## [Main Figures](https://docs.google.com/presentation/d/130-VPTGgYYqXV4SAqcZnxUvOVBqUOO4rmQOiiPjbCPY/edit?ts=585be6a0)

[**Extended Data Figures**](https://docs.google.com/presentation/d/1xx2lYoBBmfvunbOc9jvtWb1XrbHYg4XjQzPebyNp9HE/edit#slide=id.g1d513439fd_0_6)

[**Supplementary Tables**](https://drive.google.com/open?id=0B07orkTYRj9pWUZfQ1pKQ2R3R1U)

#### [**Supplementary Materials**](https://docs.google.com/document/d/1XcP3joT-P-4GqSkUz_nGktYZj3QSvBHiEFRS-gxGTOo/edit)

#### [**SCREEN**](http://zlab-annotations-v4.umassmed.edu)**: a visualizer for candidate Regulatory Elements in the Registry**

## [Experiments used for the Registry of cREs](https://docs.google.com/spreadsheets/d/1fP7-v4tnEljXCi-JxPyJyLXd9i7s24UKJYryzwI-X8g/edit) (goes into SCREEN)

## Human DNase Master Peaks at DCC: <https://www.encodeproject.org/annotations/ENCSR715GEC/>

## Mouse DNase Master Peaks at DCC: <https://www.encodeproject.org/annotations/ENCSR893GGE/>

## [Experiments used for mouse ChromHMM](https://docs.google.com/spreadsheets/d/1fw6xn9XvXctubya1luYkDAhM_QphTspyV4-Owo7zoIM/edit#gid=0) (part of the integrative level of Encyclopedia)

##

# TITLE

The ENCODE Encyclopedia for Human and Mouse

# ABSTRACT

The genomes of many humans have been sequenced, yet we still lack comprehensive maps of functional elements in the human genome, and we do not understand fully how they specify the myriad cell and tissue types throughout the life cycle. Such information is critical to assess how genomic variants affect development, aging, and susceptibility to diseases. The Encyclopedia of DNA Elements (ENCODE) project (<http://www.encodeproject.org>) is mapping these elements to elucidate the regulatory landscapes of the human genome. Phase III of the project extended its breadth and depth with new assays and an emphasis on primary cells and tissues. New components include maps of the elements recognized by RNA-binding proteins as well as comprehensive epigenome maps in a dozen tissues through eight time-points of mouse embryogenesis. Here we summarize Phase III data production and introduce the ENCODE Encyclopedia, an evolving collection of summary annotations derived from ENCODE data via assay-specific as well as integrative analyses. At the heart of the Encyclopedia is the Registry of candidate Regulatory Elements (cREs), annotated with chromatin accessibility, histone modifications and transcription factor occupancy. The Registry covers hundreds of human and mouse cell types, including primary tissues from human donors and multiple developmental stages in mouse. The Registry currently contains 1.31 M human and 0.43 M mouse cREs and is estimated to include over two-thirds of all cREs in human. Each cRE cross references its underlying data and other annotations in the Encyclopedia. The regulatory landscapes characterized by cREs recapitulate current understanding of cell lineages and developmental progression. Aided by its visualization engine SCREEN (<http://screen.umassmed.edu>), the Registry is a broadly useful resource for investigating the functions of noncoding DNA and deciphering the impacts of noncoding variants.

#

# 1. INTRODUCTION

The genome contains the blueprint for an organism’s development, behavior, and function. Deciphering genomes, particularly the vast non-protein-coding regions, is an ongoing challenge with far-reaching impacts on understanding biological mechanisms and improving human health. This challenge is the impetus for many research labs and organized consortium efforts. Among these efforts is the Encyclopedia of DNA Elements (ENCODE) Project, launched by the National Human Genome Research Institute (NHGRI) in 2003. The ENCODE Consortium develops and integrates high-throughput and cost-effective experimental technologies and computational approaches to catalog candidate functional elements in the human genome, including transcripts and their regulatory elements. The first, pilot phase of ENCODE focused on 44 carefully-chosen regions covering 1% of the human genome using array-based techniques {ENCODEProjectConsortium:2007fu}. ENCODE Phase II used deep-sequencing-based biochemical assays to interrogate the entire human genome, producing 1,640 datasets; integrative analyses of these datasets annotated an extensive set of candidate functional elements {ENCODEProjectConsortium:2012gc}. The Mouse ENCODE {Yue:2014gc} and modENCODE projects {Gerstein:2010bu, modENCODEConsortium:2010fd} performed thousands of genome-wide experiments on the model organisms mouse, fly, and worm. Complementary projects, including the NIH Epigenomics Project {RoadmapEpigenomicsConsortium:2015gq} and the International Human Epigenome Consortium have produced thousands of epigenomic maps for human cells and tissues {Stunnenberg:2016js}.

Despite the many efforts, much of the human and mouse genomes remain only partially annotated, both broadly across different cell types and narrowly within individual cell types. Accordingly, we have a limited understanding of the diversity of transcripts and the elements regulating them in the cellular context. To address these limitations, data collection for ENCODE Phase III expanded both in breadth and depth—assaying new cell and tissue types and studying new chromatin features and regulatory factors. ENCODE members have reported biological findings throughout the past five years based on new data generated in Phase III {ref. Consortium publication 2011-2016}.

The overarching goal of ENCODE is to provide an integrated resource to aid the scientific community in studying mammalian biology and human diseases. All ENCODE data are submitted to the data coordination center, reviewed for quality, and released immediately to the scientific community via the freely accessible ENCODE web portal (<http://www.encodeproject.org>). Additionally, we provide an ENCODE Encyclopedia of candidate and confirmed functional elements 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 applications.

# 2. SUMMARY OF ENCODE PHASE III DATA PRODUCTION

The ENCODE Consortium has produced large amounts of data on three aspects of genome activity—transcriptomes, DNA-based regulatory elements for transcription and replication, and RNA-based elements for post-transcriptional regulation. The efforts during Phase III have greatly expanded the number of experiments in each category, such that in total, Phase III produced and released 3,797 human and 1,106 mouse datasets (**Fig. 1**). These datasets are summarized by categories in **Table 1**. This work followed guidelines that ensure high-quality data production {[ref. experimental guidelines](https://www.encodeproject.org/about/experiment-guidelines/#guideline)}. The data, metadata, and analysis results are available through the ENCODE portal (<http://www.encodeproject.org>). In this section, we summarize highlights of new assays introduced during Phase III.

We expanded the analysis of transcriptomes to include primary cells from different body locations and different embryological origins. Single-cell long-RNA-seq was further developed for laser captured dissections of human brain tissues. To better define full-length transcripts, we analyzed captured RNAs using long-read sequencing. This effort, in collaboration with the GENCODE project, improved the annotations of gene and transcript structures in 14,667 human and 8,708 mouse long non-coding RNAs.

A 5´-complete cDNA sequencing assay called RAMPAGE was developed during Phase III to quantify gene expression, identify promoter locations, and assign 5´ capped termini to their corresponding RNA isoforms {Batut:2013kc}. RAMPAGE provides single-nucleotide resolution, and it is more accurate than RNA-seq for quantifying expression—both advances enable it to improve gene annotations. For example, the gene *ARHGAP23*, which encodes Rho GTPase Activating Protein 23 has 12 GENCODE-annotated transcripts, has 11 different transcription start sites (TSSs). RAMPAGE data revealed a novel TSS in testis (**Fig. 2a**), 9.2 kb upstream of the nearest annotated TSS, and another novel TSS in exon 7 specific to the spleen (**Fig. 2b**). As another example, different transcription start site (TSS), 824 bp apart, are annotated by GENCODE V26 and UCSC for *EP300*, which encodes a widely studied histone acetyltransferase involved in enhancer activation. RAMPAGE data across six cell and tissue types showed that while both TSSs are active, one of the TSSs is utilized far more frequently ([**Extended Data Fig. 1**](https://docs.google.com/presentation/d/1xx2lYoBBmfvunbOc9jvtWb1XrbHYg4XjQzPebyNp9HE/edit#slide=id.g1d513439fd_0_9)).

