2017-12-12

Analysis, Visualization, and Machine Learning of Epigenomic Data

Michael J. Purcaro
University of Massachusetts Medical School

ANALYSIS, VISUALIZATION, AND MACHINE LEARNING OF EPIGENOMIC DATA

A Dissertation Presented

By

MICHAEL JOSEPH PURCARO

Submitted to the Faculty of the
University of Massachusetts Graduate School of Biomedical Sciences, Worcester
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

DECEMBER 12, 2017

BIOINFORMATICS AND COMPUTATIONAL BIOLOGY

M.D., PH.D. PROGRAM

ANALYSIS, VISUALIZATION, AND MACHINE LEARNING OF EPIGENOMIC DATA

A Dissertation Presented
By
MICHAEL JOSEPH PURCARO

This work was undertaken in the Graduate School of Biomedical Sciences

Bioinformatics and Computational Biology

Under the mentorship of

Zhiping Weng, PhD, Thesis Advisor

Elinor Karlsson, PhD, Member of Committee

Manuel Garber, PhD, Member of Committee

Robert Brewster, PhD, Member of Committee

Ross Hardison, PhD, External Member of Committee

Jeffrey Bailey, MD, PhD, Chair of Committee

Anthony Carruthers, Ph.D.,

Dean of the Graduate School of Biomedical Sciences

December 12, 2017

This work is dedicated to the most loving, understanding,
and special woman in my world: my wife and best friend, Zemei.

It is also dedicated to my parents,
who have always encouraged me
to pursue my dreams,
despite the many, many years of training it has entailed.

Acknowledgements

Being involved for 4.5 years in an undertaking as complex as the ENCODE Consortium complicates writing acknowledgments. With so many moving pieces—data, metadata, code, papers, conference calls, and Google Docs—to deal with, ENCODE, and especially the Weng Lab, is a swirl of activity and ideas; keeping track of the intellectual and personal impact so many people have made on me would be a job in itself. Certain people, though, were particularly important in my journey: Arjan van der Velde, Nicholas Hathaway, Henry Pratt, Nathaniel Erskine, and Eugenio Mattei (in no particular order) were critical in my survival, allowing me to test ideas (about code, science, medicine, and life in general) and get honest, always entertaining feedback. I learned much from the intellectual prowess of our lab mates: Jill Moore, Tyler Borrman, Jack Huey, Sweta Vangaveti, Sowmya Iyer, Xiao-Ou Zhang, Shikui Tu, Junko Tsuhi, Jie Wang, Hao Chen, Thom Vreven, Yu Fu, Adam Wespiser, and Wei Wang.

From the long suffering ENCODE DCC group, I would like to particularly thank "data heroes" Cricket Sloan, Kathrina Onate, and Jason Hilton for all their work in supporting us. From the IT department, I would like to thank David Plamondon, Lewis Robbins, and Charles Davidson for help and advice. I am very appreciative of mentoring and advice from Thomas Smith, who has helped put my work in a wider clinical perspective. I am also very appreciative of my TRAC committee—Jeffrey Bailey, Elinor Karlsson, Konstantin Zeldovich, and Robert Brewster—for helping me steer through this process, as well as the MD/PhD program for giving me the opportunity to undertake training here for a career as a physician scientist.

Life in the lab would come to a crashing halt without the endless support, advice, and motherly watchfulness from Barbara Bucciaglia, Heidi Beberman, Christine Tonevski, and Maureen Schulz. And, of course, none of this would have been possible without Zhiping Weng, whose endless generosity, curiosity, and intellectual insight have been deeply inspiring, and made the lab into an intellectual—and computational—playground.

Abstract

The goal of the Encyclopedia of DNA Elements (ENCODE) project has been to characterize all the functional elements of the human genome. These elements include expressed transcripts and genomic regions bound by transcription factors (TFs), occupied by nucleosomes, occupied by nucleosomes with modified histones, or hypersensitive to DNase I cleavage, etc. Chromatin Immunoprecipitation (ChIP-seq) is an experimental technique for detecting TF binding in living cells, and the genomic regions bound by TFs are called ChIP-seq peaks. ENCODE has performed and compiled results from tens of thousands of experiments, including ChIP-seq, DNase, RNA-seq and Hi-C.

These efforts have culminated in two web-based resources from our lab—Factorbook and SCREEN—for the exploration of epigenomic data for both human and mouse. Factorbook is a peak-centric resource presenting data such as motif enrichment and histone modification profiles for transcription factor binding sites computed from ENCODE ChIP-seq data. SCREEN provides an encyclopedia of ~2 million regulatory elements, including promoters and enhancers, identified using ENCODE ChIP-seq and DNase data, with an extensive UI for searching and visualization.

While we have successfully utilized the thousands of available ENCODE ChIP-seq experiments to build the Encyclopedia and visualizers, we have also struggled with the practical and theoretical inability to assay every possible experiment on every possible biosample under every conceivable biological scenario. We have used machine learning techniques to predict TF binding sites and enhancers location, and demonstrate machine learning is critical to help decipher functional regions of the genome.

Table of Contents

List of Tables

List of Figures

List of copyrighted Materials Produced by the Author

This thesis contains no copyrighted material published by the author.

List of Third Party Copyrighted Material

This thesis contains no third party copyrighted material.

Imparo ancora

(Italian for "still learning")

I. Chapter I: Introduction

Our experience hitherto justifies us in trusting that nature is the realization of the simplest that is mathematically conceivable.

–Albert Einstein, Herbert Spencer Lecture, 1933

We, as humans, have an innate, natural curiosity about ourselves, how we work, and the myriad ways in which we malfunction. In many ways, this curiosity has been codified and matured by the scientific method into modern molecular biology. What we have found so far is that our human genome—our code—is beautifully innate and immensely complex. That there is structure in this code, though, is becoming clearer. The central dogma of molecular biology—that sequence information from DNA is transcribed into mRNA, and mRNA is ultimately translated into protein (Crick 1958)—belies an enormous amount of machinery controlling this biological flow of information.

A genome is composed of coding and non-coding regions of DNA. Coding regions get processed into protein products, while non-coding regions have a myriad of functions. Large eukaryotic genomes must be packaged and folded multiple times to fit into a cell nucleus. The first level of DNA packaging into chromatin occurs by winding DNA around histone proteins, forming structures called nucleosomes. Eight histone proteins form the core of the nucleosome, with each of 4 histone proteins (H2A, H2B, H3, and H4) found twice. Histone protein tails can undergo a large number of post-translational chemical modifications. For example, the 27th lysine residue of H3 can be acetylated (H3K27ac, for short), or the 4th lysine residue on H3 can be trimethylated (H3K4me3). These modifications have wide-ranging effects on cellular processes, regulating everything from gene expression and the cell cycle to DNA replication and apoptosis (Wang et al. 2001; Koprinarova, Schnekenburger, and Diederich 2016; Eberharter and Becker 2002). Nucleosomes can then be further packaged into increasingly compact and complex chromatin structure. This 3D structure enables gnomic elements separated by large linear genomic distance to suddenly be able to interact; almost any genomic location has a non-zero probability of interacting with any other portion of the genome (Dekker, Marti-Renom, and Mirny 2013). The winding and unwinding of chromatin all the way down to modifications of histone tails changes the accessibility of the gnome; local DNA accessibility changes influence where transcription factors can bind promoter and enhancer regions, affecting gene expression. Experimentally, chromatin accessibility is indicated by DNase I (a nuclease) digestion (Neph et al., 2012). DNase digestion followed by next generation DNA sequencing (DNase-seq) (Boyle et al., 2008) is now a widely-used and reliable technique, with experimental data available for hundreds of biosamples in ENCODE.

Certain patterns or signatures of chromatin accessibility and histone modifications have been found to have association with certain events. For example, the H3K27ac histone mark in a DNA-accessible region typically indicates that one or more activator proteins (called transcription factors) can bind and increase protein translation (Rada-Iglesias et al. 2011). These regions that increase gene expression are typically within a ~1 MB window of the gene (upstream or downstream), and are called enhancer regions (Gillies et al. 1983). Similarly, the H3K4me3 histone mark in a DNA-accessible region, within +/- 1,000 bases upstream of where transcription is initiated (the Transcription Start

Site (TSS)), and on the same strand as the gene generally indicates a promoter region (Heintzman et al. 2007).

Transcription factors (TFs) are DNA-binding proteins that regulate transcription of genetic information from DNA to RNA. TFs have activating or repressing activity via many potential mechanisms: they may complex with other TFs (Maston, Evans, and Green 2006), coactivator proteins, RNA polymerase II, chromatin remodeling complexes, and/or noncoding RNA molecules (Phillips 2008). DNA-binding TFs bind short (6-15 base pair) fragments of genomic DNA; a particular location is called a motif site or TF binding site. These sites demonstrate high evolutionary conservation (Chen & Rajewsky, 2007). Motif sites show distinct cleavage patterns (called footprints) after digestion by DNase I. While DNase-seq provides footprint data reflecting the presence of any DNA-binding proteins, additional evidence of a particular TF-DNA bound complex can be experimentally determined through chromatin immunoprecipitation (ChIP) with massively parallel DNA sequencing (ChIP-seq) (Johnson, Mortazavi, Myers, & Wold, 2007). Thousands of ChIP-seq datasets are available for hundreds of DNA-binding TFs (Wang et al., 2012). There are, however, thousands of different DNA-bound TFs in the human genome (Wilson et al. 2008), and TF binding depends upon many factors, including cell type specificity, phase of development, and/or experimental design. It is increasingly clear that many diseases are a product of genetic variations in regulatory regions of the genome, frequently in regions impact regulatory TF binding (Lee and Young).

The role of genetics in understanding disease pathology has become a central aspect of medicine, with an explosion of research occurring in the past few years. It is increasingly apparent, however, that understanding how changes “above” the genome—in the epigenome—is central to both advancement of basic science and to the translation of these findings to clinical therapy. Around 90% of disease-associated Single Nucleotide Polymorphisms (SNPs) have been found to be in intronic or intergenic regions across multiple Genome Wide Association (GWAS) studies (Hindorff et al. 2009). These genetic variants outside of protein-coding regions indicate that disease pathology may be altered by changes in regulatory regions of the genome, in functional regions such as enhancers and promoters (Hrdlickova et al. 2014). Better understanding of the epigenome, including building an encyclopedia of all functional elements that details how and why these elements work, is central to this advancement.

Deciphering this complex orchestra of histone modifications, chromatin remodelers, transcription factors, etc. is central to better understanding the epigenome. Since 2003, the Encyclopedia of DNA Elements (ENCODE) project has collecting and analyzing data in a large-scale to characterize all the functional elements of the human genome. Thus far, ENCODE has successfully collected thousands of chromatin accessibility, transcription factor, and histone modification experiments, finding hundreds of millions of regions of putative regulatory function across hundreds of different biosamples. ENCODE has also been highly influential in developing and publishing standards guidelines for DNA-seq, ChIP-seq, and RNA-seq experiments, as well as cloud-scale, open-source pipelines for processing these experiments. ENCODE has also collected hundreds of gene expression experiments, as well as developing (through GENCODE) a curated set of gene annotations.

Making an actual Encyclopedia of functional genomic elements, though, has proven difficult. The diversity of chromatin accessible regions, histone modification patterns, and TF binding sites across all the different cell and tissue types of the human body has made clear there is a combinatorial number of activation signatures in the genome, as well as millions of potential functional elements. In this thesis, we develop systematic methods of selecting cell-type specific candidate Regulatory Elements (cREs), and demonstrate the biological usefulness of these regions. We have also assigned stable IDs (called accessions) to these regions, with the intent to construct a stable, curated annotation of functional genomic elements, just as Ensembl does for genes (Birney et al. 2004; Aken et al. 2016).

As the visual analytics field has shown, just being able to display raw data is not useful: the extracted analysis products are where the real value is (Keim 2010). Just making an Encyclopedia of cREs is insufficient unless there are ways to visualize and understand the genomic and epigenetic context the putative functional elements exist in. The great importance of being able to visualize highly-multidimensional omic data is clear; there are a multitude of examples available, with even entire frameworks being developed for pathways and gene expression visualization (Streit et al. 2009). Some of the most exciting developments include a Google-maps view of 3D chromatin structure analysis products (Perkel 2017). In this thesis, we develop two visualizers—SCREEN and Factorbook—to assist users in searching and viewing the millions of elements and data points available from the ENCODE data.

While we have successfully utilized the thousands of available ENCODE ChIP-seq experiments to build the Encyclopedia and visualizers, we have also struggled with one of the core limitations of ChIP-seq TF experiments. The binding of a transcription factor at a particular motif site in the genome depends upon a diverse number of factors. While experimental methods seq have begun to shed light on these binding patterns, fully understanding regulation via transcription factor binding, though, will require an enormous number of ChIP-seq experiments. Given the wide variety of conditions affecting binding, millions of different experiments are required to comprehensively understand when and where transcription factors bind (Lee and Young). Accurately predicting transcription factor binding sites through statistical and machine learning methods could drastically reduce the number of experiments needed. Further, improved understanding of transcription factor binding would shed light on gene regulatory networks present during embryogenesis, development, and disease states.

