

To classify ccREs in a particular cell type that lacks one or more data types, we must make approximations. The scheme is illustrated in **Extended Data Fig. 6e** and summarized as follows. If both H3K4me3 and H3K27ac data are available, then we incorporate the TSS proximity information by following the classification trees (**Fig. 3d**, **Extended Data Fig. 6a**); otherwise, we classify PLS if only H3K4me3 data are available and ELS if only H3K27ac data are available, without considering TSS proximity of the ccREs. For cell types lacking DNase data, we also use the same classification scheme but without the DNase Z-score > 1.64 requirement. In these cell types, ccREs with low H3K4me3, H3K27ac, or CTCF signals are labelled “unclassified” because we are unable to definitively classify them as “inactive” without DNase data.
(*Figure 3*, *Extended Data Figure 6*)

Enrichment of TSS-distal ccREs-ELS near tissue-specific genes

Tissue specificity score {Yanai:2005dh} of each gene was calculated across 66 mouse datasets, each corresponding to a tissue at an embryonic developmental time point. In Each sample, tissue-specific and housekeeping genes were defined as the expressed genes (TPM ≥ 1) with top 1,000 and bottom 1000 tissue specificity scores, respectively. We computed the percentage of tissue-specific genes or housekeeping genes that had active ccREs-ELS within 10 kb of their TSSs, with active ccREs-ELS defined for the corresponding sample. The *p*-values of enrichment or depletion was estimated by randomly selecting 1,000 expressed genes (TPM ≥ 1) in the corresponding sample for 10,000 times.
(*Supplementary Table 7*)

Evolutionary conservation of ccREs

For the meta-ccRE conservation analysis, we calculated the conservation for groups of ccRE in reference to the ccRE center. We analyzed conservation stratifying by ccRE classification, concordance, overlap with GRO-seq/PRO-seq signal, overlap with CAGE peak, and sytney with the mouse genome. For each group of ccREs, we calculated the average phyloP score (calculated from the alignment of 46 vertebrate genomes **URL**) +/- 250 bp from the center of each ccRE.

For the ranked overlap analysis, we stratified ccRE-PLS, ELS, and CTCF-only into bins (0.1 intervals) based on their DNase max-Z. We then calculated the total number of bases in each bin that overlap GERP++regions (**download URL**).
(*Extended Data Figure 11*)

Orthologous human and mouse ccREs

Using UCSC's liftOver tool with a minimum match score of 0.5, we lifted hg19 ccREs to the mm10 genome (hg19-mm10 ccREs) and mm10 ccREs to the hg19 genome (mm10-hg19 ccREs). We labeled ccREs that did not map to the genome as "No orthology". For the ccREs that did map, we interesected them the other species ccREs (i.e. hg19-mm10 ccREs with mm10 ccREs and mm10-hg19 ccREs with hg19 ccREs). Regions that

We then intersected the liftOver mouse ccREs with human ccREs using the intersect function in bedtools{Quinlan:2010km}, requiring a minimum overlap of 1 bp, and define the overlapping human ccREs as the orthologous ccREs.
(*Extended Data Figure 11*)

Repeat and transposon contents of ccREs

Annotations of repetitive elements were downloaded from UCSC Genome Browser (human: hg19 rmsk.txt; mouse: mm10 rmsk.txt). Transposons comprise long-terminal repeats (LTRs), long interspersed elements (LINEs), and short interspersed elements (SINEs). Various transposon and repeat annotations were overlapped with all ccREs and the three groups ccREs-PLS, ccREs-ELS, ccREs-CTCF, respectively, and the total overlapping bps tallied. P-values were estimated using Chi-squared tests with 1,000 bps counted as one observation.
(*Supplementary Table 8*)

Saturation analysis of rDHSs

To estimate the percentage of all possible ccREs that have been sampled using our 440 DNA-seq datasets, we used a modified approach from ENCODE Phase II. We randomly selected *n* biosamples, where *n* is between 10–440 in intervals of 10. We then selected all corresponding DHSs for these biosamples (including their biological replicates) and calculated the number of resulting rDHS using the rDHS selection pipeline (described above). We then selected all rDHSs with a DNase maxZ > 1.64 across the selected biosamples. We performed this 100 times for bin. Since ccREs are a subset of these filtered rDHSs, we assumed this estimate to be an upper bound. Adapting the R script by Steven Wilder and Ian Dunham{ENCODEProjectConsortium:2012gc}, we calculated the complete set of rDHSs with maxZ > 1.64 to be at 74% saturation using a Weibull distribution.
(*Figure 4*)

Overlap of ccREs with H3K4me3, H3K27ac, and CTCF peaks in cell types without DNase-seq data

To determine the comprehensiveness of the Registry, we overlapped ccREs with ChIP-seq peaks (H3K4me3, H3K27ac, and CTCF) from cell types lacking DNase data. Using bedtools merge{Quinlan:2010km}, we merged all ChIP-seq peaks within 200 bp of one another and assigned each merged peak the maximal -log(FDR) score of the contributing peaks. We then filtered out all peaks with -$\log_{10}$(FDR) < 2. Using bedtools intersect with the "-u" flag, we intersected the merged peaks with ccREs and counted the number of unique peaks that overlapped at least one ccRE.
(*Extended Data Figure 12*)

Testing ccREs-ELS using transgenic mouse assays

We selected ccREs-ELS in three mouse tissues (e11.5 midbrain, hindbrain, and limb) for testing using transgenic mouse assays. We excluded ccREs-ELS that were within 2 kb of annotated TSSs or overlapped regions already in the VISTA database (www.enhancer.lbl.gov). We first ranked ccREs-ELS from the most to the least significant by the average rank of DNase and H3K27ac signals in the corresponding tissue and then selected ccREs-ELS from three segments of each tissue's ranked list (1-85, 1500-1550, and 3000-3050). We used H3K27ac peaks (called using the uniform processing pipeline) that overlapped the ccRE-ELS for the boundaries of the tested regions. In total, 151 regions were tested (**Supplementary Table 9**).
Transgenic mouse assays were performed in FVB/NCrl strain Mus musculus animals (Charles River) as described previously. Briefly, predicted enhancers were PCR amplified and cloned into a plasmid upstream of a minimal Hsp68 promoter and a lacZ reporter gene. The plasmids were pronuclear injected into fertilized mouse eggs and the transgenic embryos were implanted into surrogate mothers, collected at e11.5, and stained for β-galactosidase activity.
A predicted element is scored as positive enhancers if at least three embryos have identical β-galactosidase staining in the same tissue. Conversely, a prediction is deemed inactive if no reproducible staining was observed and at least five embryos harboring a transgene insertion were obtained.