The coverage of DNA regulatory elements for regulation of transcription was greatly expanded during ENCODE Phase III. Many new DNase accessibility maps were completed, including deeply sequenced DNase-seq datasets on hundreds of cell and tissue samples, which facilitate 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 scores of human and mouse tissues and primary cells. We expanded the application of ChIP-seq to map the locations of modified histones, histone variants, and 33 chromatin regulators and modifiers in a carefully selected collection of five human cell lines—K562, H1, GM12878, HepG2, and A549. ChIP-seq experiments have been completed for a total of 493 TFs were assayed in at least one cell type in Phase III (549 TFs in Phases II and III combined). We used both TF-specific antibodies as well as endogenously expressed, epitope-tagged TFs created through BAC transfections or CRISPR/Cas9 genome editing. Not only do we provide quality metrics for all datasets via the ENCODE Portal (<https://www.encodeproject.org/antibodies/>), but, perhaps uniquely to ENCODE, we also provide detailed information regarding the antibodies used in our experiments at the ENCODE Portal to help users evaluate and use the data most effectively.

DNA replication timing was also studied because it provides insights into gene regulation and spatiotemporal genome compartmentalization {RiveraMulia:2016dp}. Replication timing was measured during the fate commitment of human embryonic stem cells, producing 84 datasets of 26 cell types representing the embryonic layers endoderm, mesoderm, ectoderm, and neural crest {RiveraMulia:2015er} ([**Extended Data Fig. 2**](https://docs.google.com/presentation/d/1xx2lYoBBmfvunbOc9jvtWb1XrbHYg4XjQzPebyNp9HE/edit#slide=id.g1f56f8b5e5_0_305)). Because replication timing differs across cell types, we expected that clustering of these replication timing datasets would recapitulate their developmental lineages. Indeed this was observed ([**Extended Data Fig. 3**](https://docs.google.com/presentation/d/1xx2lYoBBmfvunbOc9jvtWb1XrbHYg4XjQzPebyNp9HE/edit#slide=id.g1f56f8b5e5_0_309)), providing support for the high quality of the data.

The mouse component of ENCODE Phase III was tightly organized around a developmental progression, specifically eight distinct stages from embryonic day 10.5 (e10.5) to postnatal day 0 (p0), with 6-12 tissues collected per stage. A battery of RNA-seq assays, interrogating both long and short RNAs, ChIP-seq for eight histone modifications, ATAC-seq, whole-genome bisulfite sequencing, and ChIP-seq for three TFs were performed for the samples of the mouse embryonic developmental series.

A major new project in ENCODE Phase III was the effort to comprehensively characterize the functional RNA elements bound by RNA-binding proteins (RBPs). Four types of related and integrated datasets were generated: RIP-seq and enhanced UV crosslinking and immunoprecipitation of RBPs followed by sequencing (eCLIP-seq) {vanNostrand:2016km} were performed to identify in vivo bound RNAs and to pinpoint the portions of these RNAs involved in binding interactions; RNA-seq was performed on cells depleted of RBPs by shRNA or CRISPR; RNA Bind-N-Seq {Lambert:2014jm} was performed to determine the relative binding affinity of RBPs in vitro for all possible RNA sequences; and subcellular localization of RBPs was characterized by immunostaining. The assayed RBPs are listed in [**Supplementary Table 1**](https://docs.google.com/spreadsheets/d/1k3Olfc32DXynfHCOzBdfGMCT3oQ2gGR2I67NJki2O8E/edit#gid=0) and immunostaining images are accessible at <http://rnabiology.ircm.qc.ca/RBPImage/>.

This breadth of RBP data enables integrative analyses that relate genetic variation to RBP regulation. For example, intron 66 of *UTRN* (dystrophin-related protein 1) harbors an RBFOX2 eCLIP peak (**Fig. 2c**) downstream of an alternatively spliced exon, which overlaps an variant identified by the Exome Aggregation Consortium (ExAC) {Lek:2016bi}. This variant (G→C) disrupts the RBFOX2 binding motif (GCAUG) at the first position. RNA Bind-N-Seq data reveal that this variant substantially changes the RBFOX2 binding site—the top 5mer has an enrichment value of 13.58 for the major allele, but 0.89 for the G→C variant (**Fig. 2d**), predicting that the mutation will disrupt RBFOX2 binding in vivo. To determine whether the disruption in RBFOX2 binding would lead to a shift in spliced mRNAs, we performed RNA-seq on HepG2 cells after knocking down *RBFOX2*. In wild-type cells, the upstream exon is included in 87% of messages, while the inclusion is reduced to 28% in the *RBFOX2* knockdown cells (**Fig. 2c**). In conclusion, by integrative analysis of the RBP ENCODE data, we obtained strong evidence that this G→C SNP disrupts RBFOX2 binding and leads to reduced inclusion of the upstream exon in over half of *UTRN* messages, resulting in an altered composition of protein isoforms.

# 3. THE ENCODE PORTAL

The ENCODE portal (<https://www.encodeproject.org>) is the primary interface for retrieving all ENCODE data, metadata, data standards, and experimental protocols {Sloan:2015hy}. The Consortium has designed and collected an extensive array of metadata that describes how ENCODE experiments are performed reproducibly, processed uniformly, and connected in common biological themes {Hong:2016ke}. We define an experiment as the application of a genomic assay (ChIP-seq, RNA-seq, DNase-seq, etc.) to a particular biosample type (tissue, a cell line, primary cells, stem cells, etc.). An experiment typically comprises two biological replicates. A released experiment includes the “raw” sequencing data (typically Fastq files) and all consumable output files (alignment files, signal files, peak files, etc.) from the uniform processing pipelines described in the next section.

During ENCODE Phase III, the portal was completely redesigned to enhance data access and improve metadata comprehension. The homepage now presents up-to-date summaries of numbers and types of experiments, with intuitive links to accessing the data. Experiments are annotated by key features (facets) so that the user can easily find experiments through a faceted search. A matrix view succinctly displays the search result (<https://www.encodeproject.org/matrix/?type=Experiment>; **Fig. 1**), switchable to list and table views. Entries in the matrix are hyperlinked to the underlying datasets, along with metadata and quality metrics. The Portal is the primary mode for accessing the foundational experiments of the Encyclopedia [Add figure with the levels of the Encyclopedia here], and it also provides entry to the ground and integrative levels of the ENCODE Encyclopedia (described in a later section). All data files with their metadata can be downloaded, and all are accessible via an application program interface (API). Metadata can also be retrieved via a RESTful JSON API.

# 4. UNIFORM DATA PROCESSING AND DATA QUALITY CONTROL

We have developed uniform processing pipelines for RNA-seq, DNase-seq, ATAC-seq, TF ChIP-seq, histone mark ChIP-seq, and WGBS data (**Supplementary Methods**). These pipelines are implemented in the DNAnexus cloud computing environment and are freely available on GitHub (<https://github.com/ENCODE-DCC>). We track the dependency, or provenance, of each derived or processed file via a graph describing the input files, genome references, and specific versions of software packages and parameter values used in every step {Sloan:2015hy}. Both the graph and the intermediate results of these pipelines are available at the ENCODE Portal. Three additional data types—eCLIP-seq, Hi-C, and ChIA-PET—were processed by the respective data production labs and the analysis results submitted to the ENCODE Portal.

Each ENCODE dataset is required to have two biological replicates and exceptions, typically resulting from the lack of cell or tissue samples that can serve as replicates, are noted. We developed quality control (QC) metrics for each data type (<https://www.encodeproject.org/data-standards/>), established thresholds for these metrics as the quality standards, and integrated the calculation of the metrics into the respective uniform processing pipelines. For example, the TF ChIP-seq pipeline calculates mapping statistics, library complexity, cross-correlation between signals (number of mapped reads per position) in the two strands of the genomic DNA, correlation between biological replicates, enrichment of reads in peaks (genomic regions with significantly high signals), and agreement between peaks called in the two biological replicates {Li:2011kd}. The ENCODE Portal displays the QC metrics for each dataset. Data sets that did not meet the quality standards were replaced with new experiments, and low-coverage data sets were augmented with additional sequencing whenever possible. If it is not feasible to meet the quality standards (often due to limited experimental materials), a dataset is still released if deemed valuable to the community, along with an audit flag stating the QC metrics that it did not meet.

# 5. THE ENCODE ENCYCLOPEDIA

The raw data described above and maps of signals across the human and mouse genomes have been and continue to be highly valuable for interrogating genome function. During ENCODE Phase III, we have also derived concise, comprehensive summaries of key aspects of the raw data. Such summaries are particularly valuable to users, and we have organized them into the ENCODE Encyclopedia. The Encyclopedia has two levels of annotations (**Fig. 3a**). The ground level includes peaks and quantifications produced by the uniform processing pipelines for individual data types, and the integrative level contains annotations resulting from the combined analyses across multiple data types and with ground level annotations.