Several different labs have developed predictive models for motif site binding over the last decade. Previous predictive models (such as CENTIPEDE and PIQ) demonstrate the initial feasibility of probabilistically predicting the bound state of a motif site in the genome. For experimental input data, these models primarily utilize DNase-seq data for elucidation of chromatin state and, ultimately, motif site binding probability. In this thesis, we develop several predictive models based upon supervised learning methods. These models leverage ChIP-seq data already acquired by ENCODE participants, as well as other ENCODE and Roadmap Epigenomics datasets. We also develop a large number of features for model training, using supervised approaches, and have achieved some success in predicting TF binding sites. We have also utilized some of these techniques while competing in the ENCODE-DREAM in vivo Transcription Factor Binding Site Prediction Challenge.

This thesis describes efforts toward better understanding and visualizing the epigenome. Chapter II will introduce our current version of the ENCODE Encyclopedia. In it, we demonstrate how we locate regions of the genome with open chromatin and enhancer-like or promoter like signatures based on histone modification marks and other genomic distance information. These regions—candidate Regulatory Regions (cREs)—are our first approximation of a systematic, accessioned, genomic-wide catalog detailing regions that are involved with functional control of the genome. Chapter II will introduce the latest version of an aggregated, peak-centric visualizer for transcription factor binding sites (TFBS). Chapter IV discusses our work on imputing entire epigenetic experiments, first focusing on predicting locations of TFBS. Lastly, Chapter V introduces SnoPlowPy, our tool driving metadata and job management functionality for large-scale analysis projects.

II. Chapter II: Building and Visualizing an Encyclopedia of ENCODE candidate Regulatory Elements

II.1 Preface

This research chapter encompasses work performed by Jill Moore, myself, Henry Pratt, Zhiping Weng, and >500 other collaborators in the ENCODE Consortium. The chapter combines one manuscript currently in review (as of December 2017) with another manuscript on SCREEN (with Henry Pratt) that is currently in draft.

With more than 14,000 experimental datasets, consuming more than 0.5 petabytes of storage space across hundreds of thousands of files, the ENCODE project has built a vast catalog of gene expression, chromatin accessibility, histone medication, and transcription factor binding data. While investigating these data, I found the data difficult to utilize to answer fundamental biological questions, such as where putative enhancers and promoters are located, or how gene expression levels vary across disease conditions or developmental time points. These questions were impossible to answer without manually curating, downloading, and processing the data. Just determining which files to use for such analyses was also non-intuitive and essentially undocumented, and the wide-variety of data processing techniques added many subtle problems when integrating data across labs, let alone different experiments. I decided to ameliorate these problems and allow straightforward analysis of the data and generation of actionable biological insights and hypotheses.

While individual tools to interrogate particular regions of the genome have existed for more than two decades (Kent et al. 2002), and epigenetic annotations have been available for years (Frankish et al. 2015), no tool has integrated as much genetic and epigenetic data in one location as SCREEN. SCREEN solves many of the problems (both biological and practical) described above. Inside of SCREEN, I have integrated and condensed thousands of ENCODE experiments into an easy-to-use product that allows researchers to intuitively explore the available data. I am the overall architect of SCREEN, having designed and implemented the database system, data import pipeline, and much of the software architecture. SCREEN excels at allowing the user to develop hypotheses for potential functional regulation across millions of region in the human and mouse genomes.

SCREEN offers new solutions to help navigate the vast sea of data. I am also one of the first to build an online database of hundreds of millions of DNase and ChIP-seq peaks that could then be intersected with candidate regulatory regions at the click of a button. The accessioning system for peaks I implemented is the start of a critical new stage of epigenetics, where individual regions can be tracked not just through publications, but across hundreds to thousands of experiments. Current projects like ENCODEproject.org are designed to ascension a few hundred thousand objects; the systems are not capable of supporting millions of objects. Comparing how these regions change across developmental or disease states becomes not just far more straightforward, but, in fact, doable in seconds, not hours or days it would take before. This approach of systematically cataloging regions of the genome will become as integral to the field as plant taxonomy became for the field of botany.

II.2 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 to assess how genomic variants affect development, ageing, and susceptibility to diseases. The goal of the Encyclopedia of DNA Elements (ENCODE) project is to discover and characterize the full repertoire of elements (www.encodeproject.org). Here, we summarize 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 Regulatory Elements (cREs), 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 cREs, covering hundreds of biosample types. The cRE landscape recapitulates the current understanding of cellular identity, tissue composition, developmental progression, and disease-associated genetic variants. Aided by a dedicated visualization engine called SCREEN (screen.encodeproject.org), the Registry is a resource for exploring noncoding DNA elements and their variants.

II.3 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 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 overarching goal of ENCODE is to provide an integrated resource to aid the scientific community in studying mammalian biology and human diseases.

In pursuit of this goal, ENCODE develops and applies high-throughput experimental technologies and computational approaches to catalogue candidate functional elements in the human and mouse genomes, including transcripts and their regulatory elements. The pilot phase of ENCODE focused on 44 carefully selected regions covering 1% of the human genome using array-based techniques (Birney et al. 2007). Phase II used deep-sequencing-based biochemical assays to interrogate the entire human genome, producing 1,640 datasets, and integrative analyses of these datasets identified an extensive set of candidate functional elements (Consortium 2012). The related Mouse ENCODE (Yue et al. 2014) and modENCODE projects (Gerstein et al. 2010; Roy et al. 2010) performed thousands of genome-wide experiments on the mouse, fly, and worm. Complementary projects, including the NIH Epigenomics Project (Kundaje et al. 2015) and the International Human Epigenome Consortium have also produced thousands of epigenomic maps for human cells and tissues (Stunnenberg, International Human Epigenome, and Hirst).

Despite this progress, the human and mouse genomes remain only partially annotated, limited by the depth of biochemical element types mapped for any one cell type and the breadth of cell types mapped for any single biochemical feature.

Accordingly, our understanding of the diversity of transcripts and their regulatory elements in each cellular context is far from complete. To begin to address these limitations, ENCODE Phase III expanded data collection in both depth and breadth—studying additional chromatin features, regulatory factors and RNA types with an emphasis on primary cells and tissue samples.

All data are submitted to the ENCODE Data Coordination Center, reviewed for quality and released to the scientific community via the freely accessible ENCODE web portal (www.encodeproject.org). ENCODE members have reported new findings throughout the past five years based on data generated and released during Phase III. Additionally, we have now assembled the ENCODE Encyclopedia of predicted and confirmed functional elements, based on all ENCODE data collected during Phases II and III, supplemented by data from the NIH Epigenomics Project. This chapter describes the ENCODE Encyclopedia and presents illustrative examples of its use.

A new focus in ENCODE Phase III has been to build a Registry of candidate Regulatory Elements (abbreviated as cREs). This effort is guided by the current understanding that robust biochemical signatures, including chromatin accessibility and particular histone modifications, 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 cREs that are assembled here, together with other genome annotations and their underlying experimental data, for exploration by users via a specifically designed visualization tool called SCREEN (screen.encodeproject.org).

II.4 Results

II.4.1 **Summary of Encode Phase 3 Data Production**

The ENCODE Consortium has produced data on three main aspects of genome activity—transcriptomes, DNA-based regulatory elements for transcription and replication, and RNA-based elements for post-transcriptional regulation. Phase III greatly expanded the number of experiments in each category and released 4,903 experiments (3,797 on human and 1,106 on mouse; see Figure II-1 on page 87). Table II-1 on page 78 summarizes these experiments by category. 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 type (such as a tissue, a cell line, primary cells, or stem cells). In this section, we summarize the new assays and highlight the results of Phase III data production.

New polyA and short RNA transcriptome data production has focused on primary cells from different body locations and various embryological origins. 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 analyzed captured RNAs using long-read sequencing. This effort, in collaboration with the GENCODE project, improved the annotations of gene and transcript structures for 14,667 human and 8,708 mouse long noncoding RNAs (Lagarde et al., in review).

A new 5′-complete cDNA sequencing assay called RAMPAGE quantifies gene expression, identifies promoter locations, and assigns 5′ capped termini to their corresponding RNA isoforms (Batut et al. 2013). RAMPAGE yields data at single-nucleotide resolution and is more accurate than RNA-seq for quantifying expression (Batut et al. 2013)—advances which enable it to improve transcription start site (TSS) annotation and transcript quantification. For example, the gene *ARHGAP23*, which encodes Rho GTPase-activating protein 23, has 12 GENCODE-annotated transcripts and 11 different TSSs. RAMPAGE data revealed a novel TSS in the testis (Figure II-2a on page 88), located 9.2 kb upstream of the nearest annotated TSS, and another novel TSS in exon 7 specific to the spleen (Figure II-2b on page 88). As another example, two different TSSs, 824 bp apart, are annotated by GENCODE V26 and UCSC for *EP300*, which encodes a widely studied histone acetyltransferase important for enhancer activity. RAMPAGE data across six cell and tissue types showed that although both TSSs are active, one TSS is used far more frequently than the other (Figure II-3 on page 89).

The coverage of noncoding, biochemically marked DNA elements, many of which have potential regulatory functions, has been greatly expanded during ENCODE Phase III. We completed 163 new DNase accessibility maps, including deep sequencing DNase-seq datasets on hundreds of cell and tissue samples, thus facilitating the prediction of regulatory protein occupancy by footprinting (Hesselberth et al. 2009). The ATAC-seq assay (Buenrostro et al. 2013), which assesses chromatin accessibility via insertion by the Tn5 transposome, was conducted on tens 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. Over 600

ChIP-seq experiments were completed in Phase III for 493 different transcription factors (TFs) in at least one cell type (1,622 experiments on 549 different TFs in Phases II and III combined). For these ChIP-seq experiments, we used either TF-specific antibodies or epitope-tagged TFs created by BAC transfections or CRISPR/Cas9 genome editing. ChIA-PET of Rad21 and CTCF, which are involved in the nuclear organization, along with Hi-C experiments, provide 3D linkage data that include many regulatory regions and cognate target genes. Through the ENCODE Portal (encodeproject.org/antibodies/), we provide quality metrics for all datasets as well as detailed information about the antibodies used in our experiments to help users evaluate and use the data most effectively.

DNA replication timing provides insights into gene regulation and spatiotemporal genome compartmentalization (Gilbert 2002). We measured replication timing during fate commitment of human embryonic stem cells, thus yielding 84 datasets for 26 cell types representing the embryonic layers endoderm, mesoderm, ectoderm, and neural crest (Rivera-Mulia et al. 2015) (see Figure II-4 on page 90). Because replication timing differs across cell types, we expected that clustering of these datasets would recapitulate their developmental lineages, and that was indeed observed (see Figure II-5 on page 91).

The mouse component of ENCODE Phase III focused on embryo development at daily intervals between embryonic day 10.5 (e10.5) and postnatal day 0 (p0), with 6-12 tissues sampled per day. RNA-seq of polyA RNAs 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.

A new project in ENCODE Phase III was to identify and characterize functional RNA elements bound by RNA-binding proteins (RBPs) (van Nostrand et al., in preparation). Four types of related data were generated: RIP-seq and enhanced UV crosslinking and immunoprecipitation of RBPs followed by sequencing (eCLIP-seq) (Van Nostrand et al. 2016) to identify bound RNAs in vivo and pinpoint 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 et al. 2014) 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 to RBP regulation. For the 18 RBPs with eCLIP, RBNS and, RBP-knockdown RNA-seq data, we identified 26 variants from the Exome Aggregation Consortium (ExAC)(Lek et al. 2016) 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). For example, intron 66 of *UTRN* (dystrophin-related protein 1) harbors an RBFOX2 eCLIP peak downstream of an alternatively spliced exon (Figure II-6c on page 92), which overlaps an ExAC variant (Lek et al. 2016). This G→C variant disrupts the RBFOX2 binding motif (GCAUG) at the first position. 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 (Figure II-6d on page 92), thus suggesting that the mutation disrupts RBFOX2 binding in vivo. To determine whether the disruption of RBFOX2 binding would alter splicing, we performed RNA-seq on HepG2 cells after knocking down RBFOX2. In wild-type cells, the upstream exon was included in 87% of messages, whereas the inclusion was decreased to 28% in the RBFOX2 knockdown cells (Figure II-6c on page 92). Taken together, these data argue that this G→C variant disrupts RBFOX2 binding, leading to decreased inclusion of the upstream exon in over half of *UTRN* messages, and resulting in an altered composition of protein isoforms. Overall, the actual number of variants that influence RNA metabolism is larger than the 26 ExAC variants identified in this way, because they may affect aspects of RNA biology other than splicing.