We tested the ccRE that overlapped a LD SNP (rs13031349) of a SCZ SNP (EH37E0579839; human hg19 coordinates chr2:236825204-236826931) using transgenic mouse assays and obtained a positive result in brain tissues in two out of two e11.5 mouse embryos (**Fig. 9d**).
(*Figure 5, Figure 9, Extended Data Figure 14*)

Overlap of ccREs with ChromHMM states

We compared ccREs-PLS and -ELS to chromatin states called by ChromHMM. We combined similar ChromHMM states to generate seven broad states: active promoter (state 1) and weak promoter (state 2) are combined into *TSS*; poised promoter (state 3) corresponds to *TSS bivalent*; strong enhancer (states 4 and 5) are combined into *high-signal enhancer*; weak enhancer (states 6 and 7) are combined into *low-signal enhancer*; insulator (state 8) corresponds to *insulator*; transcription transition (state 9), transcription elongation (state 10), and weak transcription (state 11) are combined into *transcription*; repressed (state 12), heterochromatin low (state 13), repetitive/CNV (state 14), and repetitive/CNV (state 15) are combined into *repressed*.

To determine the ChromHMM state of ccREs, we intersected ccREs with ChromHMM states using bedtools and for each ccRE, selected the state that overlapped the largest number of basepairs, i.e. each ccRE was assigned to its majority ChromHMM state. For human, we analysed the ChromHMM regions for GM12878 cells (*ENCFF001TDH*) {cite PMID: 21441907} . We selected all ccREs with promoter-like or enhancer-like signatures and ranked them by H3K4me3 and H3K27ac Z-scores, respectively. Then, we calculated the percentage of ccREs in each 1 k bin that overlapped regions with each ChromHMM state. For mouse, we analysed 11 tissue–time-point combinations (from e11.5 and e14.5) for which we had DNase, H3K4me3, and H3K27ac data. We overlapped ccREs with promoter-like or enhancer-like signatures with ChromHMM states derived from eight histone marks in the same tissue-time-point.
(*Extended Data Figure 15, Supplementary Table 10*)

Comparison of ccREs-ELS with FANTOM enhancers

We downloaded the set of permissive human FANTOM5 Phase 1 enhancers, numbering 43,011 (http://fantom.gsc.riken.jp/5/datafiles/phase1.3/extra/Enhancers/hg19_enhancers.bed.gz) and the set of FANTOM5 Phase 1 and 2 enhancers for human, numbering 65,423 (http://fantom.gsc.riken.jp/5/datafiles/latest/extra/Enhancers/human_permissive_enhancers_phase_1_and_2.bed.gz), from the FANTOM consortium website on January 2, 2018. We intersected FANTOM enhancers and ccREs-ELS at single bp overlap using bedtools{Quinlan:2010km}. We then binned ccREs-ELS intersecting FANTOM Phase 1 and 2 enhancers, and those not intersecting FANTOM enhancers, by DNase, H3K4me3, H3K27ac, CTCF, and POL II Z-score with bin sizes of 0.1 and plotted the percentage of ccREs in each group falling into each bin in **Extended Data Figure 16a-e**.

(Extended Data Figure 16)

Bidirectional transcription at ccREs

BigWig signal files were downloaded from GEO for GRO-seq in GM12878 (GEO accession GSM1480326) and PRO-seq in CD4+ T-cell (GEO accession GSM1613181) for the plus and minus strands. For each cell type, the active ccREs-ELS (DNase Z-score >1.64, n=25,392 for GM12878 and n=13,050 for CD4+ T-cell) and ccREs-PLS (DNase Z-score >1.64, n=34,041 for GM12878 and n=30,023 for CD4+ T-cell) were assembled, and the DNase-seq summit positions were computed for each ccRE. Proximal ccREs-PLS were then grouped according to the strand on which the nearest gene resides, with distal ccREs-PLS and ccREs-PLS proximal to both strands discarded. GRO-seq and PRO-seq signal density was computed for the remaining ccREs from 2kb upstream to 2kb downstream of the DNase-seq summit in non-overlapping 25-bp bins on each strand. The signal values for each bin were then averaged across all ccREs-ELS and ccREs-PLS (**Extended Data Fig. 17a**).

To compare cell type-specific transcription profiles, we further divided the above ccREs-ELS into three groups: those active in GM12878 but not CD4+ T-cell (n=6,953), those active in CD4+ T-cell but not GM12878 (n=18,391), and those active in both GM12878 and CD4+ T-cell (n=3,148). To estimate background signal on each strand, we shuffled each group of ccREs 100 times and computed the mean and standard deviation of the signal density across the shuffled ccREs for each strand. Active ccREs with signal density two standard deviations above the mean on a given strand were considered to exhibit evidence of transcription on that strand. We repeated this process for all three groups in both cell types; results are shown in **Extended Data Fig. 17b**.

(*Extended Data Figure 17*)

Differential gene expression analysis

We downloaded gene expression quantification results from the ENCODE portal (**Supplementary Table 1b**). To compute differentially expressed genes between all possible pairs of tissues and time points (2,145 pairs in total for 66 RNA-seq samples), we ran DESeq2 (version 1.14.1) {Love:2014ka} with an FDR cutoff < 0.01.

(*Figure 7*)

Enrichment of GWAS variants in ccREs

We curated studies from the NHGRI-EBI Catalogue as of January 1, 2018 (**Supplementary Table S11a**). We excluded studies performed on mixed populations and populations without LD information (N=29). Using 1000 genomes linkage disequilibrium (LD) values from HaploReg {Ward:2011gh} for the corresponding super population (African, Ad Mixed American, Asian, and

European), we generated LD blocks containing all SNPs with $r^2>0.7$ and uploaded these SNPs to SCREEN for ccRE intersection.