## The Ground Level of the Encyclopedia

The ground level of the Encyclopedia currently has seven components. (1) The gene and TSS expression component gives quantitative estimates of the abundance of the various types of RNA molecules in each of the assayed cell types based on ENCODE RNA-seq and RAMPAGE data. These estimates are provided at the gene and TSS levels for GENCODE-annotated genes, plus activity levels for novel TSSs identified by RAMPAGE. Gene or TSS expression profiles across cell types can be visualized using the SCREEN tool described below ([**Extended Data Fig. 4**](https://docs.google.com/presentation/d/1xx2lYoBBmfvunbOc9jvtWb1XrbHYg4XjQzPebyNp9HE/edit#slide=id.g2041a09684_0_0)). (2) The transcription factor binding component provides TF peaks, which are genomic regions significantly enriched in ChIP-seq reads. The peaks are further characterized by enriched sequence motifs (using the MEME-ChIP tool {Machanick:2011ge}) and average histone mark ChIP signals and nucleosome occupancy signal surrounding TF peaks in the matching cell type. The TF peaks and associated information can be viewed in the wiki-style web resource Factorbook (<http://factorbook.org>; {Wang:2012dk, Wang:2013fp}). (3) Locations of histone marks and histone variants are provided as histone peaks, which are regions of the genome significantly enriched in histone ChIP-seq reads. (4) The chromatin accessibility component provides DNase hypersensitive sites (DHSs)—genomic regions significantly enriched in DNase-seq reads—and their constituent DNase peaks. (5) The component for chromatin domains and compartments provides topologically associated domains (TADs) and A/B compartments that are called from Hi-C data. (6) 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. (7) 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. The eCLIP protocol and CLIPper pipeline take into account variations in transcript abundance and processing {vanNostrand:2016km}.

New data are processed and added to the ground level of the Encyclopedia as soon as they are available. Thus the ground level is continually updated ("live"), and updates do not constitute new versions. More components will be added as additional analysis pipelines are developed and existing pipelines improved.

## The Integrative Level of the Encyclopedia

We have pursued two major approaches to integration of the ENCODE data. One is unsupervised learning to identify chromatin states and assign genomic DNA segments to those states across cell types. The second approach is the supervised assignment of DNA segments to candidate functional categories based on a set of epigenetic features. Results from the former approach are described in this section; the latter approach led to the Registry of Candidate Regulatory Elements, which is described in the next section.

Comprehensive catalogs of candidate transcriptional regulatory elements defined using epigenetic signals across a large collection of cell and tissue types (henceforth abbreviated as cell types) not only can summarize the regulatory repertoire of a genome but also can delineate the realization of the repertoire in individual cell types. As described above, the ENCODE and Roadmap Epigenomics Consortia have measured epigenetic signals in hundreds of human and mouse cell types. Although the coverage of cell types by assays is broad, it is not complete. The incomplete assay coverage has motivated two schools of approaches to building catalogs of candidate transcriptional regulatory elements. Sophisticated computational models such as ChromHMM {Ernst:2010bh, Ernst:2012ii} and Segway {Hoffman:2012gn}, which maximize the types of epigenetic signals that can be integrated at the cost of cell type coverage, represent the first school of approaches; these were initiated by the Kellis and Noble labs during ENCODE Phase II when a large portion of experiments was performed on immortalized cell lines and their assay coverage was complete. ENCODE Phase III has substantially increased the number of experiments on primary cells and tissues, and their limited quantities have resulted in incomplete assay coverage, calling for the second school of approaches which utilize a small number of epigenetic signals to maximize cell type coverage. For the rest of this section, we describe extensions of ChromHMM and Segway to new cell types in ENCODE Phase III. The next section will describe our development of the Registry of candidate Regulatory Elements, which falls in the second school of approaches.

ChromHMM {Ernst:2010bh, Ernst:2012ii} and Segway {Hoffman:2012gn} are probabilistic, unsupervised models that integrate a specified number of epigenetic signals to define a complete set of chromatin states, e.g., promoters, enhancers, heterochromatin, etc. They have been applied to most ENCODE and Roadmap cell types with sufficient assay coverage and the resulting chromatin states are included in the integrative level of the Encyclopedia. We applied ChromHMM to the mouse embryonic developmental series, with 66 complete epigenomes each assayed by ChIP-seq of eight histone marks (H3K4me1, H3K4me2, H3K4me3, H3K27ac, H3K9ac, H3K36me3, H3K27me3, and H3K9me3) and RNA-seq, and defined 15 chromatin states using the histone marks {ENC101}. Both ChromHMM and Segway have also been augmented to include cell types with missing assays; for example, during the Roadmap Epigenomics project, ChromImpute was developed to impute missing epigenetic signals in a subset of cell types {Ernst:2015ep, RoadmapEpigenomicsConsortium:2015gq}, and during ENCODE Phase III, Noble and colleagues developed a strategy to train separate unsupervised annotation models on each cell type—allowing different assay coverages in different cell types—and then automatically interpret these results across all cell types using a Random Forests classifier. They annotated chromatin states of 164 human cell types using 1,615 genomics data sets {Libbrecht et al., bioRxiv 086025}.

# 6. THE REGISTRY OF CANDIDATE REGULATORY ELEMENTS

ENCODE and Roadmap data offer the opportunity to build a registry of cREs that may function in a broad range of tissues and cell types. As explained in this section, we identified cREs using a key set of epigenetic signals—chromatin accessibility, histone modifications H3K4me3 and H3K27ac, and occupancy by the insulator-binding protein CTCF. The first release of the Registry includes 1.31 M human cREs and 0.43 M mouse cREs; future versions will be released periodically. Based on the levels of these epigenetic signals and the nearest distance to an annotated TSS, we classify cREs as having promoter-like or enhancer-like signatures, or without these signatures but bound by the insulator-binding protein CTCF.

## The overall design principle of the Registry of cREs

The most direct approach to identifying cREs would be to build a comprehensive statistical model that integrates all relevant epigenetic signals and is trained using well-annotated regulatory elements. Indeed, such methods exist {Rajagopal:2013jg, Erwin:2014fo}; however, only a limited number of functional validated regulatory elements are available, and a complex model trained using insufficient data would be prone to overfitting. We evaluated the effectiveness of complex methods from laboratories in the ENCODE Consortium in the form of "blind tests," i.e., soliciting genome-wide predictions of enhancer active during mouse development and then performing mouse transgenic assays on the predicted regions (M. Gerstein et al., manuscript in preparation). None of the complex models consistently outperformed the simple approach we will describe below. While we acknowledge the potential for considerable improvement in the success of complex methods with more training data and advances in modeling, we felt that it was premature to commit to any one of these methods for our initial version of the Registry of cREs.

After rigorous evaluation using experimental data (described below), we decided to take an alternative approach to identifying cREs based on the four types of epigenetic signals which are most predictive of regulatory elements. Each cRE is centered on a DNA segment with high chromatin accessibility (measured by DNase-seq), which is well-known to delineate all of the main classes of cis-regulatory elements in a cell-type-specific manner, including promoters, enhancers, and insulators {Gross:1988bu, Thurman:2012fe}. We categorized the DHSs by overlap with histone modifications associated with promoters and enhancers, H3K4me3 and H3K27ac, respectively {Heintzman:2007ke, Visel:2009jp}, or with CTCF, the only known insulator binding protein in mammals {Kim:2007bm}. This simple model applies to all cell types with at least one of these data types, covering 301 human cell types (620 with primary cells or tissues from different donors counted separately) and 58 mouse cell types (138 with developmental time-points counted separately) when all ENCODE and Roadmap data are considered.

## DNase and H3K27ac are the best single features for predicting tissue-specific enhancers

We used mouse embryonic enhancers in the VISTA database {Visel:2007jw} to compare the effectiveness of ten types of epigenetic signals in predicting enhancers: DNase hypersensitivity, eight histone marks (H3K4me1, H3K4me2, H3K4me3, H3K27ac, H3K9ac, H3K9me3, H3K36me3, and H3K27me3), and DNA methylation. VISTA enhancers have been tested for enhancer function in vivo using mouse transgenic assays {Pennacchio:2006cy}. At the time of our evaluation (2015), over 2100 TSS-distal regions in the human and mouse genomes had been tested for reporter gene expression at embryonic day 11.5 (e11.5), and the results are available via the VISTA Enhancer database (https://enhancer.lbl.gov/). Each region was tested for enhancer activities in all mouse tissues at e11.5.