II.4.2 The Encode Portal and Uniformly Processed Data

The ENCODE portal (www.encodeproject.org) is the primary interface for retrieving all ENCODE data, metadata, data standards, and experimental protocols (Sloan et al. 2016). It also provides entry to the ground and integrative levels of the ENCODE Encyclopedia (Figure II-7 on page 93), which is described in the next section. The Portal is designed to provide users with extensive metadata that describe how ENCODE experiments were performed, processed, and connected in common biological themes (Hong et al. 2016). All experiments followed data production guidelines (www.encodeproject.org/about/experiment-guidelines/#guideline). An experiment typically comprises two biological replicates, with some exceptions; in the case of single-cell assays or assays utilizing human donor tissues of limited availability, for example, no cell is a conventional replicate of another. A released experiment includes the "raw" sequencing data (typically FASTQ files) and all analysis output files (such as alignment files, signal files, or peak files) from the uniform processing pipelines. These pipelines are central to ENCODE data, and the major pipelines are available for users to apply to their own data, either by downloading the code and running it locally or by accessing the pipelines at the DNAnexus cloud provider.

The Portal was completely redesigned during Phase III for better data access and metadata clarity. The homepage presents summaries of the numbers and types of experiments, with intuitive links for data access. Experiments are annotated by key features (called facets) so that users can easily find experiments via a faceted search. A matrix view displays the search results (encodeproject.org/matrix/?type=Experiment; see Figure II-1 on page 87), which can be switched to list or table views. Entries in the matrix are hyperlinked to underlying datasets, along with metadata and quality metrics.

II.4.3 The Encode Encyclopedia

The raw data described above and their signal maps across the human and mouse genomes are valuable for interrogating genome function in myriad 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 organized them into the ENCODE Encyclopedia. The Encyclopedia presently has two levels of annotations (Figure II-7 on page 93). The ground level includes peaks and quantifications produced by the uniform data-processing pipelines for individual data types, and the integrative level contains annotations derived from combined analyses across multiple data types and ground-level annotations.

II.4.4 Encyclopedia Ground Level

The ground level currently has nine components (Figure II-7 on page 93). The chromatin accessibility component contains DNase hypersensitive sites (DHSs)—genomic regions significantly enriched in DNase-seq reads—and their constituent DNase peaks, as well as ATAC-seq peaks. Locations of histone marks and histone variants are provided in the histone modification component as histone peaks, which are regions of the genome significantly enriched in histone ChIP-seq reads. The transcription factor binding component provides TF peaks, or genomic regions significantly enriched in TF ChIP-seq reads; these peaks are further characterized by enriched sequence motifs (identified using the MEME-ChIP tool (Machanick and Bailey 2011)) and the average histone mark ChIP signals and nucleosome occupancy signal surrounding them in each cell type. The TF peaks and associated information can be viewed in the wiki-style web resource Factorbook (see page 136). The gene and TSS expression components give 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 (see Figure II-8 on page 94).

The RNA binding protein (RBP) 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(Van Nostrand et al. 2016). The DNA methylation component analyzes whole-genome bisulfite sequencing data and provides the methylation state for each cytosine in the genome. 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.

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 these updates do not constitute new versions. More components will be added as additional analysis pipelines are developed and existing pipelines are improved. Components of the integrative level of the Encyclopedia are versioned as described below.

II.4.5 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 active or repressed in individual cell types. In pursuit of this goal, ENCODE and Roadmap Epigenomics Consortia have now produced basic epigenetic signals broadly in hundreds of human and mouse cell types and tissues. ENCODE has also examined a few cell types much more extensively for diverse transcription factor occupancy, genome-wide DNA methylation, RNA-binding protein occupancy and other “deep” assays. These differences in assay breadth versus depth have motivated two complementary computational approaches to build catalogues of candidate transcriptional regulatory elements, and our Encyclopedia offers both.

The first approach started in ENCODE II. It uses machine learning methods such as ChromHMM(Ernst and Kellis 2010; Ernst and Kellis 2012) and Segway (Hoffman et al. 2012) to integrate many different types of epigenetic signals. ChromHMM and Segway are unsupervised probabilistic models that integrate a specified number of epigenetic signals to define a large repertoire of chromatin states, many of which correlate with known functional element types and activity levels, e.g., active promoters, enhancers, or heterochromatin domains. ChromHMM have been augmented to accommodate cell types with some missing assays and then applied to the contemporary Roadmap (Ernst and Kellis 2015a) and ENCODE III cell types and tissues that achieved sufficient assay coverage. A strategy was developed in ENCODE III to train separate Segway models on each cell type—allowing for different assay coverages in different cell types—and then automatically 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 (Libbrecht et al. 2016). We similarly applied ChromHMM to the mouse embryo development series—66 complete epigenomes each assayed by ChIP-seq of eight histone marks— and defined 15 chromatin states that showed coordinated changes with gene expression measured by RNA-seq for each of the 66 samples (Gorkin et al., in preparation, Tsuji et al., in preparation). The resulting chromatin state maps from this section 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 primary cells and tissues have led to incomplete assay coverage for many of these samples. Thus we have developed an approach that uses a highly parsimonious combination of just four types of assays to maximize the coverage of cell and tissue types, though at the expense of subtler inferences about each element's possible activity. The rest of the chapter focuses on the second approach that has led to the new Registry of candidate Regulatory Elements.

II.4.6 The Registry of candidate 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 cREs in the genome. The most direct approach to identifying cREs 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 et al. 2013; Erwin et al. 2014). 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 that is based on just four epigenetic signals that we found to be most predictive of regulatory elements: chromatin accessibility (measured by DNase-seq), the histone modifications H3K4me3 and H3K27ac, and CTCF binding. This selection was initially motivated by substantial prior work in the field. DNase hypersensitive sites delineate all the main classes of cis-regulatory elements in a cell-type-specific manner, including promoters, enhancers, insulators, and locus control regions (Thurman et al. 2012). H3K4me3 and H3K27ac are the two histone marks most enriched at promoters and enhancers respectively (Heintzman et al. 2007; Visel et al. 2009). CTCF is the established insulator binding protein in mammals (Kim et al. 2007) and its binding sites are enriched at interacting chromatin loci (Rao et al. 2014).

To further test our selection, we compared the effectiveness of ten different types of epigenetic signals in predicting enhancers in the corresponding tissue: DNase hypersensitivity, eight histone marks (H3K4me1, H3K4me2, H3K4me3, H3K27ac, H3K9ac, H3K9me3, H3K36me3, and H3K27me3), and DNA methylation. These epigenetic signals were all assayed with specific mouse e11.5 tissues during ENCODE III, and the tissue-specific e11.5 enhancers tested using in vivo transgenic assays were from obtained the VISTA database (Visel et al. 2007). We found that DNase and H3K27ac were the best single features for predicting tissue-specific enhancers. We then used RNA-seq to evaluate the effectiveness of these same epigenetic signals in predicting gene expression levels and found H3K4me3 to the best single feature. We found that DNase offers high spatial precision in defining cREs: DHSs are ~350 bp long and typically correspond to the core of regulatory elements. In contrast, the H3K27ac and H3K4me3 signals are more diffuse: they tend to be low at the center of a regulatory element—presumably because of the lack of a nucleosome there—but are elevated at flanking nucleosome positions. DNase, therefore, presents the best localization of a cRE, while H3K27ac and H3K4me3 suggest the recent activity state, and the coincidence of significant signals from at least two assay types increases the overall confidence in the cRE.

To experimentally test the enhancer branch of our predictor, we used the average rank of the DNase and H3K27ac signals to identify previously untested TSS-distal (> 2 kb from the nearest TSS) candidate enhancers in the mouse e11.5 hindbrain, midbrain, and limb. The boundaries for the predicted regions were defined using the H3K27ac ChIP-seq peaks called by the MACS2 algorithm (Zhang et al. 2008) (see Figure II-9 on page 95). For each tissue, we tested 20, 15 and 15 new regions around the ranks 1-20, 1500-1520, and 3000-3020, respectively. In total, we tested 151 regions (for results, see online Supplementary Table 4). Representative e11.5 transgenic embryos for the enhancers that validated in the expected tissues are shown in Figure II-10 on page 96. Consistently, higher ranking regions were more likely than lower ranking regions to show enhancer activity in their predicted tissue (Figure II-11 on page 97; e.g., 75%, 26.6%, and 20% for the hindbrain). When enhancers were active in multiple tissues, these tissues also had high H3K27ac signals across the predicted enhancer regions (Figure II-12c-e on page 98). For example, a predicted enhancer in the hindbrain was also active in the midbrain and neural tube; accordingly, high H3K27ac signals were observed in all three tissues (Figure II-12d on page 98). In contrast, an enhancer active almost exclusively in the limb (Figure II-12e on page 98) 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.

In aggregate, our evaluations showed that combining DNase with two histone marks, H3K4me3 and H3K27ac, is an effective way to build a first version of the Registry of candidate promoters and enhancers active in specific cell types. We extended this predictor by adding CTCF, a highly conserved architectural protein that binds to insulators and contributes to the establishment and maintenance of three-dimensional chromatin structure (Ong and Corces 2014). Our final algorithm anchors cREs on a representative set of all DHSs, and then evaluates cRE types and activities based on H3K4me3, H3K27ac, and CTCF signals. To maximize coverage, we applied the algorithm to all cell types interrogated by at least one of these assays, making it possible to include data from 301 human cell types (620 when primary cells or tissues from different donors are counted separately) and 58 mouse cell types (138 with developmental time-points counted separately) with all ENCODE and Roadmap data considered. It is thus important to note that we distinguish two classes of cREs displaying no activity in a given cell type: cREs for which necessary assays are missing in the cell type, and cREs for which the necessary assays are present but the associated signals did not score as significantly positive.

The first release of the Registry presented here includes 1.31 million human cREs and 0.43 million mouse cREs; future versions will be released periodically, and are already under development. Based on the levels of the four core epigenetic signals and the distance to the nearest annotated TSS, we also classify cREs as those that have promoter-like signatures (PLS) or enhancer-like signatures (ELS) or as those that lack these signatures but are bound by the insulator-binding protein CTCF.

II.4.7 Selection of cREs for the Registry

We define cREs as DHSs supported by at least one additional type of epigenetic signal among H3K4me3, H3K27ac, and CTCF in at least one cell type. We first condensed all DHSs from individual samples into a set of non-overlapping representative DHSs (rDHSs) as described in Methods. We then filtered out the rDHSs with Z-scores less than 1.64—a threshold corresponding to the 95th percentile of a one-tailed test. Approximately 1.6 M human and 0.63 M mouse rDHSs remained. The rDHSs that have high H3K4me3, H3K27ac, or CTCF signals (a high signal is defined as a Z-score > 1.64 throughout) in at least one cell type are designated cREs. In total, there are 1,310,152 human cREs (Figure II-13 on page 99) and 431,202 mouse cREs. 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 recognized for having "concordant" support, labelled with an asterisk by their accessions in SCREEN. The remaining 585,562 human and 203,175 mouse "non-concordant" cRE result from high DNase signal in one cell type and high H3K4me3, H3K27ac, or CTCF signals in a different cell type. As more data become available, we anticipate that many of the non-concordant cREs will move into the concordant class, and will be updated to reflect that.

cREs are further designated as TSS proximal if they lie within ±2 kb of a GENCODE-annotated TSS. There are 242,739 TSS-proximal cREs in human and 92,405 in mouse. The cREs that overlap a TSS are called TSS-containing cREs; there are 46,749 and 24,549 TSS-containing cREs in human and mouse respectively. TSS-overlapping cREs are significantly longer than the rest of the TSS-proximal cREs and TSS-distal cREs (median length = 548, 317, 342 for human and 589, 320, 339 for mouse; Wilcoxon test p-values < 2.2E-16 for all tests).

II.4.8 Comprehensiveness of the current Registry of cREs

In defining the Registry of cREs based on rDHSs, our working hypothesis is that a collection of rDHSs derived from hundreds of DNase-seq experiments will represent a large fraction of all cREs in the genome and that a new cell type is likely to use as its cRE repertoire a subset of the cREs 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 current version of the Registry of cREs. To the extent that GENCODE is a mature repository of expressed RNAs across all cell types and states in the human and mouse life cycle, this test provides an informative estimate for the completeness of promoters and promoter-proximal regulatory elements in our Registry. For 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 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 overlap a cRE in the Registry.

Second, we analyzed how rapidly the total number of unique 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. The saturation curves of rDHSs continue to follow Weibull distributions, revealing at the plateau 1.66 M rDHSs with FDR < 0.1% and Z-score > 1.64. Because only a subset of such rDHSs can be cREs—those with a high H3K4me3, H3K27ac, or CTCF Z-score in at least one cell type—we have identified at least 78.9% cREs in human. We performed the same saturation analysis for mouse but could not reach a reliable estimate due to the smaller number of input tissue types.

Third, we computed the Registry's coverage of H3K27ac, H3K4me3, and CTCF peaks (FDR<0.01) 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). 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. (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. 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.