For studies with more than 25 lead SNPs, we performed cell type enrichment analysis. For each study, we generated a matching set of control SNPs as follows: for each SNP in the study (*p*-value < 1.0E-6) we selected a SNP on Illumina and Affymetrix SNP ChIPs that fell within the same population specific MAF quartile and the same distance to TSS quartile (9,553, 39,530, and 154,279 bp demarcate the first, second, and third quartile, respectively). We repeated this process 100 times, generating 100 random control SNPs for each GWAS SNP. Then, for both GWAS and control SNPs, we retrieved all SNPs in high linkage disequilibrium (LD $r^2 > 0.7$), creating LD groups.

To assess whether the ccREs in a cell type were enriched in the GWAS SNPs, we intersected GWAS and control LD groups with ccREs with an H3K27ac Z-score > 1.64 in the cell type. To avoid overcounting, we pruned the overlaps, counting each LD group once per cell type. We modified the Uncovering Enrichment through Simulation (UES) method{Hayes:2015fx} with Fisher's exact tests for performing statistical testing. We calculated enrichment for overlapping ccREs, comparing the GWAS LD groups with the 100 matched controls. Finally, we applied an FDR of 5% to each study.
(*Extended Data Figure 19*)

FIGURE CAPTIONS

Figure 1 | ENCODE Phase III data production in Human. a. As of February 1, 2018, 4,596 human ENCODE Phase III experiments are available on the ENCODE Portal, categorized by assay and biosample types. The elements of the matrix indicate the numbers of experiments performed for a given assay and biosample type. **b.** The genome browser view of ENCODE phase III data around the *KRT8* gene for *in vitro* differentiated hepatocytes. CTCF: ChIP-seq of the DNA binding protein CTCF; WGBS: whole-genome bisulfite sequencing.
Link to Figure

Figure 2 | Overview of the ENCODE Encyclopedia
The Encyclopedia consists of two levels: the ground level (in the salmon color), which contains annotations from individual data types, and the integrative level (blue), which contains annotations from multiple data types, including the Registry of candidate cis-Regulatory Elements (ccREs; green). Both levels use data processed by the uniform processing pipelines (purple), which are available at the ENCODE portal. SCREEN integrates these data and annotations and allows users to visualise them in the UCSC Genome Browser.
Link to Figure
Link to a google doc for updating the ground level of Encyclopedia at the ENCODE Portal.

Figure 3 | Selection of ccREs and assignment of ccREs to seven states and three groups across all human cell types. a, Selection of human ccREs. We begin by clustering significant DHSs (FDR < 0.1%) to create representative DHSs (rDHSs). The rDHSs with high DNase max-Z and high max-Z for at least one other assay (H3K4me3, H3K27ac or CTCF) are ccREs (n = 1,310,152). **b,** Breakdown of ccREs by concordant support, missing data, and discordant support. **c,** Seven states of ccREs, defined by the combination of high or low H3K4me3, H3K27ac or CTCF max-Zs. DNase max-Z must be high. High max-Zs are indicated in colour and low max-Zs in gray. For ccREs in each state, the counts for TSS-proximal (within 2 kb of an GENCODE-annotated TSS; squares) and TSS-distal (circles) ccREs are shown in a bar plot. **d,**

We classify ccREs into three cell-type-agnostic groups according to their Max-Z and proximity to the nearest TSS, including ccREs with promoter-like signatures (cREs-PLS, n = 254,880), ccREs with enhancer-like signatures (cREs-ELS, n = 991,173), and ccREs bound by CTCF only (n = 64,099). **e,** Percentages of the mappable portion of the human genome occupied by the three groups of ccREs.
Link to Figure

Figure 4 | The impact of ENCODE Phase III data on the coverage of the Registry. a, To estimate the coverage of the current Registry of human ccREs, we generated rDHSs using varying numbers of cell types, randomly selecting the datasets each time. After performing this randomization 100 times for 10 to 440 cell types, we estimated the number of rDHSs at 95% saturation using a Weibull distribution (r^2=0.99), which revealed that the plateau corresponded to 1.76 M rDHSs with FDR < 0.1% and Z-score > 1.64. The green dot indicates that using only ENCODE phase II data would yield 0.75 M rDHSs and the green triangle indicates that using only Roadmap Epigenomics data would yield 0.78 M rDHSs. Note that the Roadmap data point falls below the curve because the Road Epigenomics project assayed tissues from multiple donors, resulting in a less diverse panel of cell types compared to the ENCODE Project.
Link to Figure

Figure 5 | **Prediction and validation of mouse embryonic enhancers. a,** Precision-Recall (PR) curves of 151 predicted enhancers in midbrain (blue), hindbrain (green), and limb (orange). Roughly 50 regions were predicted to be active in each tissue, and they were tested by transgenic mouse assays, which assessed activity in all tissues. Thus, the PR curves are for all 151 regions in each tissue. **b,** Validation rates of the 151 regions stratified by their prediction ranks in each tissue. The bottom portions of the bars with darker colours indicate validation of activity in the predicted tissue, and the top portions of the bars with lighter colours indicate a lack of activity in the predicted tissue but detection of activity in other tissues. **c–e,** (Top) Predicted enhancers (yellow boxes with their ccRE accessions) using DNase signal (green) and H3K27ac signal (yellow) in **c,** midbrain, **d,** hindbrain and **e,** limb. (Middle) Mouse staining images highlight the tissues in which each region was found to be active. The regions in **c** and **d** were each active in several tissues related to the primary predicted tissue. (Bottom) H3K27ac signal profiles across tissues accurately predict additional observed activity in related tissues (no available H3K27ac data for eye). EM10E0304559 is not active in forebrain despite a low level of H3K27ac signal. Tested regions are indicated in dashed boxes.
Link to Figure