The ENCODE Phase III data across the mouse developmental series are an ideal source of epigenetic data for evaluating enhancer prediction since they are from the same tissues and stage of development assayed for reporter gene expression by mouse transgenic assays. Four tissues at e11.5—the midbrain, hindbrain, neural tube, and limb—were covered by all epigenetic assays (DNase-seq, the eight histone marks, and DNA methylation), and there were hundreds of VISTA regions active in each of these tissues. We asked which of the ten epigenetic signals were predictive of VISTA enhancers in each tissue (details of the comparison are in **Supplementary Methods**). Our positive test set comprised all VISTA enhancers in that tissue; our negative set contained the remaining VISTA regions (i.e., tested but showed no activity in that tissue). The VISTA regions used in our analysis are listed in [**Supplementary Table 2**](https://docs.google.com/spreadsheets/d/1lLQOT5m13M3CW_fuZvlIBFYhsGb9ifxPiPFdPLkL0go/edit?usp=sharing). The average DNase signal and the average H3K27ac signal in a window centered at their respective peaks were consistently the most predictive single features for enhancer activity. In most tissues, the DNase signal was the most predictive feature, and in the midbrain, H3K27ac was slightly better than DNase ([**Extended Data Fig. 5**](https://docs.google.com/presentation/d/1xx2lYoBBmfvunbOc9jvtWb1XrbHYg4XjQzPebyNp9HE/edit#slide=id.g1f4632a01f_0_0); [**Supplementary Table 3**](https://docs.google.com/spreadsheets/d/10IPKf8MatKovGNd4zWP2XIkVONGwI2YDBDgQuqmCAjs/edit#gid=1763487930)).

DNase offers high spatial precision in defining regulatory elements: DHSs are ~300 bp long and often correspond to the core of regulatory elements. In contrast, the H3K27ac signal is more diffuse: it tends to be low at the center of a regulatory element—due to the lack of a nucleosome—but is high right before and after, where two flanking nucleosomes reside. Defining enhancer predictions as windows centered on DHSs, we tested the performance of DNase, histone marks, and DNA methylation signals averaged within each window. On average, ranking using the DNase signal slightly outperformed ranking using the H3K27ac signal ([**Extended Data Fig 6**](https://docs.google.com/presentation/d/1xx2lYoBBmfvunbOc9jvtWb1XrbHYg4XjQzPebyNp9HE/edit#slide=id.g1f56f8b5e5_0_313)**;** [**Supplementary Table 3**](https://docs.google.com/spreadsheets/d/10IPKf8MatKovGNd4zWP2XIkVONGwI2YDBDgQuqmCAjs/edit#gid=1763487930)). Signals for other histone marks and DNA methylation individually were far less predictive ([**Supplementary Table 3**](https://docs.google.com/spreadsheets/d/10IPKf8MatKovGNd4zWP2XIkVONGwI2YDBDgQuqmCAjs/edit#gid=1763487930)). The average rank of DNase and H3K27ac signals was slightly better than DNase. Incorporating additional histone marks or DNA methylation using a linear model did not improve the performance further. We did not test more complex models because of the small number of VISTA enhancers—only 200-300 genomic regions were tested positive in each tissue.

## Combining DNase and H3K27ac signals accurately predicts new tissue-specific enhancers

To further test the simple model of combining DNase and H3K27ac, we used the average rank of DNase and H3K27ac signals to predict new TSS-distal (>2kb from the nearest TSS), candidate enhancers in the hindbrain, midbrain, and limb ([**Extended Data Fig. 7**](https://docs.google.com/presentation/d/1xx2lYoBBmfvunbOc9jvtWb1XrbHYg4XjQzPebyNp9HE/edit#slide=id.g1f72a930f5_1_8)). The boundaries for these regions were defined using H3K27ac ChIP-seq peaks (**Fig. 3a**). For each tissue, we used mouse transgenic assays to test 20, 15 and 15 new regions around the ranks 1-20, 1500-1520, and 3000-3200, respectively. In total, we tested 151 regions, and the results are in [**Supplementary Table 4**](https://docs.google.com/spreadsheets/d/1V1__9vi4noCl909Co2oRKXzusksOIEQdphpBFxl-Y9k/edit?usp=sharing). Consistently, higher ranking regions were more likely to show enhancer activity in their predicted tissue compared with lower ranking regions (**Fig. 3b**; e.g., 75%, 26.6%, and 20% for hindbrain). When enhancers were active in multiple tissues, these tissues also had H3K27ac signal across the predicted enhancer region (**Fig. 3c-e**). For example, a predicted enhancer in the hindbrain was also active in midbrain and neural tube; accordingly, high H3K27ac signals were observed in all three tissues (**Fig. 3d**). In contrast, an enhancer active almost exclusively in the limb (**Fig. 3e**) did not show high H3K27ac signals in other tissues assayed.These results suggest that combining DNase and H3K27ac not only can identify active enhancers in a particular tissue but also can quantify the tissue selectivity of these enhancers.

## Combining DNase and H3K4me3 signals accurately predicts active promoters

We further evaluated whether an adaptation of the above-described enhancer prediction model could perform well for predicting cell-type-specific promoters, using transcript expression levels measured by RNA-seq in the e11.5 midbrain, limb, neural tube, and hindbrain. The single most predictive feature was H3K4me3 signal, which when averaged over a ±1.5 kb window centered on TSS-proximal DHSs, correlated with expression levels at *r* = 0.75 averaged over the four tissues ([**Supplementary Table 5**](https://docs.google.com/spreadsheets/d/1dXorMI_a0EJ7_a3yTqU78Ahs0ld1Ic13dHpeYuSU6Eg/edit#gid=1761182185); [**Extended Data Fig. 8**](https://docs.google.com/presentation/d/1xx2lYoBBmfvunbOc9jvtWb1XrbHYg4XjQzPebyNp9HE/edit#slide=id.g1f56f8b5e5_0_313) for hindbrain). This correlation is substantially higher than that of the H3K4me3 signal centered on H3K4me3 peaks (*r* = 0.57) or the DNase signal centered on TSS-proximal DHSs (*r* = 0.39). Repeating this analysis using human RNA-seq data in GM12878, K562, and HepG2 yielded consistent results (*r* = 0.72, 0.73, 0.71). In conclusion, the high spatial precision offered by DHSs improves the accuracy of H3K4me3 for predicting gene expression.

## Selection of cREs for the Registry

These evaluations showed that combining DNase with two histone marks H3K4me3 and H3K27ac was an effective way of building the Registry of candidate promoters and enhancers active in specific cell types. We further extended our model by adding CTCF, a highly conserved architectural protein that binds to most insulators and contributes to the establishment of the three-dimensional chromatin structure {Ong:2014ku}.

The cREs in the registry derive from the DHSs that are supported by at least one additional type of epigenetic signal among H3K4me3, H3K27ac, or CTCF. Because DHSs were identified in individual samples, we first condensed them into a set of non-overlapping representative DHSs (rDHSs) using the following iterative process (**Fig. 4a**). 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 repeated until it finally resulted in a list of non-overlapping rDHSs representing all DHSs. We then filtered out rDHSs with Z-scores less than 1.64—a threshold corresponding to the 95th percentile of a one-tailed test. Roughly 1.6 M rDHSs in human and 0.63 M rDHSs in the mouse remained. The subset of rDHSs with support from at least one additional type of epigenetic signal are designated cREs, i.e., the rDHSs that are also cREs have high H3K4me3, H3K27ac, or CTCF-binding signals (high signal is defined as Z-score>1.64 throughout) in at least one cell type. In total, there are 1,310,152 human cREs and 431,202 mouse cREs (**Fig. 4a**). Among them, 724,590 human cREs and 228,027 mouse cREs have high DNase and high H3K4me3, H3K27ac, or CTCF in the same cell type, and these cREs are further designated as having "concordant" support.

cREs are further designated as TSS-proximal if they lie within 2 kb of a TSS. There are 242,739 such cREs in human and 92,405 in mouse. Among these, cREs overlapping TSSs are called TSS-containing cREs; there are 46,749 and 24,549 TSS-overlapping cREs in human and mouse respectively. TSS-overlapping cREs are significantly longer than the rest of the TSS-proximal cREs and TSS-distal cREs ([**Extended Data Fig. 9**](https://docs.google.com/presentation/d/1xx2lYoBBmfvunbOc9jvtWb1XrbHYg4XjQzPebyNp9HE/edit#slide=id.g2041a09684_0_12), median length = 548, 317, 342 for human and 589, 320, 339 for mouse; Wilcoxon test p-values < 2.2e-16 for all tests).