In conclusion, the human Registry appears to be comprehensive: by the above criteria, it covers two-thirds of all cREs and 85% of elements marked by H3K4me3 or H3K27ac or bound by CTCF in any cell type. A cautionary note is that we do not yet know the extent of coverage on highly cell-type-specific cREs 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.

II.4.9 Classifying cREs 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 cREs based on the maximal Z-score of each feature across all cell types with ENCODE and Roadmap data, abbreviated henceforth as max-Z. The goal is to provide a useful overview of the entire cRE landscape by integrating all input cell types for the four epigenetic features. We then classify cREs according to these four features at two levels of detail—the state classification and group classification—described below in turn.

As described above, all cREs must have a high DNase max-Z and furthermore 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 (max-Z ≥ 1.64) or low (max-Z < 1.64) H3K4me3, H3K27ac, and CTCF signals, with each combination called a state. This classification captures the fact that while some cREs are marked by just one high signal (41% of human and 59% of mouse cRE), many cREs have two or three high signals (Figure II-14 on page 100). Because the all-low state is not allowed, a cRE can adopt one of seven states. Furthermore, each cRE is classified as being proximal or distal (within or outside the ±2 kb window) to the nearest GENCODE-annotated TSS. The state classification simply indicates which of the seven states a cRE is in and is displayed in SCREEN with a color code alongside the information on TSS proximity and whether the cRE is supported by concordant signals from the same cell type.

The group classification is an abbreviated abstraction that assigns each cRE to a group according to its biochemically dominant signature. As reported above for transgenic mouse enhancer assays, the intensity of biochemical signals is positively but modestly predictive of functional enhancer activity. We define broad, mutually exclusive groups of elements in the expectation that they will be enriched in the respective promoter-like, enhancer-like, or CTCF-mediated functions. We currently use three groups assigned in the following order (Figure II-13 on page 99):

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 groups and thus by definition must have high CTCF max-Zs to qualify as cREs.

Classifications are assigned in the above order; thus, a cRE possessing high histone mark signals and a high CTCF signal will be classified as either PLS or ELS. This simplified classification scheme is designed to give users a first-cut idea of the most likely function for each cRE, 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 SNP that interrupts the binding is associated with severe influenza risk in humans

(Allen et al. 2017). Massively parallel reporter assays indicate that some promoters also have enhancer activities while some enhancers also have promoter activities (Nguyen et al. 2016). A tiling-deletion-based CRISPR screen of the 2-Mb *POU5F1* locus identified 45 cis-regulatory elements, among which 17 are promoters of functionally unrelated genes (Diao et al. 2017). CapStarr-seq data revealed that 2-3% of the coding-gene promoters display enhancer activity in a given mammalian cell line (Diao et al. 2017). Thus, the group classification is intended to ease analysis and simplify discussion, and we emphasize that many cREs belong to multiple groups.

As currently formulated, the Registry does not explicitly define negative elements, but we aim to include them in the next version of the Registry. We note that some of the cREs 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, attracting repressive epigenetic regulator such as Polycomb group proteins, or establishing 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 et al. 2017). Such findings underscore the notion that many cREs belong to multiple groups.

We analyzed the fraction 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. The lower coverage for mouse is due to the smaller number of cell types with data available with which to define cREs.

The state classification scheme extends naturally to a specific cell type, by characterizing the biochemical activity of each cRE with the DNase, H3K4me3, H3K27ac, and CTCF data in that cell type. 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 H3K4me3, H3K27ac, and CTCF Z-scores.

The group classification scheme also extends naturally to a specific cell type, but two additional groups are needed: an inactive group, containing all cREs with low DNase Z-scores, and a DNase-only group, containing cREs with high DNase Z-scores but low H3K4me3, low H3K27ac, and low CTCF Z-scores within that cell type.

Using GM12878 lymphoblastoid cells as an example, Figure II-15 on page 101 summarises the cREs states, with each state further stratified by TSS proximity. The bar graph reveals that cREs with high H3K4me3 are mostly TSS-proximal, regardless of whether or not 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 the group classification of cREs in this cell type (Figure II-14 on page 100): 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. cREs-PLS have the highest median Pol II signal, cREs-ELS have the highest median EP300 signal, and CTCF-only cREs have the highest median RAD21 signal.

II.4.10 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. The higher abundance of cREs-PLS over cREs-ELS may be surprising, given the widely held perception that enhancers outnumber promoters. It is important to note that different cell types share many promoters but they share far fewer enhancers. Indeed, we identify far more cREs-ELS than cREs-PLS across all available cell types (991,173 vs. 254,880, respectively). Also, many of the cREs-PLS could be TSS-proximal enhancers. Among the 34K TSS-proximal cREs-PLS, over 14K directly overlap a TSS, making them strong candidates to contain promoters active in GM12878. The remaining 20K lack a known TSS, and these are likely to be a mixture of active promoters and enhancers. Enhancers near basal promoters are common in gene architecture, effectively creating a functional continuum (Nguyen et al. 2016) that are likely to carry both enhancer and promoter-like marks. Furthermore, the H3K4me3 histone mark is known to spread around active TSSs, and thus a TSS-proximal regulatory region, including an enhancer, can be reproducibly marked by H3K4me3 but still lack promoter activity. Users can inspect the contributing signals individually in the state classification and also integrate other richer data-types, including RNA-seq and RAMPAGE, to refine their understanding of these cREs-PLS and cREs-ELS.

II.4.11 Comparison between cREs and the corresponding ChromHMM states

As described above, there are two approaches to building catalogues 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 incorporates more histone marks.

The cREs-PLS and cREs-ELS in GM12878 are consistent with the respective chromatin states called by ChromHMM using eight histone marks and CTCF in this cell type (Consortium 2012). Figure II-16a on 102 shows that 90% of top cREs-PLS (ranked by H3K4me3 Z-scores) overlap with ChromHMM promoters. Figure II-16b on 102 reveals that over 85% of the top cREs-ELS (ranked by H3K27ac Z-scores) overlap with ChromHMM high-signal enhancers. The overlap decreases for lower ranking cREs-ELS, but the overlap with ChromHMM low-signal enhancers increases; 82% 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, as described in a companion paper (Tsuji et al., in preparation). We observed that 95 ± 2% of cREs-PLS overlapped ChromHMM-annotated promoters and 78 ± 3% of cREs-ELS overlapped ChromHMM-annotated enhancers in the corresponding tissue and time point.

II.4.12 Cell and tissue type clustering

To examine whether the Registry of cREs captured the regulatory landscapes, we clustered primary cell types and tissues on the basis of 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 (see online Extended Data Fig. 13-15), in agreement with findings from a previous report (Stergachis et al. 2013). As described in a companion paper, we analyzed gene expression data in a panel of primary cells and uncovered four clusters of primary cells with similar expression patterns, and we related these expression patterns to those of histological tissues (Breschi et al., in preparation).

II.5 SCREEN: A Web Engine for Searching and Visualizing cREs

With millions of cREs in the Registry, providing an easy access to these data to the end users is the next challenge. ENCODE has four major goals in making these annotations available to end users:

1. dynamic and interactive interfaces: provide a search interface allowing users to filter cREs in real time without leaving or reloading the search page, and display dynamically-generated analyses related to subsets of cREs with interactive plots
2. integrated: provide the ability to view all the low-level annotations, such as transcription factor ChIP-seq, RNA-seq, RAMPAGE, and histone ChIP-seq results, associated with cREs through one interface, and make customizable visualizations of these annotations easily accessible in the UCSC Genome Browser
3. reproducible: provide dynamic and interactive plots for published figures relating to the cREs within the web interface, and facilitate users’ generation of these figures with custom data

4. extensible: provide for the inclusion of analyses involving external annotations and user-submitted annotations

Meeting these goals posed numerous challenges for the existing ENCODE Portal architecture. The scale of the annotations far exceeds the scale of the datasets and experiments cataloged by the Portal, which number in the tens of thousands, posing challenges both for accessioning of the annotations and for rapid dynamic searching. Further, the dataset-centric model of the portal is not easily generalized to the concepts of genomic coordinates, activity profiles, and the like, nor to analysis involving intersection with large collections of peaks or with external datasets. Finally, the Portal is not equipped to accept external or user-provided data for the generation of analysis plots.

Numerous tools exist for cataloging epigenomic annotations and regulatory elements; however, to our knowledge, none of the existing tools catalog epigenomic annotations on the scale of the ENCODE Encyclopedia, nor provide the feature set required to meet ENCODE's goals for visualization of the Encyclopedia's annotations. The Broad Institute's HaploReg provides an interface for searching SNPs and associated annotations and can tailor results to uploaded user input; however, the search results may not be dynamically filtered and no analytical plots are provided (Ward and Kellis 2012). The WashU and Roadmap Epigenome Browsers provide rich feature sets for dynamic plot generation and exploration of low-level annotations, but they are not engineered for cataloging or searching higher-level annotations like the cREs (Zhou et al. 2011). DENdb (Ashoor et al. 2015) and Enhancer Atlas (Gao et al. 2016) both provide enhancer databases, but the former does not incorporate visualizations of search results, the latter provides only static searching, and neither provides for the display of peak intersections or element overlap with external datasets.

GeneCards (Stelzer et al. 2002) and Ensembl (Yates et al. 2016) provide more complete epigenomic catalogs. GeneCards offers detailed information on individual genes, complete with a list of candidate enhancers predicted to regulate them, provided by the recently-published GeneHancer feature (Fishilevich et al. 2017) which integrates enhancers from various sources including ENCODE enhancer predictions. Similarly, the Ensembl Regulatory Build (Zerbino et al. 2015) uses Segway to create genome-wide epigenomic annotations on the basis of DNase-seq, CTCF ChIP-seq, and various histone ChIP-seq results; the database containing these annotations is directly accessible via the BioMart web UI and the annotations are also presented in association with SNPs. Both these projects provide resources similar to the ENCODE Encyclopedia, but the Encyclopedia expands on both in scale, with 1.3 million human cREs compared to approximately 285,000 human regulatory elements in GeneHancer and approximately 447,000 elements in the Ensembl Regulatory Build. Additionally, GeneCards is predominantly gene-centric and Ensembl is predominantly SNP-centric; although both provide locus-based searches for regulatory elements, neither provides the rich, regulatory-element-centric features ENCODE aims to incorporate into its Encyclopedia viewer, particularly the ability to filter regulatory elements by activity within particular cell types and to view intersecting low-level annotations from the thousands of ENCODE ChIP-seq experiments. Additionally, although both tools are well-suited for identifying regulatory elements associated with particular genes and SNPs of interest, they do not provide features allowing for coanalysis of external datasets.

In order to meet our goals for cRE visualization, we designed an original web-based visualizer, SCREEN (*Search Candidate Regulatory Elements by ENCODE*). SCREEN is divided into three *apps*, each providing a unique perspective on the cREs and a unique approach to searching and interpreting them. The locus-centric search app provides a keyword search and a dynamic filtering interface for browsing cREs both by genomic coordinates and activity profiles across available cell types. The gene-centric expression app provides dynamic plots of gene and TSS expression data derived from ENCODE RNA-seq and RAMPAGE datasets, as well as a differential expression plot which compares differential gene expression with differential activity of nearby cREs. Finally, the SNP-centric GWAS app is SCREEN's first app involving external datasets, displaying the interaction of SNPs from GWAS studies and cREs.

In order to catalog all the Encyclopedia's annotations and provide rapid, efficient searching, SCREEN utilizes two separate databases and an extensive pre-processing pipeline implemented in C++ and Python. SCREEN's frontend takes advantage of HTML5's support for embedded vector graphics to produce dynamically-generated and interactive plots which reproduce published figures relating to the cREs; these figures provide easy access to the underlying annotations, update automatically to include new data as they become available at ENCODE, and are extensible to support external and user-submitted annotation sets. SCREEN's Genome Browser configuration feature allows users to select custom subsets of cell types and pre-loaded cRE-related tracks to view in the UCSC Genome Browser, making visualization of both low-level and integrated annotations dramatically easier. Finally, SCREEN's design permits easy extension to new data types and analyses via the addition of new apps. Together, these features represent a new paradigm for the dynamic visualization and analysis of epigenetic annotations genome-wide.

II.5.1 SCREEN Methods

II.5.1.1 General architecture

SCREEN consists of a CherryPy-based webserver backed by PostgreSQL and Apache Cassandra for data storage (Figure II-17 on page 103). Its front-end UI is based upon the ReactJS Javascript library and utilizes the Redux Javsacript library to manage application state. The front-end is written in ES6 Javascript; package management and compilation to pure Javascript is handled by Facebook's yarn. In order to permit dynamic searching and visualizations, page updates after the initial page load are handled via JSON-based AJAX requests. Plots are rendered dynamically on the client side as scalable vector graphics (SVG). Computationally expensive operations are pre-computed using a C++ and Python pipeline.