Figure 6 | Overview of SCREEN. a, SCREEN's ccRE-centric search view. Using the facets on the main search page (top), users can retrieve ccREs (centre) according to genomic coordinates and signal profiles in a particular cell type. The four-quarter square to the right of each cell type indicates data availability, with colored quarters indicating available data—DNase-seq (green top-left quarter), H3K4me3 ChIP-seq (red top-right quarter), H3K27ac ChIP-seq (yellow bottom-left quarter), and CTCF ChIP-seq (blue bottom right quarter)—and white indicating lack of data. Two ccREs active in K562 are shown on chromosome 11 out of a total of 6,453 results for the shown criteria. Both ccREs are marked with blue stars indicating that they have high DNase and high H3K4me3, H3K27ac, or CTCF in the same cell type, i.e., they have "concordant" support. The top ccRE is marked with a "P" indicating that it is TSS-proximal (within 2 kb of a GENCODE TSS); the bottom ccRE is marked with a "D" indicating that it is TSS-distal. Four colours correspond to high values (≥1.64) for the four following epigenetic signals: DNase (green), H3K4me3 (red), H3K27ac (yellow), CTCF (blue). Grey indicates a low Z-score (<1.64) for the given mark. The ccRE details view shows neighbouring genes, bound transcription factors, and mini-peaks for epigenetic signals (bottom left, shown here for the top

ccRE in the search table). A trackhub is custom built for visualizing a ccRE or a gene and the supporting data using the UCSC Genome Browser (bottom right, top ccRE from the table highlighted in blue). TPM and FPKM are two units of gene expression level. **b,** SCREEN's gene-centric view provides RNA-seq and RAMPAGE derived expression levels for the genes and TSSs near the ccRE of interest. **c,** SCREEN's SNP-centric view displays ccREs that overlap SNPs from published GWAS and lends insight into which cell types may be relevant to a particular phenotype. The cell type is shown for an inflammatory bowel disease GWAS, along with one ccRE active in CD4+ T cells that contains a SNP from that study.
Link to Figure

Figure 7 | Analysing differential gene expression and ccRE activity across mouse developmental time-points. a, Comparison between liver e11.5 and P0 gene expression and ccRE activity at the *Apoe* locus. Green bars indicate differentially expressed genes, and red and yellow dots indicate ccREs-PLS and ccREs-ELS. The widths of the green bars represent gene lengths. The lines beneath the green bars and above the gene names indicate the positions and orientations of the genes—red for plus genomic strand and blue for minus strand. The heights of bars or dots indicate changes—Log_2(fold change) or difference in Z-score—between the two time-points. The ccRE-ELS EM10E0289438 that overlaps a hepatic control region (HCR) is outlined in black. Outlined in blue is another ccRE-ELS EM10E0289437 that has correlated H3K27ac signal levels with EM10E0289438. **b,** Genome browser view of the *Apoe* locus with H3K27ac, DNase, and RNA-seq signals in liver across all surveyed time-points. The two ccREs-ELSs outlined in **a** are shown as yellow boxes at the top of the figure. **c,** *Apoe* gene expression and HCR ccRE H3K27ac levels increase coordinately during development.
Link to Figure

Figure 8 | Annotating variants associated with red blood cell traits using SCREEN. a, Overlap of red blood cell (RBC) SNP LD Blocks with ccREs: 41 of the 45 LD blocks contain at least one SNP that overlaps a ccRE. These ccREs are enriched for H3K27ac signal in blood cell types, particularly K562. **b,** Overlap of ccREs with MPRA positive (MPRA+) and negative (MPRA−) regions containing RBC variants{Ulirsch:2016gm}. MRPA+ regions more frequently overlap ccREs than MRPA- regions (*p* = 1.8E-2), particularly ccREs active in K562 (*p* = 9.7E-3). **c,** Genome browser view of RBC SNP rs737092, which is downstream of the *RBM38* gene. This variant overlaps a human ccRE-ELS (EH37E0606160), which has high DNase signal (green) and high H3K27ac signal (yellow) in K562. This ccRE-ELS also has high H3K27ac signals in other blood cell types, indicated by the table shaded in yellow. rs737092 also overlaps an orthologous mouse ccRE (EM10E0214140, brown) which has high H3K27ac signal in fetal liver (table shaded in brown). **d,** Genome browser view of the *RBM38* locus. The ccRE-ELS EH37E0606160 is linked to the promoter of *RBM38* by CHi-C assay.
Link to Figure

Figure 9 | Interpreting GWAS variants using SCREEN. a, H3K4me3 and H3K27ac Z-scores for ccREs containing SNPs in LD with the schizophrenia-associated SNP rs13025591. In the table below the SNPs and ccREs, H3K4me3 Z-scores and H3K27ac Z-scores are displayed in red and yellow for ccREs with promoter-like and enhancer-like signatures, respectively. Only ccREs that overlap a SNP are included in the table and Z-scores lower than 1.64 are omitted. Of particular interest is ccRE-ELS EH37E0579839 (in bold), which has high H3K27ac levels in neural cells and bipolar neurons and harbours the SNP rs13031349. **b,** SCREEN's Signal Profile tool allows users to view DNase peaks at ccREs across all cell types with data. Both the human ccRE EH37E0579839 and its orthologous mouse ccRE EM10E0042440 show high DNase signals in developing brain and eye tissues. **c,** The H3K27ac signal at EM10E0042440 over developmental time in mouse forebrain (red), midbrain (green) and hindbrain (blue). **d,**

(Left) Zoomed-in view of EH37E0579839. The SNP rs13031349 overlaps both EH37E0579839 and EM10E0042440. The SNP also overlaps an SP3 motif and results in a change in the motif score. (Right) Transgenic mouse assays indicate that the human ccRE is active in mouse e11.5 brain tissues, with two stained mouse embryos shown.
Link to Figure

CAPTIONS FOR SUPPLEMENTARY TABLES

Table S1 | Summary of data produced during ENCODE3
Link to Table S1

Table S2 | VISTA regions used for evaluating epigenetic signals
Link to Table S2

Table S3 | The performance (AUPR) of using epigenetic signals to predict VISTA enhancers
Link to Table S3

Table S4 | Using epigenetic signals to predict transcript expression
Link to Table S4