## Comprehensiveness of the current Registry of cREs

Defining the registry of cREs based on rDHSs, we have assumed that the existing collection of rDHSs distilled from hundreds of DNase-seq experiments can represent cREs in a majority of cell types, and a new cell type is likely to use a subset of the cREs in the registry as its regulatory landscape. To test this hypothesis, we set out to analyze 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 current version of the Registry of cREs. For human, 67% (121,692/181,177) of all annotated TSSs and 72% (105,196/145,671) of the TSSs of protein-coding genes overlapped a cRE in the Registry. For mouse, 61% (57,459/93,719) of all annotated TSSs and 66% (52,066/78,782) of the TSSs of protein-coding genes overlapped a cRE in the Registry.

Second, we analyzed how rapidly the total number of unique rDHSs saturated with more and more cell types. In ENCODE Phase II, we modeled DHS saturation using a Weibull distribution and estimated that we had discovered around half of the total DHSs. We performed this analysis using all human data generated by ENCODE Phase III and Roadmap projects. The saturation curves of rDHSs also follow Weibull distributions, revealing 1.66 M rDHSs with FDR<0.1% and Z-score>1.64 at plateau. Because only a subset of such rDHSs can be cREs—those with at least one max-Z > 1.64—we have identified at least 78.9% cREs in human ([**Extended Data Fig. 10**](https://docs.google.com/presentation/d/1xx2lYoBBmfvunbOc9jvtWb1XrbHYg4XjQzPebyNp9HE/edit#slide=id.g1b03e2385d_0_8)). We performed the same saturation analysis for mouse but could not reach a reliable estimate due to the small number of cell types—we obtained 0.62 M cREs at the plateau when we treated the same tissue at different embryonic time-points separately but 2.41 M cREs when we kept only one time-point for each tissue type.

Third, we computed the Registry's coverages of H3K27ac, H3K4me3, and CTCF peaks (FDR<0.01) of cell types with the corresponding ChIP-seq data but without DNase-seq data ([**Supplementary Table 6**](https://drive.google.com/open?id=1Eam70P0ZHsoNIhW62sDy1Zui1kMuP9gHweDHa8TK1vE)). 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. 11**](https://docs.google.com/presentation/d/1xx2lYoBBmfvunbOc9jvtWb1XrbHYg4XjQzPebyNp9HE/edit#slide=id.g1f578edfd6_5_19)). 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. 12**](https://docs.google.com/presentation/d/1xx2lYoBBmfvunbOc9jvtWb1XrbHYg4XjQzPebyNp9HE/edit#slide=id.g1f578edfd6_5_28)). (There were no cell types with CTCF but without DNase data for mouse.) The coverages for H3K4me3 peaks were low for several human and mouse cell types. The average -log(FDR) of the H3K4me3 peaks in these datasets were low ([**Extended Data Fig. 13**](https://docs.google.com/presentation/d/1xx2lYoBBmfvunbOc9jvtWb1XrbHYg4XjQzPebyNp9HE/edit#slide=id.g1f578edfd6_5_37)). 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 positives by the peak calling algorithm.

## Classification of cREs in the Registry

Gene catalogs 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 cREs based on the maximal Z-scores across all cell types with ENCODE and Roadmap data, abbreviated as max-Z henceforth (**Fig. 4a**; [**Extended Data Fig. 14**](https://docs.google.com/presentation/d/1xx2lYoBBmfvunbOc9jvtWb1XrbHYg4XjQzPebyNp9HE/edit#slide=id.g1f578edfd6_5_45)). As described above, all cREs must have a DNase max-Z greater than 1.64, and furthermore must have a max-Z greater than 1.64 for least one of three epigenetic signals—H3K4me3, H3K27ac, or CTCF. Since each of the three epigenetic max-Z values can either be above (≥ 1.64) or below the threshold, a cRE can adopt one of eight states. Furthermore, each cRE is classified as being proximal (≤ 2 kb) or distal (> 2 kb) to the nearest GENCODE annotated TSS. We further combine these sixteen states into three broader mutually-exclusive groups, assigned in the following order:

1. cREs with promoter-like signatures (cRE-PLS) must have high H3K4me3 max-Zs. If they are TSS-distal, they must also have low H3K27ac max-Zs.
2. cREs with enhancer-like signatures (cRE-ELS) must have high H3K27ac max-Zs. If they are TSS-proximal, they must also have low H3K4me3 max-Zs.
3. CTCF-only cREs are the remaining cREs. They do not fall into either of the first two categories and thus by definition must have high CTCF max-Zs to qualify as cREs.

We analyzed the percentage of the genome covered by each group of cREs, considering only regions of the genome which are mappable by 36-nt long sequences in DNase-seq experiments (~2.65 billion bases for human and 2.29 billion bases for mouse). In total, 20.8% of the mappable genome is covered by cREs (4.2% by cREs-PLS, 15.9% by cREs-ELS, and 0.7% by CTCF-only cREs) and 8.8% of the mappable mouse genome is covered by cREs ([**Extended Data Fig. 15**](https://docs.google.com/presentation/d/1xx2lYoBBmfvunbOc9jvtWb1XrbHYg4XjQzPebyNp9HE/edit#slide=id.g1f56f8b5e5_0_4)). The lower coverage for mouse is due to the smaller number of cell types with data to define cREs.

In addition to the cell type agnostic classification described above, we evaluated the biochemical activity of each cRE in each individual cell type using the corresponding DNase, H3K4me3, H3K27ac, and CTCF data. All cREs with low DNase Z-scores in a particular cell type are bundled into one “inactive” state for that cell type; the remaining “active” cREs are divided into eight states according to their epigenetic signal Z-scores, producing nine possible states in total. The three groups described above—cRE-PLS, cRE-ELS, and CTCF-only cRE—apply to the active cREs within a particular cell type. Two additional groups are defined with respect to individual cell types: an inactive group, containing all cREs in the inactive state, and a DNase-only group, containing cREs with high DNase Z-scores but low H3K4me3, H3K27ac, and CTCF Z-scores within the cell type.

Using GM12878 lymphoblastoid cells as an example, **Fig. 4b** summarizes the cREs in the nine states, with each state further divided into TSS-proximal and TSS-distal. The bar graph reveals that cREs with high H3K4me3 are mostly TSS-proximal, regardless of whether they have high H3K27ac or CTCF, while cREs with low H3K4me3 are mostly TSS-distal.

We used additional ChIP-seq data for three factors in GM12878 to evaluate our five-group classification of cREs in this cell type (**Fig. 4c**): RNA polymerase II (POL2), which binds most active promoters, EP300, a histone acetyltransferase that binds many enhancers, and RAD21, another component of the cohesin complex which includes CTCF. The TSS-proximal cREs in the high-H3K4me3, high-H3K27ac, high-CTCF state had the highest median POL2 signal (25.0; compared with 7.3 for the second highest state; [**Extended Data Fig. 16**](https://docs.google.com/presentation/d/1xx2lYoBBmfvunbOc9jvtWb1XrbHYg4XjQzPebyNp9HE/edit#slide=id.g1f578edfd6_5_64)), yet moderately high EP300 signals (median = 10.9; [**Extended Data Fig. 17**](https://docs.google.com/presentation/d/1xx2lYoBBmfvunbOc9jvtWb1XrbHYg4XjQzPebyNp9HE/edit#slide=id.g1f578edfd6_5_102)). Some of these cREs may function as both promoters and enhancers, but collectively they are more promoter-like than enhancer-like judged by their POL2 and EP300 signals. In contrast, the high-H3K27ac, low-H3K4me3 states, regardless of CTCF status or proximity to TSS, showed the highest EP300 signals but low POL2 signals, supporting their assignment as cREs with enhancer-like signatures ([**Extended Data Fig. 18**](https://docs.google.com/presentation/d/1xx2lYoBBmfvunbOc9jvtWb1XrbHYg4XjQzPebyNp9HE/edit#slide=id.g1f72a930f5_1_12)**)**. The most challenging assignments were for the relatively few high-H3K4me3, low-H3K27ac, TSS-distal cREs (450 high-CTCF and 1,584 low-CTCF). These cREs had slightly, yet significantly higher POL2 binding than DNase-only cREs, which supported a promoter-like classification ([**Extended Data Fig. 19**](https://docs.google.com/presentation/d/1xx2lYoBBmfvunbOc9jvtWb1XrbHYg4XjQzPebyNp9HE/edit#slide=id.g1e46acddc4_0_7)). **Fig. 4c** summarizes the five-group classification of cREs in GM12878.