II.5.1.2 Storage of cREs and supporting analyses

Storing the cREs, performing searches, and retrieving requested subsets are the core challenges underlying SCREEN's design. SCREEN aims to do all three dynamically: as ENCODE data expands, so must SCREEN's capacity to catalog regulatory elements and cell types, yet SCREEN's storage system must be able to respond to user queries in real time to support filtering of cREs without a full page reload. In order to support both dynamic expansion and dynamic searching, we use PostgreSQL, an open-source object-relational database system, to store the cREs. Postgresql's integer range operations allow for rapid searching by genomic coordinates, its array data types allow for scaling of cell-type-specific data points as new cell types are added, and its JSON and binary JSON data types facilitate representation of nested metadata and precomputed parameters for SVG plots and visualizations. The core of the database schema, showing the representation of cREs with associated gene and cell type information, is presented in Figure II-18 on page 104. Fields marked with a capital **I** are indexed; indexing cREs by chromosome dramatically improves search performance when searching for cREs on a single chromosome, and indexing the cREs by their maximum Z-scores allows rapid searching for elements of a particular activity profile, such as strong promoters or enhancers, across all cell types. We also use PostgreSQL to store several supporting datasets, including RAMPAGE and RNA-seq results, differential gene expression datapoints, TAD information, and cRE-SNP intersection.

II.5.1.3 Ground level annotations and pre-computed analyses

Analyses involving the ground-level annotations are too large scale to perform dynamically. Reading signal and peaks directly from BigWig and Bed format files, and performing intersections between cREs and hundreds of millions of peaks, is prohibitively slow. Because these analyses are highly computationally expensive, we designed an extensive pre-processing pipeline, written predominantly in C++ with some supporting Python; this pipeline performs analyses involving the cREs and ground-level annotations, and then formats the output for import into SCREEN's databases. The pipeline utilizes OpenMP for parallelization of its most expensive analyses, and utilizing a 64-core server with 512 GB of RAM the pipeline can complete its computations in approximately 6 hours. This allows SCREEN to be rapidly updated as new data become available at ENCODE.

Overlap between cREs and peaks from all available ENCODE transcription factor and histone ChIP-seq experiments is pre-computed by SCREEN's *peak_intersection* package, a Python package utilizing bedtools (Quinlan 2014). Support for intersection with peaks from Cistrome (Liu et al. 2011) is also included by default, and the package is easily extensible to other external datasets as well. SCREEN's Signal Profile display provides snapshots of the signal from H3K4me3, H3K27ac, and CTCF ChIP-seq experiments in the region surrounding a selected cRE. These signal snapshots are pre-computed by SCREEN's *minipeaks* package, which down-samples and bins signal extracted from downloaded ENCODE signal files in BigWig format. The *minipeaks* package also utilizes *ZentLib*, a C++ wrapper of ENCODE's kentUtils (Kent et al. 2010), to process signal files.

The output from the *peak_intersection* package is stored in PostgreSQL and is available to the user in real time. The output from the *Minipeaks* package, however, includes hundreds of signal datapoints across thousands of experiments, and is too large scale to permit storage, indexing, and querying using PostgreSQL. Instead, the signal values are stored using Apache Cassandra, an open-source NoSQL data store offering linear scalability. Cassandra allows SCREEN to present snapshots of the signal within a cRE and the associated summit signal values in real-time; these profiles are rendered dynamically on the client side using SVG.

II.5.1.4 Web server

SCREEN's web server is implemented in Python and is based on the CherryPy framework. Support for dynamic filtering of search results and generation of figures requires numerous AJAX requests and large volumes of data to be transferred between the server and the client; to support this volume, the current production site consists of several simultaneous Docker instances. SCREEN supports sharing of necessary state information between Docker instances using *redis*, an in-memory data store.

The server employs a model-view-controller design, which separates the logic of processing data, rendering content to HTML, and handling user input; this improves both security and code maintainability. The controllers primarily receive user input from CherryPy and retrieve data from the database models; the AJAX service, for example, consists primarily of the *data_ws* controller. The models perform any processing and reformatting necessary before data may be returned and rendered by the client. The *gene_expression* model, for example, pre-sorts expression data from the database according to several different criteria and performs log-transformation before the data are returned to the client; this reduces the amount of processing necessary within the user's web browser to improve page responsiveness, and also allows SCREEN's backend to incorporate caching mechanisms to reduce database queries and thus the time required to perform searches. The view component of the controller handles initial page requests only, which are rendered from templates using Jinja2; subsequent updates via AJAX are handled dynamically on the client side by ReactJS and Redux, which allows search results and plots to update without requiring a full page reload.

SCREEN's web user interface code is written in ES6 Javascript; package management and compilation of the ES6 code into pure Javascript are performed by Webpack. ES6 provides numerous advantages over pure Javascript, including better cross-browser compatibility, standardized incorporation of external libraries, and support for embedded HTML and rich object-oriented code. Currently, the server outputs a static Javascript bundle which is served by CherryPy; future work aims to take full advantage of the hot module reloading functionality provided by Webpack.

II.5.1.5 User interface and dynamic plots

In order to provide dynamic searching and plot generation, SCREEN's user interface is rendered nearly entirely on the client side using ReactJS, an open-source library for building component-based applications. ReactJS handles updates efficiently using a virtual representation of the web page's browser's internal *document object model* (DOM), which incorporates embedded SVG figures as well. This efficient update model is critical for SCREEN's ability to update its dynamically-generated plots, some of which render several thousand datapoints, with little to no observable delay for the user.

Because of the complexity of SCREEN's state, we use the Redux library to handle component-generated actions. Redux provides a central *store* which manages the application's state; components *connect* to the store to receive properties and may *dispatch* actions to the store to update the state. Actions dispatched to the store are handled by the store's *reducers*, which operate on the current state to produce a new state. Reducers never alter the current state directly in order to prevent race conditions, and components never communicate with each other directly, which improves code maintainability. The use of Redux also allows SCREEN’s dynamic plots to interact with other components of the page, allowing datapoints within figures to link directly to more information about their underlying annotations, for example.

II.5.1.6 Custom trackhubs

In addition to its own visualizations, SCREEN aims to provide easy access to the full collection of ENCODE’s ground-level annotations. SCREEN provides custom trackhubs for the UCSC Genome Browser (Kent et al. 2002), which include tracks for every ENCODE DNase-seq, H3K4me3 ChIP-seq, and H3K27ac ChIP-seq dataset available. SCREEN offers a configuration view to allow users to select which cell types to view and to create custom orderings of the tracks before being redirected to the Genome Browser. Tracks are also available in two separate formats for visualizing the cREs. The 9-state format displays four tracks, one each for DNase-seq, H3K4me3 ChIP-seq, H3K27ac ChIP-seq, and CTCF ChIP-seq Z-scores; cREs with a Z-score >1.64 for a given mark are colored and the remaining cREs are gray. The 5-group format displays a single condensed track, with active cREs colored according to their activity (red for promoter-like cREs, yellow for enhancer-like cREs, etc.) and inactive cREs colored gray. Trackhubs are accessed via buttons available across SCREEN. Future plans include expansion of the tracks to include other data types, such as transcription factor ChIP-seq, as well as external data sources such as Cistrome.

II.5.2 SCREEN Usage

Usage of SCREEN begins at the homepage, which provides access to the locus-centric search app (see Figure II-19 on page 105) and the gene-centric expression app (see Figure II-20 on page 106) via a keyword search box and the SNP-centric GWAS app via a "browse GWAS" button.

Here we present a use case of SCREEN which explores how a user might use the GWAS app to form a hypothesis about the functional role of a SNP in a particular disease state. Clicking the "browse GWAS" button on the SCREEN homepage produces the GWAS app (Figure II-21 on page 107). The user first selects a GWAS study of interest; we have selected a 2012 study on inflammatory bowel disease for illustration (Jostins et al. 2012). The app will display how many SNPs from the study, as well as SNPs in linkage disequilibrium with SNPs from the study, overlap cREs (Figure II-22 on page 108), along with a list of the ENCODE cell types with the most active overlapping cREs. In the case of the selected inflammatory bowel disease study, the top ten cell types include nine leukocyte cell types, as might be expected given IBD's autoimmune nature, along with a cell type from the rectal mucosa. When the user clicks to select a cell type, a list of the cell type's active cREs is displayed; we select the top cell type, a T-cell line from an adult male donor, for illustration (Figure II-23 on page 109).

The first search result is cRE EH37E1089569, overlapping SNP rs11041476. The symbols in the left two columns reveal that the region surrounding this cRE is enriched in H3K4me3 ChIP-seq, H3K27ac ChIP-seq, and CTCF ChIP-seq signal in at least one ENCODE cell type each, and that the former two are enriched in the selected T-cell line.

The “P” symbol indicates that this cRE is within 2kb of a TSS for *LSP1*; following the *LSP1* link in the second column from the right to GeneCards reveals that this gene is thought to play a role in immune cell chemotaxis, adherence to matrix proteins, and migration through the epithelium, and following the rs11041476 link to Ensembl reveals that the A allele of this SNP is correlated with *LSP1* expression changes. Clicking the UCSC button in the far right column allows the user to visualize the region surrounding the cRE in the UCSC genome browser. Raw signal and cRE annotations are available for all ENCODE datasets, and subsets may be selected using the configuration view (Figure II-24 on page 110, top); here, we have selected GM12878 and T-cells from an adult male donor. Visualized in the UCSC Genome Browser (Figure II-24 on page 110, bottom), the data suggest that the cRE represents methylation of the histone adjacent to the first promoter for *LSP1*, resulting in increased chromatin accessibility for transcription.

To obtain more information about EH37E1089569, the user may click the link in the leftmost column, which produces the search results page for EH37E1089569 (Figure II-25a on page 111). Clicking the cRE’s row in the results table produces the details view. The default view displays the cRE’s Z-scores for the four core marks across all available cell types, which are highest in various immune cell types (H3K4me3 and CTCF shown). The user may use the signal profile tab at the far right to view the raw data contributing to these Z-scores as well (Figure II-25b on page 111). The user may use the *Nearby Genomic Features* tab to identify the nearest genes, SNPs, and other cREs to the selected cRE, and the *TF and his-mod intersection* tab to view ENCODE transcription factor ChIP-seq and histone mark ChIP-seq experiments with peaks intersecting the cRE (not shown).

The *Associated Gene Expression* tab displays a component of the expression app, here showing *LSP1* expression. Grouping by tissue indicates that this gene is most strongly expressed in immune cells (*blood* tissue) as well as the adrenal gland and spleen according to ENCODE RNA-seq results (Figure II-26 on page 112). The *RAMPAGE* tab gives greater insight into this expression profile by displaying transcription activity at all of LSP1's transcription start sites; grouping RAMPAGE signal by tissue max suggests that the TSS closest to EH37E1089569 is most strongly expressed in the spleen, with adrenal *LSP1* expression arising predominantly from other TSSs.

Together, these results suggest that EH37E1089569 is the first promoter for *LSP1*, that the corresponding TSS is most active in immune cells and tissues, and that an A allele at rs11041476 likely impacts expression of isoforms of *LSP1* which include the first exon, which are likely expressed predominantly in immune tissues. If the user is interested in studying *Lsp1* in mouse, for example as a component of a mouse IBD model, the *orthologous cREs in mm10* tab reveals that there is an orthologous cRE, EM10E0419598, in mouse. A similar workflow reveals that this cRE as well as *Lsp1* have similar activity profiles within immune cells as their human counterparts.

Also available in SCREEN mm10 is the ability to duplicate the differential gene expression plot for developmental tissues, accessed via delta symbols adjacent to gene names in the search results table; for further discussion, see "II.6.1 Comparing cREs across mouse developmental timepoints" (below) and Figure II-27 on page 113.

II.5.3 Testing SCREEN User Interface

The user interface of SCREEN went through many iterations. To gauge usability, we developed a series of tasks we requested beta testing users to complete. The tests covered as much of the functionality of SCREEN as possible. We requested feedback on how difficult each task was to complete; if the tasks were unclear; if finding the functionalities in SCREEN for performing the task were difficulty, and how many attempts were made to perform the task. As recommended by Paul Flicek at Ensembl, users were encouraged to record their SCREEN session using a screen capture application, and, optionally, to also record a user voice-over as the user talked out-loud, to future give insight on the users' experience.

The user tasks were as follows:

- Visit the ENCODE Encyclopedia page
    - Do you see a description about the Registry of candidate Regulatory Elements (cREs)?
    - Do you see an entry to SCREEN?
    - What is the relationship between the ENCODE Encyclopedia, the Registry of cREs, and SCREEN?
- Use SCREEN to find all candidate Regulatory Elements (cREs) in the beta globin locus by its genomic location in the human genome (build hg19): chr11:5226493-5403124.
    - How many cREs are there?