Table S5 | Input datasets for building the Registry of ccREs
Link to Table S5

Table S6 | The relative abundance of ccREs-PLS and ccREs-ELS
Link to Table S6

Table S7 | ccREs are near tissue-specific genes.
Link to Table S7

Table S8 | The transposon and repeat content of ccREs
Link to Table S8

Table S9 | Testing candidate enhancers with transgenic mouse assays
Link to Table S9

Table S10 | Regions and genomic positions of ChromHMM promoter and enhancer states that overlap ccREs.
Link to Table S10

Table S11 | a, List of 2,927 GWAS that are included in SCREEN. b, the GWAS that uncovered SNP rs12740374
Link to Table S11

CAPTIONS FOR EXTENDED DATA FIGURES

Extended Data Figure 1 | ENCODE data production. a. As of February 1, 2018, 7,385 human ENCODE Phase II & III experiments are available on the ENCODE Portal, categorized by assay and biosample types. **b.** As of February 1, 2018, 1,141 mouse ENCODE Phase III experiments

are available on the ENCODE Portal, categorized by assay and biosample types.
Link to Figure

Extended Data Figure 2 | Using RAMPAGE data for identifying TSSs and quantifying tissue-specific transcript levels. a, RAMPAGE signals (purple) and poly-A RNA-seq signals (green) across four human tissues (heart, skin, spleen, and testis) at *EP300*. RAMPAGE signals reveal that although both the GENCODE- (TSS1, black) and UCSC-annotated (TSS2, grey) TSSs for *EP300* are active, one TSS is used far more frequently than the other. Each of the two TSSs is in a dashed-line rectangle. **b,** A scatter plot shows the RAMPAGE signals of the two *EP300* TSSs across cell types, indicating that the upstream TSS is less active than the downstream TSS in most cell types. Each point corresponds to a RAMPAGE dataset in a particular cell type. **c,** RAMPAGE signals (purple) and poly-A RNA-seq signals (green) across four human tissues at *ARHGAP23*. RAMPAGE data identified a cell-type-specific TSS for *ARHGAP23*, located 9.2 kb upstream of a ubiquitous TSS. Each of the two TSSs is in a dashed-line rectangle. **d,** A scatter plot shows the RAMPAGE signals of the two *ARHGAP23* TSSs across cell types, indicating that the upstream TSS is active mostly in testis and two brain regions (occipital lobe and temporal lobe), while the downstream is active in most cell types.
Link to Figure

Extended Data Figure 3 | DNA replication timing programmes are cell type-specific. a, Genome-wide RT programs were obtained for distinct human cell types, including embryonic stem cell (hESC)-derived cells, primary cells and established cell lines representing intermediate stages of endoderm, mesoderm, ectoderm, and neural crest (NC) development. Solid arrow lines depict the in vitro differentiation pathways of the distinct cell types from hESCs; dashed arrows depict the embryonic origin of the cell types not derived from hESCs (primary cells and cell lines). ENCODE IDs are shown for datasets and protocols for each cell type. **b,** Hierarchical clustering of RT programs from the distinct human cell types. Branches of the dendrogram were constructed based on the correlation values between distinct cell types (distance = 1 – Pearson correlation value). Specific clusters of cell types are indicated at the bottom: pluripotent, definitive endoderm (DE), liver and pancreas, neural crest and mesoderm cell types, neural precursors (NPC), myeloid and erythroid progenitors and lymphoid cells, neural crest (NC), mesendoderm (MED), definitive endoderm (DE), lateral plate mesoderm(LPM), splanchnic mesoderm (Splanc), mesothelium (Mesothel), smooth muscle (SM), myoblasts (Myob), fibroblasts (Fibrob), mesenchymal stem cells (MSC), and neural progenitor cells (NPC).
Link to Figure

Extended Data Figure 4 | Integrative analyses of RBP data can identify genetic variants that may impact RBP regulation. Intron 66 of *UTRN* (dystrophin-related protein 1) harbours an RBFOX2 eCLIP peak downstream of an alternatively spliced exon, which overlaps an ExAC variant {Lek:2016bi}, and this G→C variant disrupts the RBFOX2 binding motif (GCAUG) at the first position (**Bottom right**). RBNS data reveal that this variant substantially changes the RBFOX2 binding site—the top 5-mer has an enrichment value of 13.58 for the major G allele but 0.89 for the C variant (**Bottom left**), thus suggesting that the mutation disrupts RBFOX2 binding in vivo. RBFOX2-knockdown and control RNA-seq data on HepG2 cells indeed support this hypothesis (**Top left**): inclusion of the alternatively spliced exon 66 was reduced from 87% in control cells to 29% in RBFOX2 knock-down (KD) cells. Exon inclusion is measured as Percent Spliced In (Ψ) by the MISO tool {Katz:2010iv} and averaged over two biological replicates (**Top right**).
Link to Figure

Extended Data Figure 5 | Testing methods of enhancer and promoter prediction. a, PR curves for midbrain, hindbrain, neural tube, and limb enhancers at e11.5. Colours indicate peaks and signals used for anchoring and ranking the enhancer predictions. All peaks were set to 300 bp centred on their summits. **b,** PR curves for midbrain, hindbrain, neural tube, and limb enhancers at e11.5. All predictions were anchored on DHSs in the respective tissue. Colours indicate signals used for ranking predictions; gray indicates the average of DNase and H3K27ac signals. **c,** Scatter plots demonstrating correlation of transcript expression in midbrain at e11.5 with H3K4me3 peaks ranked according to the H3K4me3 signal (*r* = 0.59), H3K4me3 peaks ranked according to the DNase signal (*r* = 0.19), DHSs ranked according to the H3K4me3 signal (*r* = 0.76), DHSs ranked according to the DNase signal (*r* = 0.45)**.**
Link to Figure