[**Extended Data Fig. 20-22**](https://docs.google.com/presentation/d/1xx2lYoBBmfvunbOc9jvtWb1XrbHYg4XjQzPebyNp9HE/edit#slide=id.g22d7ccb98d_0_14) show the nine-state and five-group classifications of cREs and the underlying DNase-seq, H3K4me3, H3K27ac and CTCF data for three cell types—hepatocytes, B cells, and bipolar spindle neurons. Three loci are displayed, each specifically active in one of the three cell types as indicated by RNA-seq data: hepatocyte nuclear factor 4 (*HNF4a*) ([**Extended Data Fig. 20**](https://docs.google.com/presentation/d/1xx2lYoBBmfvunbOc9jvtWb1XrbHYg4XjQzPebyNp9HE/edit#slide=id.g22d7ccb98d_0_14)) , active in hepatocytes; hematopoietic transcription factor PU.1 (*SPI1*) ([**Extended Data Fig. 21**](https://docs.google.com/presentation/d/1xx2lYoBBmfvunbOc9jvtWb1XrbHYg4XjQzPebyNp9HE/edit#slide=id.g22d7ccb98d_0_157)), active in B cells; and neuronal PAS domain protein 4 (*NPAS4*) ([**Extended Data Fig. 22**](https://docs.google.com/presentation/d/1xx2lYoBBmfvunbOc9jvtWb1XrbHYg4XjQzPebyNp9HE/edit#slide=id.g22d7ccb98d_0_240)), active in bipolar spindle neurons. Both the general, cell-type-agnostic classification of cREs and the classifications in each cell type are shown. The cREs surrounding each locus are active specifically in the corresponding cell type.

## Relative abundance of cREs-PLS vs. cREs-ELS

In GM12878, there are 36,022 cREs with promoter-like signatures, 27,739 cREs with enhancer-like signatures, 10,913 CTCF-only cREs, 16,085 DNase-only cREs, and 1,219,393 inactive cREs ([**Extended Data Fig. 23**](https://docs.google.com/presentation/d/1xx2lYoBBmfvunbOc9jvtWb1XrbHYg4XjQzPebyNp9HE/edit#slide=id.g1d566b2282_1_277)). The higher abundance of cREs-PLS over cREs-ELS may surprise the reader, given the common perception that enhancers outnumber promoters. Two factors contributed to our results. First, some of the cREs-PLS may in fact be enhancers proximal to a TSS, which we are unable to reliably identify given the 2 kb resolution of our approach. Among the 34 k TSS-proximal cREs-PLS, over 14 k directly overlap a TSS; [**Extended Data Fig. 24a**](https://docs.google.com/presentation/d/1xx2lYoBBmfvunbOc9jvtWb1XrbHYg4XjQzPebyNp9HE/edit#slide=id.g1f42c5876b_0_1) shows that 7.6 k TSSs have two or more cREs-PLS. The "extra" cREs-PLS for these TSSs are good candidates for TSS-proximal enhancers, and because the promoter and enhancer functions exist in a continuum {Nguyen:2016kt}, some of these "extra" cREs-PLS may function as both promoters and enhancers. Second, it is important to note that different cell types share many of their promoters but many fewer of their enhancers ([**Extended Data Fig. 24b**](https://docs.google.com/presentation/d/1xx2lYoBBmfvunbOc9jvtWb1XrbHYg4XjQzPebyNp9HE/edit#slide=id.g1f42c5876b_0_1)). Indeed, we define 254,880 cREs-PLS and 991,173 cREs-ELS across all available cell types.

## Comparison between cREs and the corresponding ChromHMM states

As described above, there are two schools of approaches to building catalogs of regulatory elements, with the Registry of cREs representing one and ChromHMM representing the other. We asked how the simple, rDHS-anchored, one additional support approach of defining cREs compared with the more sophisticated, hidden Markov model based approach of chromHMM which also incorporated more histone marks.

The cREs-PLS and cREs-ELS in GM12878 agree with the respective chromatin states called by ChromHMM using eight histone marks and CTCF in this cell type {ENCODEProjectConsortium:2012gc}. [**Extended Data Fig. 25a**](https://docs.google.com/presentation/d/1xx2lYoBBmfvunbOc9jvtWb1XrbHYg4XjQzPebyNp9HE/edit#slide=id.g1d513439fd_0_6) shows significant overlap between top-ranking cREs-PLS with ChromHMM promoters. [**Extended Data Fig. 25b**](https://docs.google.com/presentation/d/1xx2lYoBBmfvunbOc9jvtWb1XrbHYg4XjQzPebyNp9HE/edit#slide=id.g1d513439fd_0_6) reveals that over 85% of the top ranking cREs-ELS overlap with ChromHMM enhancers. The overlap decreases for lower ranking cREs-ELS, but the overlap with ChromHMM low-signal enhancers increases; 97% of the cREs-ELS ranked above 20,000 overlap with ChromHMM enhancers or low-signal enhancers.

We also compared the cREs for five e11.5 and six e14.5 mouse tissues with the ChromHMM states called using eight histone marks in the corresponding tissues (described in the above "The Integrative Level of the Encyclopedia" section) {ENC101}. We observed substantial agreement between cREs-PLS and ChromHMM-annotated promoters and between cREs-ELS and ChromHMM-annotated enhancers ([**Extended Data Fig. 25c, d**](https://docs.google.com/presentation/d/1xx2lYoBBmfvunbOc9jvtWb1XrbHYg4XjQzPebyNp9HE/edit#slide=id.g1d513439fd_0_6)).

## Cell and tissue types clustering by DNase or H3K27ac signal at cREs

Clustering of primary cell types according to their DHS landscapes reflects established relationships among lineages and embryonic origins {Stergachis:2013hw}. To examine whether the Registry of cREs captured the regulatory landscapes, we clustered primary cell types and tissues by the DNase or H3K27ac Z-scores at each cRE as being either high or low in the particular cell type. The dendrograms indeed recapitulate relationships among cell and tissue lineages (Supplementary Materials; [**Extended Data Fig. 26-28**](https://docs.google.com/presentation/d/1xx2lYoBBmfvunbOc9jvtWb1XrbHYg4XjQzPebyNp9HE/edit#slide=id.g1f72a930f5_1_31)). A companion paper {ref Breschi, ...Gingeras, Guigo} reports consistent results based on RNA-seq data.

# 7. SCREEN: A WEB ENGINE FOR SEARCHING AND VISUALIZING cREs

We have built a web-based tool called SCREEN (Search Candidate Regulatory Elements by ENCODE; <http://screen.umassmed.edu>) for searching and visualizing the 1.31 M human cREs and 0.43 M mouse cREs in the Registry and connecting them to the underlying ENCODE data and other ENCODE annotations. SCREEN is divided into three “apps”, each of which provides a different perspective on the cREs (**Fig. 5**). The core app is the cRE-centric search, where the user can retrieve a subset of cREs by various characteristics, including genomic coordinates and activity profile within a selected cell type (**Fig. 5a**). The search app also provides a built-in details view, which lists a cRE’s interaction with genomic features such as transcription factor binding sites, SNPs, topologically-associated domains, genes, and TSSs, as well as clips of the DNase, H3K4me3, H3K27ac, and CTCF signals at the cRE’s locus (mini-peaks) across all available cell types. The gene-centric app plots RNA-seq and RAMPAGE expression data for genes and TSSs (**Fig. 5b**), and, for mouse data, plots differentially-expressed genes and differential cRE activity across cell types and developmental time points (see the first use case below). Finally, the SNP-centric GWAS app intersects cREs with SNPs from more than 50 published GWAS studies, provides a list of the cell types with the most enriched cREs for each study, and allows the user to browse the cREs by cell type (**Fig. 5c**). All three apps have links to the UCSC genome browser, facilitating visualization of the epigenetic signals at a cRE’s or a gene’s locus (**Fig. 5a**).

# 8. USE CASES OF ENCODE ENCYCLOPEDIA

We foresee many applications for the ENCODE Encyclopedia. The various annotations at the ground level of the Encyclopedia can be downloaded from the ENCODE Portal and further analyzed along with the user’s own data. SCREEN allows the user to directly search for cREs in the Registry and explore all associated annotations. Here we provide three use cases for the Registry of cREs through SCREEN. The first use case explores the 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 cREs with promoter-like or enhancer-like signatures. One major application of the Encyclopedia is to interpret GWAS variants; the other two use cases illustrate how to characterize GWAS SNPs using the Registry of cREs.