  - Download these cREs as a file in the comma-separated-values (CSV) format (an Excel friendly format).
- Identify the cRE in the beta globin locus (identified in task 1) that has a DNase Z-score > 1.64 AND the highest H3K27ac Z-score in K562.
  - What is the accession for this cRE?
  - Is this cRE classified as a promoter-like cRE, enhancer-like cRE, or CTCF-only cRE?
  - What is its nearest protein coding gene? How far is this gene?
  - In which tissues does this cRE have the highest DNase Z-score?
- Find all the cREs within the gene body of human *Actin alpha 1*.
  - How many cREs are there?
  - What is the official gene symbol of *Actin alpha 1*?
  - Can you find all the cREs within 25 kb upstream of *Actin alpha 1* transcription start site?
- Examine the expression profile of *HNF1A* across cell types in human.
  - Can you choose only tissues, and not other types of samples such as primary cells or cell lines?
  - Can you ask SCREEN to display expressions only in the nuclear compartment?
- Examine the activity profiles of the multiple transcription start sites (TSSs) of *HNF1A* across cell types in human, measured by RAMPAGE.

  - Do some TSSs of *HNF1A* have different activity profiles than other TSSs of this gene?
  - Are the activity profiles for some of *HNF1A*'s TSSs similar to the expression profile of the *HNF1A* gene?
- Use SCREEN to find the human cRE with accession EH37E0579839.
  - Can you find its "Signal Profile", i.e., cropped out signal peaks across cell types?
  - Can you find other cREs that are within a topologically associated domain (TAD) of this cRE?
  - Which transcription factors bind to this cRE? In which cell types?
  - How many H3K4me1 ChIP-seq experiments have overlapping peaks at this cRE? In which cell types are these ChIP-seq experiments?
- Use SCREEN to find the human cRE with accession EH37E0579839.
  - What is the nearest human gene of this cRE?
  - Can you find the mouse ortholog of this human cRE? What is the nearest mouse gene of the mouse cRE? Is the mouse gene the ortholog of the human gene?
  - Are the cRE classifications (i.e. promoter-like cRE, enhancer-like cRE, or CTCF-only) consistent between human and mouse?
  - Do the two orthologous genes have similar expression profiles across cell types between human and mouse?
- Compare the differential expression levels of the mouse *Ogn* gene.

    - Compare embryonic day 11.5 limb vs. embryonic day 15.5 limb.
    - Compare between limb and forebrain, both at embryonic day 14.5.
- Search SCREEN twice, with the first time searching for cREs in the beta globin locus (human genome build hg19, chr11:5226493-5403124) and the second time searching for cREs in the *HNF1A* gene body.
    - Can you download the cREs that you found in these two searches in two CSV files (one for each search)?
    - Can you download the cREs that you found in these two searches in a single CSV file?
- Browse the list of genome-wide association studies (GWAS) in SCREEN and find the study on QT interval by Arkin et al.
    - What is the cell type with enhancer-like cREs that most significantly overlaps SNPs identified in this GWAS?
- Please read the About page and watch the tutorial videos.
    - Are they informative?
    - Is there additional information we should add?
- Please feel free to browse and give other comments. Is SCREEN easy enough to use? What improvements would you like to see? Which ones are essential to have before this is launched?

II.6 Use Cases of the Encode Encyclopedia and SCREEN

We foresee many applications for the ENCODE Encyclopedia and SCREEN. The various annotations at the ground level of the Encyclopedia can be downloaded from the

ENCODE Portal and further analyzed along with users' own data. SCREEN allows users 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 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.

II.6.1 Comparing cREs across mouse 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 mouse embryonic day 12 (Taher et al. 2011). SCREEN displays *Ogn* and nearby differentially expressed genes in the limb between e11.5 and e15.5 (identified using DESeq2 (Love, Huber, and Anders 2014), 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 (Figure II-27a on page 113). 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 to identify cREs that might account for the increase in *Ogn* expression—specifically, cREs proximal to *Ogn* are likely to play a role in regulation, because their increase in signal is 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 (Figure II-28 on page 114), which is correlated with increases in H3K27ac, H3K4me4, and DNase signals. *Ogn* expression increases most notably after e12.5, in agreement with previous findings (Taher et al. 2011). This increase in gene expression correlates with the increases in H3K27ac and H3K4me3 Z-scores of nearby cREs (Figure II-27c on page 113).

II.6.2 Using the Registry of cREs to annotate GWAS SNPs

Previous studies have repeatedly demonstrated that most GWAS variants reside outside exons. Furthermore, independent annotation of noncoding regions can be used to guide the interpretation of GWAS variants by predicting disease-relevant cell types and regulatory factors (Ernst et al. 2011; Maurano et al. 2012; Consortium 2012; Andersson et al. 2014; Farh et al. 2015; Dickel et al. 2016). With the broad coverage of cell types and rich epigenetic and transcription factor binding data associated with the cREs, the Registry can be particularly useful for annotating GWAS SNPs.

To facilitate GWAS exploration, we have preloaded SCREEN with a subset of studies from the NHGRI-EBI GWAS catalogue (Hindorff et al. 2009; MacArthur et al.

2017) that were performed on the Caucasian-European (CEU) population (see online Supplementary Table 7), and we plan to include other populations in the near future. For each GWAS, we tested each cell type for whether its set of cREs-ELS was significantly enriched in the GWAS SNPs after accounting for SNPs in linkage disequilibrium (LD). SCREEN displays the cell types in descending order of enrichment, and users can browse the cREs in each cell type that overlap with GWAS SNPs. Figure II-29 on page 115 shows a heat map of the enriched cell types for a subset of GWAS, and the results are summarized in Figure II-30 on 116.

The user can first select a GWAS study of interest (Figure II-21 on page 107 and Figure II-31a on page 117), 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 (Figure II-31b on page 117). A list of cell and tissue types is provided on the basis of enrichment in the H3K27ac signal. The user can narrow the search by selecting a cell type, such as GM12878 for multiple sclerosis, the left ventricle tissue for QT interval, or HepG2 for cholesterol levels (Figure II-31c on page 117). After a cell type is selected, SCREEN updates 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 with promoter-like or enhancer-like signatures (Figure II-31d on page 117). SCREEN also returns a list of SNPs for users to search and view in a genome browser along with the cRE and other supporting data, thus aiding in fine annotation of the SNPs and prediction of their functional impact.

As an example, rs1250568 is in LD (r^2=0.7) with two SNPs associated with multiple sclerosis, rs1250542 (Patsopoulos et al. 2011) and rs1250540 (De Jager et al. 2009). rs1250568 is predicted to be a causal SNP by the deltaSVM algorithm (Lee et al. 2015). GM12878 has previously been suggested to be a relevant cell type for multiple sclerosis (Maurano et al. 2015), and SCREEN computes 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 (Figure II-31d on page 117). It overlaps a ChIP-seq peak for the transcription factor ELF1 and disrupts an ELF1 motif site (Figure II-32e on page 118). 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 et al. 1992; Vaknin-Dembinsky, Balashov, and Weiner 2006). 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 than in controls (Fewings et al. 2017). *ZMIZ1* is highly expressed in neurons and cardiac muscle cells (Figure II-33 on page 119) and has been reported in the GWAS but *PPIF* has not (Patsopoulos et al. 2011; De Jager et al. 2009). *PPIF* encodes a mitochondrial permeability transition pore protein and is expressed in cardiac muscle cells, lymphocytes and hepatocytes (Figure II-34 on page 120). We predict that *PPIF* performs functions in lymphocytes which are associated with the demyelination of neighboring neurons. Knockdown or knockout of *Ppif* leads to neuroprotective effects in murine disease models of multiple sclerosis (Forte et al. 2007; Warne et al. 2016). 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 cREs and SNPs may cause the disease.

II.6.3 Combining orthologous cREs to fine-map GWAS SNPs

One particular strength of the Registry is its inclusion of both human and mouse cREs and the definition of orthologous cREs in these two species. Mouse cREs are mostly defined using tissues during embryonic development; 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 affect the brain.

For example, rs13025591 has been reported by two studies to be associated with schizophrenia (p-values 8E-8 and 6E-6) (2011; Bergen et al. 2012). rs13025591 lies in the intron of the *AGAP1* gene, and 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 (Figure II-35 on page 121). rs13025591 does not lie within a cRE. Therefore, we hypothesized that the signal driving this genetic association arises from SNPs in high LD with rs13025591. There are five cREs that overlap such SNPs (Figure II-36 on page 122), four of these cREs show enhancer-like signatures and one shows a promoter-like signature. None of the five cREs 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 et al. 2000; Nierenberg et al. 2005; Weinberger, Berman, and Zec 1986); nevertheless, EH37E0579839 has high H3K27ac signals 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 tissues, but the signals disappear in older fetal brain and adult brain tissues (Figure II-37b on page 123). EH37E0579839 is orthologous to the mouse cRE EM10E0042108, which shows enhancer-like signatures in brain tissues. Consistently with its human ortholog, EM10E0042108 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 (Figure II-38 on page 124). In the forebrain, midbrain, and hindbrain, H3K27ac signals increase over time, reaching a maximum at e13.5. Then, similarly to those of the human ortholog, H3K27ac signals at the cRE decrease after this time-point through birth (Figure II-37c on page 123). These results indicate that this cRE is active only during a narrow window of brain development.

The region harboring these two orthologous cREs is conserved across mammals (Figure II-37d on page 123). Although we do not have TF ChIP-seq data in fetal brain or mouse embryonic brain tissues, motif analysis using both HaploReg and RegulomeDB (Ward and Kellis 2012; Boyle et al. 2012) 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 using

SCREEN and the ENCODE Encyclopedia, we were able to narrow down a region to guide experimental testing for biological function.

II.7 Methods

II.7.1 Identifying rDHSs

We used all DNase-seq datasets as of February 1, 2017 with HOTSPOT2 calls on the hg19 or mm10 genomes (see online Supplementary Table S8). 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. This pipeline is available on GitHub (Create-rDHSs.sh).

II.7.2 Normalizing epigenomic signals

For each rDHS, we computed the Z-scores of the log of DNase, H3K4me3, H3K27ac, and CTCF signals. Z-score computation is necessary for the signals to be comparable across all cell and tissue types, because the uniform processing pipelines of 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 100k randomly selected rDHSs across five different cell-types (Figure II-39a on page 125). 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 of the signals (Figure II-39b on page 125), we observed that the distribution in each dataset roughly follows a normal distribution. The Z-scores of log(signal) values have the same distribution across cell types (Figure II-39c on page 125).

To implement this normalization, we used the UCSC tool *BigWigAverageOverBed* to compute the signal for each cRE (averaged across the entire cRE for DNase and CTCF and across the entire cRE plus 500 base-pairs on each end for H3K4me3 and H3K27ac), and, 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. This pipeline is available at GitHub (Process-rDHS-Signals.sh).

II.7.3 Saturation analysis of rDHSs

To determine the percentage of all possible rDHSs that have been sampled using our 440 DNA-seq datasets, we used a modified approach from ENCODE Phase II. We randomly selected *X* cell types, where *X* is between 10–440 in intervals of 10. We then selected all corresponding DHSs for these cell types (including their biological replicates) and calculated the number of resulting rDHS using the rDHS selection pipeline (described above). Adapting the R script by Steven Wilder and Ian Dunham (Consortium 2012), we calculated the complete set of rDHSs to be at 95% saturation for each curve using a Weibull distribution.

II.7.4 Overlap of cREs in cell types without DNase-seq data

To determine the comprehensiveness of the Registry, we overlapped cREs with ChIP-seq peaks (H3K4me3, H3K27ac, and CTCF) from cell types lacking DNase data. Using bedtools merge, we merged all ChIP-seq peaks within 200 bp of one another and assigned each merged peak the maximal -log(FDR) score of the original peaks. We then filtered out all peaks with -log(FDR) < 2. Using bedtools intersect with the "-u" flag, we intersected the merged peaks with cREs and counted the number of unique peaks that overlapped at least one cRE. This pipeline is available at GitHub (Calculate-Peak-Overlap.sh).

II.7.5 Classifying cREs

For cell type agnostic classification of cREs 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 Figure II-13 on page 99. The rDHSs that were classified as having promoter-like, enhancer-like, or CTCF-only signatures are deemed cREs and assigned an accession; the rDHS that are not classified in any of these categories are discarded. This pipeline is available at GitHub (Create-cREs.sh).

To classify cREs 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 cREs 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: cRE-PLS, cRE-ELS, and CTCF-only are assigned according to the classification trees, the DNase-only group contains cREs with high DNase Z-scores but low H3K4me3, H3K27ac, and CTCF Z-scores, and the inactive group coincides with the inactive state, containing cREs with low DNase Z-scores regardless of their H3K4me3, H3K27ac, and CTCF Z-scores.

To classify cREs in a particular cell type that lacks one or more data types, we must make approximations. The scheme is summarized as follows. If both H3K4me3 and H3K27ac data are available, then we incorporate the TSS proximity information by following the (above) classification trees; otherwise, we classify PLS if only H3K4me3 data are available and ELS if only H3K27ac data are available, without considering TSS proximity of the cREs. 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, cREs with low H3K4me3, H3K27ac, or CTCF signals are labelled “unclassified” because we are unable to definitively classify them as “inactive” without DNase data.