Extended Data Figure 6 | Selection and classification of ccREs. a, Methods for classifying ccREs in mouse. This panel corresponds to Fig. 3a and Fig. 3d combined, but for mouse. We begin by clustering high quality DHSs (FDR < 0.1%) to create representative DHSs (rDHSs). We then select ccREs from rDHSs and classify the ccREs by max-Zs of DNase, H3K4me3, H3K27ac, and CTCF. **b,** Percent of the DNase-mappable (36 nt, single-end reads) mouse genome covered by each group of ccREs. **c,** Number of human cell types with complete or partial assay coverage, and the ccRE groups that can be classified with available data. **d,** Number of human ccREs in each of the nine states in GM12878 cells, assigned according to whether they have high Z-scores (> 1.64, or 95th percentile) for H3K4me3, H3K27ac, CTCF, and DNase in that cell type. ccREs with low DNase are classified as inactive regardless of the Z-scores for the other marks. Each ccRE is further classified as being either proximal (≤ 2 kb) or distal (> 2 kb) to the nearest GENCODE-annotated TSS. Icons mark the states to the left of the bars. Coloured boxes (for proximal ccREs) and pie quarters (for distal ccREs) represent high Z-scores, and gray represents low Z-scores. **e,** Number of human ccREs in each of the five groups in GM12878 cells: with promoter-like signatures (PLS), with enhancer-like signatures (ELS), CTCF-only, DNase-only, and inactive.
Link to Figure

Extended Data Figure 7 | Clustering of human biosamples on the basis of DNase or H3K27ac signals at ccREs. Human biosamples separated by biosample type (tissues **a** and **b**; primary cells **c** and **d**) clustered according to the Jaccard similarity coefficients of ccREs with high DNase (**a** and **d**) or H3K27ac (**b** and **c**) signals. **a,** The tissues clustered using DNase signal segregate according to their organs of origin, with each given a different colour. **b**, When clustered using H3K27ac signal, tissues from different regions of the same organ cluster together, e.g., the various brain regions (pink). Foetal and adult tissues often aggregate together (e.g., foetal and adult adrenal gland, black). The samples from the gastrointestinal tract form two clusters, one reflecting smooth muscles (the purple and maroon samples at the top) and the other reflecting mucosa (the maroon samples at the centre). **c.** Human primary cells, when clustered using H3K27ac signal, segregate perfectly into three groups coloured by their embryonic origins, including blood (red), non-blood mesoderm (yellow), and ectoderm (blue). Even the endothelial cells of the umbilical vein, which are derived from the extraembryonic mesoderm, clustered with the cell types derived from the embryonic mesoderm (fibroblasts, myoblasts, osteoblasts, and astrocytes). **d.** Human primary cells hierarchically clustered using DNase signal segregate into large clusters corresponding to lineages. The left cluster (red) composed entirely of blood cells, subdivided into to the myeloid and lymphoid lineages. The leftmost subcluster of the right cluster contains the four trophoblast samples (in black), thus reflecting their extraembryonic fate. The rightmost subcluster contains mostly fibroblasts, and

the middle subcluster contains endothelial cells, epithelial cells, keratinocytes, and melanocytes. The fibroblasts aggregate together regardless of their anatomical locations, as do most of the endothelial cells, in agreement with their common mesodermal origin. Most of the epithelial cells also cluster together, despite their different embryonic germ layers.
Link to Figure

Extended Data Figure 8 | Clustering of mouse tissues by epigenetic signals at ccREs. Mouse embryonic tissues were hierarchically clustered according to the Jaccard similarity coefficient of ccREs with high **a,** H3K27ac **b,** H3K4me3 **c,** DNase and **d,** CTCF (Z-score > 1.64). Colours indicate the organs of origin of the tissues. When clustered according to H3K27ac signals at ccREs (**a**), the tissues segregate completely according to their organs of origin.
Link to Figure

Extended Data Figure 9 | Transcription factor ChIP-seq signals support the five-group classification. a, Violin plots show the average Pol II, EP300, and RAD21 ChIP-seq signals for ccREs belonging to each of the nine ccRE states. ccREs proximal and distal to the nearest TSSs are displayed separately. Median values are displayed along with the number of ccREs in each state. Colors in the boxes on the left indicate high epigenomic signal. Colors of violins indicate five group classifications (PLS, red; ELS, yellow; CTCF-only, blue; DNase-only, green; inactive, gray) Scatterplots of **b,** the median EP300 signal or **c,** the median RAD21 signal vs. the median Pol II signal for each ccRE state in GM12878. The size of an icon is proportional to the number of ccREs in that state except for the inactive state. Proximal ccREs are represented by square icons. Distal ccREs are represented by circular icons. Five group classifications are shown in red (PLS), yellow (ELS) and blue (CTCF-only)
Link to Figure

Extended Data Figure 10 | UCSC Genome Browser views of ccREs by state and group classifications. Browser views of hepatocyte, bipolar spindle neuron, and B cell ccREs near the TSSs of **a,** *SPI1* **b,** *NPAS4*, and **c,** *HNF4A*. The cell type showing high activity at the promoter region of each gene is highlighted in larger font. Right beneath the gene symbols, the cell-type agnostic seven-state and three-group classifications are shown for each locus. Further down, a group of tracks are shown for each individual cell type: the nine-state classification, the five-group classification, the signals of DNase (green), H3K4me3 (red), H3K27ac (yellow), and CTCF (blue). For a state classification, coloured boxes indicate high Z-scores (> 1.64) and gray boxes indicate low Z-scores for the respective assays. For a group classification, red indicates ccREs-PLS, yellow ccREs-ELS, and blue CTCF-only ccREs.
Link to Figure

Extended Data Figure 11 | Conservation of human ccREs. a, Average phyloP score across +/- 250 bp from the center of each ccRE. ccREs-PLS (reds), ccREs-ELS (yellows), and CTCF-only ccREs (blues) are each plotted in two groups: those with concordant support (dark colors) and others (light colors). In gray are 500k 300 bp regions randomly selected from mappable regions of the human genome. **b,** The percentage of positions in ccREs-PLS (red), ccREs-ELS (yellow), and CTCF-only ccREs (blue) that overlap the GERP++ set of evolutionarily conserved regions{Davydov:2010dg} binned by DNase maxZ score. Bins with fewer than 20 ccREs are omitted. **c,** Fraction of human and mouse ccREs with orthology to the mouse or human genome, respectively. In black (no orthology) are ccREs that do not map to the reciprocal genome. In dark blue (orthology only) are ccREs that map to the reciprocal genome but do not overlap a ccRE in that genome. In light blue (orthology & ccRE) are ccREs that map to ccREs that reciprocal map to the original genome. **d,** Average phyloP scores (as described in **a**) across

human ccREs stratified by orthology categories defined in **c**. **e,** Average phyloP scores (as described in **a**) across human ccREs stratified by ccRE group and presence of a FANTOM CAGE peak. ccREs overlapping a CAGE peak are designated by a thin black line. **f,** Average phyloP scores (as described in **a**) across ccREs-ELS active in CD4+ T cells and ccREs-ELS active in GM12878 respectively further stratified by whether they show transcription, as measured by GRO-seq data in GM12878 and PRO-seq data in T cells.
Link to Figure