##

## Comparing cRE activities between mouse tissues across developmental timepoints

We have performed differential gene expression analysis for all GENCODE-annotated mouse genes between all available pairs of tissues and time-points. SCREEN displays differentially expressed genes in a locus alongside the differential activities of cREs within 500 kb of the gene of interest—activity here is defined as the H3K4me3 Z-score for cREs-PLS and the H3K27ac Z-score for cREs-ELS. As an example, *Ogn* encodes osteoglycin, a protein involved in bone formation. *Ogn* exhibits a dramatic increase in expression corresponding to bone development which occurs on embryonic day 12 of mouse {Taher, 2011}. SCREEN displays *Ogn* and nearby differentially expressed genes between e11.5 and e15.5 of limb (identified using DESeq2 {Love:2014ka}, FDR < 0.01) as bars, with the heights of the bars corresponding to the log2 fold change in expression between the two time-points and the widths representing the lengths of the genes in base pairs (**Fig. 6a**). cREs-PLS and cREs-ELS are shown in the plot as red and yellow dots respectively, with the y-coordinates of the cREs designating the differences in activity Z-scores between the two time-points. This view over a large domain helps identify cREs that could account for the increase in Ogn expression—specifically, cREs proximal to Ogn are likely to play a role in the regulation, as they increase in signals concomitant with the increase in Ogn expression. The UCSC genome browser view of the *Ogn* locus across six time-points, which can be directly launched from SCREEN, reveals the change in *Ogn* expression over developmental time (**Fig. 6b**), correlated with increases in H3K27ac, H3K4me4, and DNase signals. *Ogn* expression increases most notably after e12.5, consistent with previous findings {Taher, 2011}. Intriguingly, this increase in gene expression lags behind the increases in H3K27ac and H3K4me3 Z-scores of nearby cREs (**Fig. 6c**).

## Using the registry of cREs to annotate GWAS SNPs

Previous studies have repeatedly demonstrated that the majority of GWAS variants reside outside exons and that annotation efforts in non-coding 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 epigenetics and transcription factor binding data associated with the cREs, the Registry can be particularly useful for annotating GWAS SNPs.

We have preloaded SCREEN with the subset of studies from the NHGRI-EBI GWAS catalog {Hindorff:2009cc, MacArthur:2017kz} that were performed on the Caucasian-European (CEU) population ([**Supplementary Table 9**](https://drive.google.com/open?id=1JRc8T0rMqeFDOQIFJW4abm4qIGsrR5HwmJhobzglB4A)), and we use CEU-specific data of linkage disequilibrium (LD; correlation coefficient r2>0.7) to perform statistical tests. We plan to include other populations shortly. For each GWAS, we tested each cell type for whether its cREs-ELS were significantly enriched in the GWAS SNPs after accounting for SNPs in LD. We adapted and modified the Uncovering Enrichment through Simulation (UES) method {Hayes et al. 2015} with Fisher's exact tests for performing statistical testing. SCREEN displays the cell types in the descending order of enrichment, and the user can browse the cREs in each cell type that overlap with GWAS SNPs. [**Extended Data Fig. 29**](https://docs.google.com/presentation/d/1xx2lYoBBmfvunbOc9jvtWb1XrbHYg4XjQzPebyNp9HE/edit#slide=id.g1f578edfd6_5_142) shows a heat map for the enriched cell types for a subset of GWAS, and the results are summarized in [**Extended Data Fig. 30**](https://docs.google.com/presentation/d/1xx2lYoBBmfvunbOc9jvtWb1XrbHYg4XjQzPebyNp9HE/edit#slide=id.g1f578edfd6_5_146).

The user can first select a GWAS study of interest (**Fig. 7a**), and SCREEN displays the fraction of LD blocks with at least one GWAS SNP overlapping cREs, which estimates the portion of GWAS signal that can be explained by cREs in the Registry using all available cell types (**Fig. 7b**). A list of cell and tissue types is provided based on enrichment in H3K27ac signal (see **Supplementary Methods**). The user can narrow their search by selecting a cell type, such as GM12878 for multiple sclerosis, the left ventricle tissue for QT interval, or HepG2 for cholesterol levels (**Fig. 7c**). Once a cell type is selected, SCREEN will update to show the list of cREs in that cell type overlapping the LD blocks (e.g., 473 GM12878 cREs overlap multiple sclerosis SNPs), and denote the cREs as having promoter-like or enhancer-like signatures (**Fig. 7d**). SCREEN also returns a list of SNPs for the user to search and view in a genome browser along with the cRE and other supporting data, which can help finely annotate the SNPs and predict their functional impact.

As an example, rs1250568 is in LD (r2=0.7) with two SNPs associated with multiple sclerosis, rs1250542 {Patsopoulos:2011bq} and rs1250540 {DeJager:2009bn}. rs1250568 is predicted to be a causal SNP by the deltaSVM algorithm {Lee:2015hi}. GM12878 was previously suggested to be a relevant cell type for multiple sclerosis {Maurano:2015ef}, and SCREEN computed an FDR of 2.6E-7 for the enrichment of GM12878’s cREs. rs1250568 lies in cRE EH37E0182314, which has a high H3K27ac Z-score in GM12878 (**Fig. 7d**). It overlaps a ChIP-seq peak for the transcription factor ELF1 and disrupts an ELF1 motif site (**Fig. 7e**). 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 in an intron, and *PPIF*, a downstream gene also known as Cyclophilin D. *ZMIZ1* is in the androgen receptor signaling pathway and is expressed at lower levels in patients with multiple sclerosis {Fewings:2017jw}. It is highly expressed in neurons and cardiac muscle cells ([**Extended Data Fig. 31**](https://docs.google.com/presentation/d/1xx2lYoBBmfvunbOc9jvtWb1XrbHYg4XjQzPebyNp9HE/edit#slide=id.g1f4e223a70_0_0)). *ZMIZ1* has been reported in the GWAS studies but *PPIF* has not {Patsopoulos:2011bq, DeJager:2009bn}. *PPIF* encodes a mitochondrial permeability transition pore protein, and it is expressed in cardiac muscle cells, lymphocytes and hepatocytes ([**Extended Data Fig. 32**](https://docs.google.com/presentation/d/1xx2lYoBBmfvunbOc9jvtWb1XrbHYg4XjQzPebyNp9HE/edit#slide=id.g1f4e223a70_0_6)). We predict that the functions that *PPIF* performs in lymphocytes are associated with the demyelination of neighboring neurons. Knocking down or out *Ppif* led to neuronal protective effects in murine disease models of multiple sclerosis {Forte:2007df, Warne:2016de}. In summary, SCREEN enables the user both to identify the cell types that are likely to be implicated in a disease and explore possible mechanisms by which cREs and SNPs may cause the disease.

## Combining human and mouse orthologous cREs to fine map GWAS SNPs

One particular strength of the Registry is the inclusion of both human and mouse cREs and the definition of orthologous cREs in these two species. The mouse cREs are mostly defined using tissues during embryonic development, in particular, four brain regions forebrain, midbrain, hindbrain, and neural tube; such developmental tissues are impractical to obtain for humans. Thus the orthologous mouse cREs can complement the human cREs in applications such as interpreting GWAS variants associated with developmental diseases especially those that afflict the brain.

For example, rs13025591 was reported by two studies to be associated with Schizophrenia (p-values 8E-8 and 6E-6) {SchizophreniaPsychiatricGenomeWideAssociationStudyGWASConsortium:2011bm, Bergen:2012il}. It lies in the intron of the *AGAP1* gene, most highly expressed in bipolar spindle neurons and the eye in human and all assayed embryonic brain regions in mouse ([**Extended Data Fig. 33**](https://docs.google.com/presentation/d/1xx2lYoBBmfvunbOc9jvtWb1XrbHYg4XjQzPebyNp9HE/edit#slide=id.g1b03e2385d_0_46)). rs13025591 does not lie within a cRE; therefore, we hypothesized that the signal driving this genetic association was from SNPs in high LD with rs13025591, and there are five cREs that overlapped such SNPs (**Fig. 8a**). Four of these cREs showed enhancer-like signatures, and one showed promoter-like signatures. None of the five cREs showed 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, Weinberger:1986up}; nevertheless, EH37E0579839 had high H3K27ac signal in neural cells and bipolar spindle neurons.

SCREEN's Activity Profile tool, which displays DNase or histone modification signals at cREs as “mini-peaks” across cell types, reveals that EH37E0579839 has high DNase signals in human fetal brain and eye but the signals disappear in older fetal brain and adult brain (**Fig. 8b**). EH37E0579839 is orthologous to the mouse cRE EM10E0042108, which shows enhancer-like signatures in brain tissues. Consistent with its human ortholog, EM10E0042108 has high DNase signals in embryonic brain and retina. Across the twelve tissues at eight time-points of embryonic development, EM10E0042440 has the highest H3K27ac signals in brain regions ([**Extended Data Fig. 34**](https://docs.google.com/presentation/d/1xx2lYoBBmfvunbOc9jvtWb1XrbHYg4XjQzPebyNp9HE/edit#slide=id.g1deab68534_0_14)). In forebrain, midbrain, and hindbrain, H3K27ac signals increase over time, reaching a maximum at e13.5. Then, like the human ortholog, it decreases after this time-point through birth (**Fig. 8c**). These results indicate that this cRE is only active during a narrow window of brain development.