II.7.6 Saturation of cREs group with increasing numbers of cell types

To determine the relative saturation of cREs with promoter-like, enhancer-like or CTCF-only signatures, we used 21 cell types with all four core epigenomic marks (DNase, H3K4me3, H3K27ac, and CTCF). For *X* in the range of 1–21, we randomly selected *X* cell types 100 times. For each selection, we calculated the number of cREs in each of the three groups—promoter-like, enhancer-like, and CTCF-only signatures. Then, using the R script adapted from Steven Wilder and Ian Dunham (Consortium 2012), we calculated the cREs in each group to be at 95% saturation for each curve using a Weibull distribution. This pipeline is available on GitHub (Run-Calculate-cRE-Saturation.sh).

II.7.7 Overlap of cREs with ChromHMM states

We compared cREs with promoter-like and enhancer-like signatures to the chromatin states called by ChromHMM. We combined similar chromHMM states to generate seven broad states, as seen in Table II-2 on page 81. Each cRE was assigned to only one chromHMM state—the state that overlapped the largest number of basepairs.

For human, we analyzed the chromHMM regions for GM12878 cells from the ENCODE 2012 paper (ENCFF001TDH). We selected all cREs with promoter-like or enhancer-like signatures and ranked them by H3K4me3 and H3K27ac Z-scores, respectively. Then, we calculated the percentage of cREs in each 1 k bin that overlapped regions with each chromHMM state. This pipeline is available at GitHub (Ranked-ChromHMM-Overlap.sh).

For mouse, we analyzed 11 tissue–time-point combinations (from e11.5 and e14.5) for which we had DNase, H3K4me3, and H3K27ac data. We overlapped cREs with promoter-like or enhancer-like signatures with chromHMM states derived from eight histone marks in the same tissue–time-point. This part of the pipeline is available at GitHub (Overall-ChromHMM-Overlap.sh).

II.7.8 Clustering cell types on the basis of their cRE activities

To examine whether the Registry of cREs captured the regulatory landscapes of cell and tissue types, we performed hierarchical clustering on all primary cells and tissues with DNase-seq data by classifying the DNase Z-score at each cRE as either high (Z-score > 1.64) or low within each cell type. We also performed the same analysis using the Z-scores of H3K27ac, H3K4me3, or CTCF. We clustered tissues and primary cells separately because each tissue comprises multiple types of primary cells with different embryonic origins. For each cell or tissue type, we selected all cREs with a Z-score > 1.64 for each epigenomic mark and then calculated the Jaccard index for pairwise tissue or cell type comparisons. We clustered the tissues according to the pairwise Jaccard index using the hclust function in R. This pipeline is available at GitHub (Cluster-Cell-Types.sh).

II.7.9 Enrichment of GWAS variants in cREs

We curated studies from the NHGRI-EBI Catalogue (see Table II-3 on page 82) that were performed on European populations and used minor allele frequencies (MAF) and linkage disequilibrium (LD) of these populations to generate control SNPs. Because MAF and LD differ across populations, we limited the scope of our initial analysis to the populations with the most data. We used CEU-specific data of linkage disequilibrium (LD; correlation coefficient r^2>0.7) to perform statistical tests.

For each study, we generated a matching set of control SNPs as follows: for each SNP in the study (p < 1E-6) we selected a SNP on Illumina and Affymetrix SNP ChIPs that fell within the same minor allele frequency (MAF) quartile and the same distance to TSS quartile (Table II-4 on page 84). 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 cREs in a cell type were enriched in the GWAS SNPs, we intersected GWAS and control LD groups with cREs with an H3K27ac Z-score > 1.64 in the cell type. To avoid over-counting, we pruned the overlaps, counting each LD group once per cell type. We modified the Uncovering Enrichment through Simulation (UES) method (Hayes et al. 2015) with Fisher's exact tests for performing statistical testing. We calculated enrichment for overlapping cREs, comparing the GWAS LD groups with the 100 matched controls. Finally, we applied an FDR of 5% to each study.

II.7.10 Best single features for predicting tissue-specific enhancers

We used mouse embryonic enhancers in the VISTA database (Visel et al. 2007) to compare the effectiveness of the following ten types of epigenetic signals in predicting enhancers: DNase hypersensitivity, eight histone marks (H3K4me1, H3K4me2, H3K4me3, H3K27ac, H3K9me3, H3K36me3, and H3K27me3), and DNA methylation. VISTA enhancers have been tested with mouse transgenic assays, a widely acknowledged in vivo test for enhancer function (Pennacchio et al. 2006). 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 (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 because 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 assays, and there were hundreds of VISTA regions active in each of the tissues. We asked which of the ten epigenetic signals were predictive of VISTA enhancers in each tissue. Our positive test set comprised all VISTA enhancers in that tissue; our negative set contained the remaining VISTA regions (i.e., those that were tested but showed no activity in that tissue). The VISTA regions used in our analysis are listed in online Supplementary Table 2. Among four epigenetic signals (H3K4me3, H3K4me1, H3K27ac and DNase), the average DNase signal in a window anchored at DHSs was the most predictive feature for enhancer activity in the hindbrain, limb, and neural tube, with the area under the precision-recall curve AUPR=0.38, 0.39, and 0.29, respectively (Figure II-40 on page 126; Table II-5 on page 85). H3K27ac, anchored at H3K27ac peaks was the second-most predictive feature in these three tissues: AUPR = 0.33, 0.33, and 0.26, 9-17% lower than DNase. For midbrain enhancers, H3K27ac (AUPR=0.41) was the most predictive features followed by DNase (AUPR = 0.37).

DNase performs better than H3K27ac due to its higher 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, which lacks a nucleosome, but is high at the two flanking nucleosomes. Anchoring predictions on DHSs, we then tested different methods of ranking predictions testing both histone mark and DNase signals. On average, ranking with H3K27ac signal outperformed DNase signal in predicting of enhancer activity when averaged over a window centered DHSs (Figure II-41 on page 127**;** Table II-5 on page 85**)**. Signals for other histone marks and DNA methylation individually were far less predictive (Table II-5 on page 85). The average rank of the DNase and H3K27ac signals was slightly better than that of DNase. Incorporating additional histone marks or DNA methylation using a linear model did not further improve performance. We did not test more complex models because of the small number of VISTA enhancers—only 200-300 genomic regions tested positive in each tissue.

II.7.11 Combining signals accurately predicts active promoters

We further evaluated whether an adaptation of the above-described enhancer prediction model could be used to map cell-type-specific promoter regions. Judged by transcript expression levels measured by RNA-seq in the e11.5 midbrain, limb, neural tube, and hindbrain, the single most predictive feature among the ten we evaluated (DNase, eight histone marks, and DNA methylation) was the H3K4me3 signal. When averaged over a ±1.5 kb window centered on TSS-proximal DHSs, H3K4me3 correlated with expression levels at $r = 0.75$ averaged over the four tissues (Table II-6 on page 86; Figure II-42 on page 128 for the 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 with 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.

II.7.12 Evaluating the group classification of cREs in GM12878 cells

Figure II-15 on page 101 summarizes the five-group classification of cREs in GM12878. We used ChIP-seq data of RNA Pol II, EP300, and RAD21 in GM12878 to evaluate the group classification of cREs in this cell type. 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; Figure II-43 on page 129), yet moderately high EP300 signals (median = 10.9; Figure II-44 on page 130). Some of these cREs may function as both promoters and enhancers, but collectively they are more promoter-like than enhancer-like, as 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 (Figure II-45 on page 131**)**. 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 (Figure II-46 on page 132).

Figure II-47, Figure II-48, and Figure II-49 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*) (Figure II-47) , active in hepatocytes; hematopoietic transcription factor PU.1 (*SPI1*) (Figure II-48), active in B cells; and neuronal PAS domain protein 4 (*NPAS4*) (Figure II-49), 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.

II.7.13 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. These pipelines are implemented in the DNAnexus cloud computing environment and are freely available on GitHub (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 et al. 2016). 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 have submitted to the ENCODE Portal. 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.

Each ENCODE dataset is required to have two biological replicates; exceptions, typically resulting from a lack of cell or tissue samples that can serve as replicates, are noted. We have developed quality control (QC) metrics for each data type (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 et al. 2011). 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 because of limited experimental material), a dataset is still released if deemed valuable to the community, along with an audit flag stating the QC metrics that were not met.

Four ENCODE uniform processing pipelines are summarized below. More information can be found at the GitHub (github.com/ENCODE-DCC).

II.7.13.1 DNase-seq

The ENCODE DNase-seq processing pipeline consumes raw sequencing reads from technical replicates of experiments in the form of FASTQ files. Indexing and alignment of the FASTQ reads is performed with the Burrows-Wheeler Aligner (BWA) (Li and Durbin 2009), which outputs alignments in BAM format. Alignments from sets of technical replicates are merged and filtered prior to peak calling with HOTSPOT2, which generates peaks in BED format. Input FASTQs must meet minimum criteria to be processed, and various quality control metrics are also generated at each step. Further detail and basic workflows are available at github.com/ENCODE-DCC/dnase_pipeline.

II.7.13.2 ChIP-seq

The ENCODE consortium has developed two distinct ChIP-seq pipelines, one for transcription factor (TF) ChIP-seq data and one for histone ChIP-seq data, which take into account the different binding distributions of the respective immunoprecipitation targets across the genome. The ChIP-seq pipelines consume raw reads in FASTQ format; alignment of the reads is performed with BWA to generate alignment BAMs. Signal tracks are produced from the alignments using MACS2; these are output in two separate BigWigs, which represent fold-change over control and signal p-value.

Peaks are also called from the alignments, using MACS2 in the case of histone data and SPP in the case of TF data. Additionally, the pipelines call peaks from the pooled alignments of each experiment’s isogenic replicates. For TF experiments, the pooled peaks are compared with the peaks called for each replicate individually using IDR and thresholded to generate a conservative set of peaks and an optimal set of peaks;

for histone data, sets of replicated peaks are generated by comparing the pooled and individual peaks using overlap_peaks. Further detail and basic workflows are available at github.com/ENCODE-DCC/chip-seq-pipeline.

II.7.13.3RNA-seq

There are two distinct ENCODE uniform RNA-seq pipelines, one for RNAs longer than 200 bp and the other for RNAs shorter than 200 bp. The long RNA pipeline is appropriate for processing libraries generated from mRNA, rRNA-depleted total RNA, or poly-A(–) RNA. The pipeline consumes RNA-seq reads in FASTQ format; alignment is performed with STAR and gene and transcript quantification is performed by RSEM against a gene annotation file, which contains by default GENCODE annotations. STAR also outputs normalized RNA-seq signal for both the (+) and (–) strands. Further details are available at github.com/ENCODE-DCC/long-rna-seq-pipeline.

II.7.13.4RAMPAGE

Like the long RNA-seq pipeline, the ENCODE RAMPAGE pipeline is appropriate for libraries generated with RNAs longer than 200bp, and it consumes reads in FASTQ format and produces alignments and normalized signal for both the (+) and (–) strands with STAR. PeaOverlap of cREs with H3K4me3, H3K27ac, and CTCF peaks in cell types without DNase-seq data. To determine the comprehensiveness of the Registry, we overlapped cREs with ChIP-seq peaks (H3K4me3, H3K27ac, and CTCF) from cell types lacking DNase data. Using bedtools merge, we merged all ChIP-seq peaks within 200 bp of one another and assigned each merged peak the maximal -log(FDR) score of the original peaks. We then filtered out all peaks with -log(FDR) < 2. Using bedtools intersect with the "-u" flag, we intersected the merged peaks with cREs and counted the number of unique peaks that overlapped at least one cRE. This pipeline is available on GitHub at Calculate-Peak-Overlap.sh. ks, representing transcription start sites, are called from the alignments using GRIT, and output in BED, bigBED, and GFF formats. QC is performed for the peaks, and IDR is used to identify reproducible peaks between replicates.

II.7.14 Testing single features for predicting tissue-specific enhancers

We downloaded all regions from the VISTA Enhancer database in November, 2015. Merging overlapping regions yielded 1,994 unique regions. Because we had histone mark ChIP-seq, DNase-seq, and RNA-seq data for the midbrain, hindbrain, limb, or neural tube at embryonic day 11.5, we selected all regions active in these four tissues at e11.5, thus resulting in 301, 271, and 193 active regions, respectively (see online Supplementary Table 2).

We determined the best method for anchoring enhancer predictions (i.e., which peaks should be used to center the genomic regions as predicted enhancers) and then tested metrics for ranking these predictions. We tested the metrics using DHSs and H3K4me3, H3K4me1, and H3K27ac peaks for anchors. To make comparisons across the different data types and to account for differences in their genome coverage, we developed a uniform comparison pipeline with the following requirements:

1. *Uniform number of peaks across cell types.* We restricted the DHSs and histone mark peaks to the same number in each cell type, using the minimal number of peaks and DHSs across all datasets. For example, in the midbrain, there are 168 k

DHSs, 28 k H3K27ac peaks, 81 k H3K4me1 peaks, and 21 k H3K4me3 peaks, and we selected the top 21 k peaks of all four datasets for analysis.