Extended Data Figure 12 | Coverage of the current Registry of ccREs. a-c, Percentages of human H3K4me3, H3K27ac and CTCF ChIP-seq peaks from cell types without DNase data that are covered by human ccREs. On average 89.7%, 86.8%, and 99.1% of H3K4me3, H3K27ac, and CTCF peaks overlap a ccRE, respectively. **d-e,** Percentages of mouse H3K4me3 and H3K27ac ChIP-seq peaks from cell types without DNase data that are covered by mouse ccREs. On average 95.8% and 87.6% of H3K4me3 and H3K27ac peaks overlap a ccRE, respectively. **f-g**, in human (**f**) and mouse (**g**), cell-types with peaks that have a lower average -$\log_{10}$(FDR) across all peaks tend to have a lower percentage of peaks covered by ccREs. Manual inspection reveals that this low coverage is due to low-signal, false-positive peaks called by the algorithm for these datasets.
Link to Figure

Extended Data Figure 13 | Estimation of ccRE validation rate using published MPRA results. a, percentage of MPRA-positive (black) and MPRA-negative (grey) regions from lymphoblastoid cell lines (LCLs) intersecting cell type-agnostic ccREs (left) and ccREs active in various immune (center) and non-immune (right) cell types. MPRA-positive regions are significantly more likely to intersect a cell type agnostic ccRE than MPRA-negative regions (Chi-square p=1e-79); additionally, significantly more MPRA-positive regions are intersected by immune ccREs than non-immune ccREs (Wilcoxon rank sum p=1.3e-17). **b,** PR curves for the prediction of MPRA activity in LCLs using ccREs ranked by max-Z (blue), GM12878 DNase Z-score (red), and trophoblast Z-score (green), highlighting that LCL ccREs are more predictive of MPRA activity in LCLs than cell type agnostic ccREs or non-immune ccREs. **c,** average fold enrichment for MPRA-positive intersection over MPRA-negative intersection for ccREs-ELS (yellow) and ccREs-PLS (red) active in LCLs (left), other immune cell types (center), or non-immune cell types (right). ccREs-ELS show cell type-specific enrichment, while ccREs-PLS do not. **d,** MPRA-positive intersection by ccREs-ELS for all cell types ranked by fold enrichment over MPRA-negative intersection; the top two ranks are LCLs, and the top 150 ranks are enriched for immune cell types (Kolmogorov-Smirnov p=2.2e-12).
Link to Figure

Extended Data Figure 14 | In vivo testing of ccREs-ELS. Shown are representative transgenic embryonic day 11.5 (e11.5) mouse images for the predicted enhancers that displayed reproducible activity in the expected tissue type. Enhancer predictions were performed using a combination of H3K27ac and DNase profiling for mouse e11.5 midbrain (blue), hindbrain (green), or limb (orange) tissue. Predicted enhancers were selected from three rank classes (Top; Middle: ~ 1,500; Bottom: ~ 3,000) and tested for activity using transgenic mouse assays (see **Methods**). Blue staining indicates enhancer activity, and the unique identifier below each embryo (mm number) corresponds to the accession of the enhancer in the VISTA Enhancer Browser (www.enhancer.lbl.gov).
Link to Figure

Extended Data Figure 15 | Overlap of ccREs with ChromHMM states. **a-b,** Percentages of GM12878 ccREs that overlap ChromHMM states. We first ranked ccREs-PLS (**a**) and ccREs-

ELS (**b**) by their H3K4me3 and H3K27ac Z-scores respectively. For each bin of 1 k ccREs, we then calculated the percentage of ccREs overlapping each ChromHMM state in GM12878. **c-d,** Percentages of mouse ccREs that overlap ChromHMM states in the corresponding tissue. We included all tissue–time-point combinations with both DNase and histone data. We calculated the percentages of ccREs-PLS (**c**) and ccREs-PLS (**d**) that overlapped each ChromHMM state in the same tissue and time point.
Link to Figure

Extended Data Figure 16 | Overlap of ccREs with FANTOM Enhancers and FANTOM CAT genes. Density plots of the Z-scores of ccREs intersecting FANTOM enhancers (orange) and not intersecting FANTOM enhancers (blue). Z-scores are plotted for **a,** DNase; **b,** H3K4me3; **c,** H3K27ac; **d,** H3K4me1; and **e,** Pol II. **f,** percentages of FANTOM CAT transcripts in the eleven FANTOM-defined categories that overlap ccREs-PLS (red) or ccREs-ELS (yellow) within 2 kb of their TSSs. Categories are, from left to right, **i**, divergent lncRNAs; **ii**, protein coding mRNAs; **iii**, small RNAs; **iv**, structural RNAs; **v**, pseudogenes; **vi**, uncertain coding; **vii**, short ncRNAs; **viii**, sense overlap RNAs; **ix**, intergenic lncRNAs; **x**, sense intronic lncRNAs; and **xi**, antisense lncRNAs. More than 80% of coding-associated genes (**i**, lncRNAs that are divergent from a protein-coding mRNA and **ii**, protein-coding mRNAs) are annotated by a ccRE-PLS, and more than 70% of eRNA-like non-coding RNAs (**vii-xi**, short ncRNAs, sense overlap lncRNAs, intergenic lncRNAs, sense intronic lncRNAs, and antisense lncRNAs) are annotated by a ccRE-ELS.
Link to Figure