The region harboring these two orthologous cREs is conserved across mammals (**Fig. 8d**). Although we do not have TF ChIP-seq data in fetal brain or mouse embryonic brain, motif analysis using both HaploReg and RegulomeDB {Ward:2011gh, Boyle:2012em} reveals that the LD SNP rs13031349 overlaps an SP3 motif and improves the match from a log-odds score of 8.1 to 19. Additional experiments are needed to test whether the SNP improves SP3 binding, but by using SCREEN and Encyclopedia, we were able to narrow down a region to guide the experimental testing for biological function.

In conclusion, we have integrated ENCODE and Roadmap Epigenomics data to produce an “Encyclopedia,” a compendium of candidate functional elements in human and mouse. Since the precise biological functions of a tiny fraction of cREs have been experimentally tested to-date, the Encyclopedia empowers the exploration of the regulatory roles of these elements in diverse cell types. It serves as a continually improving resource to enable the broader scientific community to seek new understandings in biological processes and diseases.

# FIGURE LEGENDS

**Figure 1** | **ENCODE Phase III data production as of February 1, 2017.** **a,** Human and **b,** Mouse ENCODE phase III experiments available on the ENCODE portal. Experiments are categorized by assay and biosample type: blue for immortalized cell line, red for tissue, teal for in vitro differentiated cells, orange for primary cells, purple for stem cells and pink for iPSCs.

**Figure 2** | **New assays used in ENCODE Phase III.** **a–b**, Using the 5' ends of RAMPAGE reads, we can identify TSSs and quantify tissue- and transcript-specific transcription. **a,** In testis, we identified a novel, tissue-specific TSS for *ARHGAP23* upstream of previous annotated TSSs. **b,** In spleen, we identified a novel TSS within exon 7 of *ARHGAP23*. **c–d,** Integrative analyses of RBP data can identify genetic variants that may impact RBP regulation. **c,** Control and RBFOX2 knockdown RNA-seq of exons 65–67 of the *UTRN* gene in HepG2 cells. Inclusion of the alternatively spliced exon 66 is reduced from 87% in control cells to 29% in RBFOX2 KD cells. **d,** (right) A strong RBFOX2 eCLIP binding peak in the downstream intron is consistent with this splicing factor enhancing inclusion of the upstream alternative exon. The minor allele of an ExAC SNP in the eCLIP peak in is expected to abrogate RBFOX2 binding as it abolishes the high affinity binding site determined from RNA Bind-n-Seq (RBNS). (left) Effect of the ExAC variant on the RBFOX2 binding site as determined from RBNS data. The G->C SNP in the eCLIP peak changes the most enriched 5-mer that likely mediates RBFOX2 binding (GCAUG *R* = 13.78) to a 5-mer with no detectable *in vitro* binding (CCAUG *R* = 0.89).

**Figure 3** | **Overview of the ENCODE Encyclopedia and prediction and validation of mouse embryonic enhancers.** **a,** Overview of the ENCODE Encyclopedia. The Encyclopedia consists of two levels (ground and integrative) which utilize data processed by the uniform processing pipelines. SCREEN integrates these data and annotations and allows users to visualize them on the UCSC genome browser **b,** Validation rates of 151 enhancer-like regions tested using transgenic mouse assays. Dark color indicates the region was active in the predicted tissue while light color indicates a lack of activity in the predicted tissue but with activity in other tissues. **c–e,** Examples of enhancers (orange boxes) that were predicted based on DNase signal (green) and H3K27ac signal (orange) and validated in **c,** midbrain, **d,** hindbrain and **e,** limb. H3K27ac signal (yellow) in across tissues accurately predicts additional observed activity.

**Figure 4 | Selection of cREs and assignment of cREs to nine states and five groups in a particular cell type.** **a,** Method for the section of cREs. We begin by clustering high quality DHSs (FDR > 0.1%) to create representative DHSs (rDHSs). For each assay (DNase, H3K4me3, H3K27ac or CTCF), we calculate a Z-score for every rDHS in a particular cell or tissue type. We then obtain the maximum Z-score across all cell types, known as the Max-Z. Using the Max-Z as well as the distance to the nearest TSS, we classify cREs into three cell-type agnostic groups using the decision tree: cREs with promoter-like signatures, cREs with enhancer-like signatures, and cREs bound by CTCF only. **b,** Given a cell type (shown for GM12878), we assign cREs into nine states based on whether they have high Z-scores (> 1.64) for H3K4me3, H3K27ac, CTCF, and DNase in that cell type. Each cRE is either proximal (≤ 2 kb) or distal (> 2 kb) to the nearest GENCODE-annotated TSS, and the bar graph shows the tally for each state in GM12878. Icons mark the states to the left of the bars. Colored boxes (for proximal cREs) and pies (for distal cREs) represent high Z-scores while white ones represent low Z-scores. **c,** Assignment of cRE states to five groups: with promoter-like signatures, with enhancer-like signatures, CTCF-only, DNase-only, and inactive. The bar plot shows the median ChIP-seq signal for POL2, EP300 and RAD21 in GM12878 for cREs in each category.

**Figure 5 | Overview of SCREEN's interface and utilities. a,** SCREEN’s regulatory-element-centric search view. Using the facets on the main search page (top), the user can retrieve candidate regulatory elements (center) by genomic coordinates and activity profiles in a particular cell type; here, four cREs in chromosome 11 are shown active in K562 cells. The cRE details view shows neighboring genes, bound transcription factor, and mini-peaks epigenetic signals (bottom left, shown here for the top cRE in the search table). A trackhub is custom built for visualizing a cRE or a gene and the supporting data using the UCSC genome browser (bottom right, top cRE from the table highlighted in blue). **b,** SCREEN’s gene-centric view provides RNA-seq and RAMPAGE derived expression levels for the genes and TSSs near the cRE of interest. **c,** SCREEN’s SNP-centric view displays cREs that overlap SNPs from published GWAS studies and lends insight into which cell types may be relevant to a particular phenotype. The top two cell types are shown for an inflammatory bowel disease GWAS study, along with two cREs active in CD4+ T-cells which contain SNPs from the study.

**Figure 6 | Analyzing differential gene expression and cRE activity across developmental time-points. a,** Comparison between Limb e11.5 and e15.5 gene expression and cRE activity. Blue bars indicate differentially expressed genes while red and yellow dots indicate cREs promoter-like and enhancer-like signatures. The heights of bars or dots indicate changes (Log2 FC or difference in Z-score) between time-points. **b,** Genome browser view of the *Ogn* locus with H3K27ac, H3K4me4, DNase, and RNA-seq signals for the limb across all surveyed time-points. Promoter-like cREs are designated by red bars and enhancer-like cREs are designated by orange bars. **c,** Ogn gene expression and nearby cRE activity increase coordinately across time-points. The increase in gene expression lags behind the increases in cRE-PLS and cRE-ELS activities.

**Figure 7 | Annotating GWAS variants using SCREEN. a,** The user can select from a preloaded list of GWAS. For each study, we included all tagged SNPs it reported and all SNPs in LD with them (*r*2 > 0.7). **b,** SCREEN reports the percent of LD blocks of a GWAS with at least one SNP overlapping a cRE. **c,** SCREEN ranks cell and tissue types based on enrichment in H3K27ac signals. The top 5 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 analyze the overlapping cREs. **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*.

**Figure 8 | Fine mapping GWAS variants using SCREEN. a,** H3K4me3 and H3K27ac Z-scores for cREs containing SNPs in LD with the schizophrenia-associated SNP rs13025591. H3K4me3 Z-scores and H3K27ac Z-scores are displayed in red and yellow, for cREs with promoter-like and enhancer-like signatures respectively. **b,** SCREEN's Activity Profile tool allows the user to view DNase peaks at cREs across all cell types. Both the human cRE EH37E0579839 and its orthologous mouse cRE EM10E0042440 show high DNase signals in developing brain and eye tissues. **c,** H3K27ac signal at EM10E0042440 over developmental time in mouse forebrain (red), midbrain (green) and hindbrain (blue). **d,** Zoomed-in view of EH37E0579839. The SNP rs13031349 overlaps both EH37E0579839 and the orthologous mouse cRE EM10E0042440. The SNP also overlaps an SP3 motif, resulting in a change in the motif score.