Extended Data Figure 17 | Transcription patterns at ccREs. a, GRO-seq signal in GM12878 (left) and PRO-seq signal in CD4+ T cells (right) averaged over all ccREs-ELS and ccREs-PLS active in the given cell type, in a ± 2 kb window centred on the DNase signal summit of each ccRE. ccREs-PLS were grouped by the orientations of their associated genes, with distal ccREs-PLS and ccREs-PLS proximal to genes on both genomic strands excluded. ccREs-ELS were grouped by their own genomic strands. Genomic background signal, computed as described in supplemental methods, is shown by the grey tick marks and was approximately 0.02 for both strands in GM12878 and 0.03 for both strands in CD4+ T-cell. **b,** The percentage of ccREs-ELS with GRO-seq or PRO-seq signal significantly above background on both strands (black), the plus strand only (gray), the minus strand only (gray), or neither strand (unfilled) in GM12878 (left four bars in each plot) and CD4+ T cell (right four bars in each plot). Left, ccREs-ELS active in GM12878 but not CD4+ T cells; center, ccREs-ELS active in CD4+ T cells but not GM12878; right, ccREs-ELS active in both GM12878 and CD4+ T cells.
Link to Figure

Extended Data Figure 18 | Activity of human and mouse ccREs that overlap hepatic control region (HCR). a, H3K27ac and DNase Z-scores in mouse biosamples at EM10E0289438. **b,** Genome browser view of HCR.1 and HCR.2 in human, their overlapping ccREs, and epigenomic signals in liver tissue. **c,** H3K27ac activity in human biosamples for EH37E0492388 (HCR.1), EH37E0492390 (HCR.2) and EH37E0492391 (HCR.2). Biosamples are grouped by tissue ontology.
Link to Figure

Extended Data Figure 19 | Annotating GWAS variants using SCREEN. a, Users can select from a preloaded list of GWAS. For each study, we included all tagged SNPs reported in the study and all SNPs in LD with the tagged SNPs (r^2 > 0.7). **b,** SCREEN reports the percentage of LD blocks of a GWAS with at least one SNP overlapping a ccRE. **c,** SCREEN ranks cell and tissue types by enrichment in H3K27ac signals. Top cell and tissue types are displayed here for

each study. **d,** The user can narrow the search by selecting a cell type, such as GM12878, for multiple sclerosis (MS), and analyse the overlapping ccREs. In particular, two SNPs (rs1250566 and rs1250568) overlap two ccREs-ELS (EH37E0182316 and EH37E0182314, respectively) with high H3K27ac signals in GM12878 (highlighted in yellow). **e,** The two ccREs in **d** show strong DNase and H3K27ac signals in GM12878 and overlap the ChIP-seq peaks of the transcription factor ELF1. In particular, rs1250568 overlaps a high-scoring position of the ELF1 motif, with the major allele disrupting the motif score. **f,** A zoomed-out genome browser view of the *ZMIZ1* and *PPIF* locus, with ChIA-PET data (purple) linking the two ccREs-ELS (EH37E0182316 and EH37E0182314) to the promoters of both genes.**e,** Zoomed-in genome browser view of MS-associated SNP rs1250568, which overlaps an ELF1 ChIP-seq peak (blue box) and an ELF1 motif. **f,** Zoomed-out genome browser view of the locus showing POL2 ChIA-PET links between rs1250568 and two genes *ZMIZ1* and *PPIF*. **a,** Gene expression of *ZMIZ1* from whole-cell RNA-seq assays shown in tags per million (TPM). **b,** The RAMPAGE signal at the TSS of ENST00000472035.1 (averaged over ± 50 bp window). **c,** Gene expression of *PPIF* from whole-cell RNA-seq assays shown in tags per million (TPM). **d,** The RAMPAGE signal at the TSS of ENST00000225174.3 (averaged over ± 50 bp window). Bars are coloured according to the tissue of origin indicated on the left.
Link to Figure

Extended Data Figure 20 | Functional SNP examples related to prostate cancer and liver traits. a, Genome browser view of rs2742624, which was previously shown to regulate expression of *UPK3A* with epigenomic signals from LNCaP and PC-3 (prostate cancer cell lines) and prostate gland tissue. **b-c,** Top ten biosamples ranked by Z-score for (b) DNase and (c) H3K27ac at EH37E0636504. **d,** Genes linked to EH37E0636504 by eQTLs and ChIA-PET. **e,** Genome browser view of rs12740374 with epigenomic signals from three cell types demonstrating the ubiquitous activity of EH37E0107819. **f-g,** Top ten biosamples ranked by Z-score for (b) DNase and (c) H3K27ac at EH37E0107819. **h,** Genes linked to EH37E0107819 by eQTLs and ChIA-PET.
Link to Figure

Extended Data Figure 21 | SCREEN display of *AGAP1* expression levels. a, In human, *AGAP1* is expressed across many adult tissues. **b,** In mouse, *Agap1* is primarily expressed in embryonic brain tissues. Expression values were calculated from whole-cell RNA-seq experiments and are displayed in tags per million (TPM).**c,** H3K27ac signals measured as fold-change between ChIP and input are displayed across 12 tissues and 8 time-points. Tissues without H3K27ac ChIP-seq data are left blank. The maximal height of the signals is 10.
Link to Figure

Extended Data Figure 22 | Method for normalizing epigenomics signals. **a,** Distribution of the H3K27ac signals of 100 k randomly selected rDHSs from five cell types (B cell, liver, K562, T cell, and GM12878; shown in different colours). rDHSs with H3K27ac signals higher than 5 (n < 5.7 k) are omitted from this histogram. **b,** Distributions of the $\log_{10}$ of the H3K27ac signals in **a**. The log(signal) values of the rDHSs in each cell type roughly follow a normal distribution. **c,** Distribution of the Z-scores corresponding to the $\log_{10}$(signal) values in **b**. Signal values of zero are assigned a Z-score of −10.
Link to Figure

REFERENCES

{papers2_bibliography}

THE ENCODE CONSORTIUM (complete list of authors)

https://docs.google.com/document/d/1dk2lhwL9H8pf5hV9nGK5WBvJZFE5geONJ98oQrck-eY/edit?usp=sharing

AFFILIATIONS

SUPPLEMENTARY INFORMATION