2. *Uniform width for predicted enhancers.* We resized each DHS or histone mark peak to the same length of 300 bps, centered on the midpoint of DHSs and the summit of histone peaks (the position with the highest ChIP signal), and used these as enhancer predictions.

We intersected DHSs and histone mark peaks with all VISTA regions. If a VISTA region overlapped a DHS or peak, we assigned the region the score of the DHS or peak, i.e., its –log(p-value) or signal. If a VISTA region overlapped multiple DHSs or peaks, we assigned it the maximal score of the overlapping DHSs or peaks. If a VISTA region did not overlap any DHSs or peaks, we assigned it a score of 0. To evaluate the performance of each method, we calculated the area under the Precision-Recall Curve (AUPRC) using the ROCR package and custom R scripts. This pipeline is available on GitHub at: Evaluate-VISTA-Enhancers.sh.

Averaged over the four tissues, DHSs performed the best as anchors for enhancer predictions, followed by H3K27ac peaks (Figure II-40 on page 126). Anchoring all enhancer predictions on DHSs, we tested different metrics for ranking the regions. Overall, the best performing metric was the average rank of H3K27ac and DNase signals (Figure II-41 on page 127).

II.7.15 Prediction of expression levels

To test methods of promoter prediction, we used transcript expression values from the RNA-seq uniform processing pipeline. We computed Pearson correlations between the ranks of TSS-proximal (± 2 kb) DHSs or H3K4me3 peaks (by DNase or H3K4me3 signal) and the ranks of the expression levels of nearby transcripts. We tested all four combinations of ranking schemes (DHSs ranked by DNase signal, H3K4me3 peaks ranked by DNase signal, DHSs ranked by H3K4me3 signal, and H3K4me3 peaks ranked by H3K4me3 signal). The method with the highest correlation was centering predictions on DHSs and ranking by H3K4me3 signal. This pipeline is available on GitHub**.**

II.8 Discussion

The genetics revolution has fundamentally changed our understanding of medicine over the past several decades. More recently, however, we are learning that epigenetics is as important, if not more so, to our understanding of disease. There are already catalogs of hundreds of epigenetic changes affecting disease (Mirabella, Foster, and Bartke 2016). With ~90% of disease-associated Single Nucleotide Polymorphisms (SNPs) occurring to be in intronic or intergenic regions (Hindorff et al. 2009), the systematic location and study of genetic variants outside of protein-coding regions is critical to better understanding and potentially treating disease pathology. There are millions of these potentially functional regions in the genome, working as enhancer, promoters, repressors, or insulators; having a catalog of these elements, and a way to easily interrogate these regions, will be essential to keep researchers afloat in an ocean of data. Researchers require tools to aid in both biological questions and practical bioinformatics problems; users require an encyclopedia synthesizing the low-level data into a more manageable product that can be analyzed and effectively investigated.

We have identified the first real catalog of putative regulatory regions, locating nearly 2 millions cREs across human and mouse genomes in regions with open chromatin and enhancer-like or promoter-like signatures (based on histone modification marks and other genomic distance information). These elements are numbered and versioned, permitting direct reference in future papers. cREs are anchored on DNase-seq representative DHSs (rDHSs), condensed first from >30 million DHS sites across more than 400 individual samples into a set of non-overlapping regions number ~1.3 million in human and ~400 thousand in mice. We have found a wealth of putative regulatory regions that correlate with experimental datasets within and outside of ENCODE. These regions greatly simplify the search for putative regulatory regions genes or SNPs of interest.

To interrogate these regions, we needed a tool with several requirements, including dynamic and interactive interfaces that allow users to search and filter cREs in real time, as well as display interactive plots of cREs. The tool needed to integrate and view all the low-level data in some sort of genome browser. The interactive plots also needed to be reproducible, and near-publication quality. Finally, we also needed a mechanism to facilitate users' generation of these figures with custom data. The tool needed to be extensible, providing for the inclusion of analyses involving external annotations and user-submitted annotations.

To fulfill these requirements, we developed SCREEN, the visualizer for cREs. This tool is maturing into an integrated approach to examining cREs in the context of the genome and epigenome. We already have a multi-part visualization platform, with mechanisms to search and investigate cREs, show gene expression information, and explore GWAS studies for SNP overlap with cREs. SCREEN is the start of central repository for accessing information on functional regions of human and mouse genomes, integrating cRE searching, sorting, and visualization. It will (hopefully) become an easily adaptable tool widely utilized.

II.9 Tables

Table II-1 | ENCODE Project data production as of June 20, 2017

Category	Assay	# Tissues	# primary cell	# Cell Line	# IPSC	# in vitro	# Stem cell	Number of Experiments (Phase III)	Number of Experiments (All ENCODE)	Number of Experiments (All ENCODE + ROADMAP)
Human										
Transcriptome	Bru-seq	0	0	1	0	0	0	0	1	1
	CAGE	1	32	37		1	7	0	78	78
	CRISPR genome editing followed by RNA-seq	0	0	27	0	0	0	27	27	27
	CRISPRi followed by RNA-seq	0	0	147	0	0	0	0	147	147
	RAMPAGE	104	16	27	1	6	1	155	155	155
	RNA-PET	1	7	22	0	0	1	0	31	31
	polyA depleted RNA-seq	0	13	15	0	1	3	1	32	32
	polyA mRNA RNA-seq	197	73	89	3	24	19	27	132	405
	small RNA-seq	68	33	57	1	8	6	86	173	173
	total RNA-seq	114	60	47	2	13	5	210	239	241
	microRNA counts	24	2	5	1	5	1	37	38	38
	microRNA-seq	52	38	5	1	5	6	36	38	107
	shRNA knockdown followed by RNA-seq	0	0	526	0	0	0	524	526	526
	siRNA knockdown followed by RNA-seq	0	0	55	0	0	0	50	55	55
	single cell isolation followed by RNA-seq	0	26	13	1	1	0	41	41	41

	RNA microarray	78	108	82	10	7	6	0	179	291
Transcriptional regulation and replication	DNase-seq	322	168	160	11	28	14	163	372	703
	ATAC-seq	34	0	0	0	0	0	33	0	34
	DNAme array	124	54	71	1	6	3	0	259	259
	FAIRE-seq	7	4	25	0	0	1	0	37	37
	MNase-seq	0	0	2	0	0	0	0	2	2
	MRE-seq	3	35	2	0	0	5	0	2	45
	MeDIP-seq	8	37	2	0	0	4	0	2	51
	RRBS	31	44	48	6	7	7	0	103	143
	WGBS	74	35	3	2	13	6	9	9	133
	ChIP-seq, histone	727	534	360	49	274	127	495	852	2071
	ChIP-seq, TF	225	76	1222	5	15	79	615	1622	1622
	ChIP-seq, RNA binding protein	0	1	76	1	1	4	48	83	83
	ChIP-seq, recombinant	0	0	218	0	0	0	178	218	218
	ChIP-seq, control	359	138	404	15	46	33	393	742	995
	ChIP-seq, other post-translational modification	0	0	3	0	2	0	4	4	5
	Repli-ChIP	0	5	5	2	27	6	36	45	45
	Repli-seq	0	24	60	0	0	6	0	90	90
	5C	0	1	11	0	0	1	0	13	13
	ChIA-PET	0	0	39	0	0	0	31	39	39
	HiC	0	2	12	0	0	0	12	14	14
Post-transcriptional regulation via RBPs	RIP-chip	0	0	29	0	0	3	0	32	32
	RIP-seq	0	0	43	0	0	0	35	43	43
	RNA Bind-N-Seq	0	0	0	0	158	0	158	158	158
	eCLIP	3	0	318	0	0	0	320	321	321
	iCLIP	0	0	5	0	0	0	5	5	5

genotyping	DNA-PET	0	0	6	0	0	0	0	6	6
	genotyping array	16	54	62	2	4	4	59	123	142
	genotyping HTS	8	1	0	0	0	0	9	9	9
Human Total								3797	7096	9665
Mouse										
Transcriptome	polyA mRNA RNA-seq	78	8	21	0	2	2	0	111	
	total RNA-seq	83	14	8	0	0	3	104	108	
	microRNA counts	77	0	0	0	0	0	77	77	
	microRNA-seq	65	0	0	0	0	0	65	65	
	single cell RNA-seq	12	59	0	0	0	0	71	71	
Transcriptional regulation and replication	DNase-seq	49	12	15	0	4	8	33	88	
	ATAC-seq	27	7	2	0	0	1	10	37	
	MRE-seq	0	0	2	0	0	0	0	2	
	MeDIP-seq	0	0	2	0	0	0	0	2	
	WGBS	72	0	0	0	0	0	72	72	
	ChIP-seq, histone	638	18	46	0	6	20	564	728	
	ChIP-seq, TF	45	9	106	0	15	10	16	185	
	ChIP-seq, control	113	5	25	0	4	4	94	151	
	Repli-ChIP	0	1	4	0	8	5	0	18	
Mouse Total								1106	1715	
Grand Total								4903	8811	11380

Table II-2 | Consolidated ChromHMM States

Combined State				
TSS	1 Active Promoter	2 Weak Promoter		
TSS Bivalent	3 Poised Promoter			
High Signal Enhancer	4 Strong Enhancer	5 Strong Enhancer		
Low Signal Enhancer	6 Weak Enhancer	7 Weak Enhancer		
Insulator	8 Insulator			
Transcription	9 Txn Transition	10 Txn Elongation	11 Weak Txn	
Repressed	12 Repressed	13 Heterochrom/lo	14 Repetitive/CNV	15 Repetitive/CNV

Table II-3 | GWAS Studies

First Author	*PMID*	*Phenotype*
Anderson	21297633	Ulcerative colitis
Anttila	23793025	Migraine
Arking	24952745	QT Interval
Barrett	19430480	Type 1 Diabetes
Baurecht	25574825	Inflammatory skin disease
Baurecht	25574825	Psoriasis
Bentham	26502338	Systemic lupus erythematosus
Berndt	23563607	Height
Berndt	23563607	Obesity
Cai	25130324	Heschl's gyrus morphology
Chasman	19936222	Lipid metabolism phenotypes
deVries	26561523	Fibrinogen levels
Dubois	20190752	Celiac disease
Dupuis	20081858	Fasting glucose-related traits
Fox	22589738	Subcutaneous adipose tissue
Fox	22589738	Visceral adipose tissue adjusted for BMI
Fox	22589738	Visceral adipose tissue/subcutaneous adipose tissue ratio
Fox	22589738	Visceral fat
Franke	21102463	Crohn's disease
Gieger	22139419	Platelet count
Gieger	22139419	Mean platelet volume
Gudbjartsson	18391951	Height
Hromatka	25628336	Motion sickness
Imboden	22424883	Pulmonary function decline
Jostins	23128233	Inflammatory bowel disease
Kaplan	21216879	Insulin-like growth factors
Kapoor	24962325	Alcohol dependence (age at onset)
Kottgen	23263486	Urate levels
Lango	20881960	Height
Lemaitre	21829377	Phospholipid levels (plasma)
Lesch	18839057	Attention deficit hyperactivity disorder
Li	26252872	Cognitive decline rate in late mild cognitive impairment
Li	26301688	Pediatric autoimmune diseases
Liu	26192919	Crohn's disease
Liu	26192919	Inflammatory bowel disease
Liu	26192919	Ulcerative colitis

Michailidou	23535729	Breast cancer
Mozaffarian	25646338	Trans fatty acid levels
Patsopoulos	22190364	Multiple sclerosis
Perry	25231870	Menarche (age at onset)
Porcu	23408906	Thyroid hormone levels
Rietveld	25201988	Educational attainment
Ripke	25056061	Schizophrenia
Sawcer	21833088	Multiple sclerosis
Shin	24816252	Blood metabolite levels
Shin	24816252	Blood metabolite ratios
Speedy	24292274	Chronic lymphocytic leukemia
Suhre	21886157	Metabolic traits
Surakka	25961943	Cholesterol, total
Surakka	25961943	HDL cholesterol
Surakka	25961943	LDL cholesterol
Surakka	25961943	Triglycerides
Teslovich	20686565	Cholesterol, total
Teslovich	20686565	HDL cholesterol
Teslovich	20686565	LDL cholesterol
Teslovich	20686565	Triglycerides
vanderHarst	23222517	Red blood cell traits
Wain	21909110	Blood pressure
Wang	20889312	Bipolar disorder and schizophrenia
Willer	24097068	Cholesterol, total
Willer	24097068	HDL cholesterol
Willer	24097068	LDL cholesterol
Willer	24097068	Triglycerides
Wood	25282103	Height
Yucesoy	25918132	Diisocyanate-induced asthma

Table II-4 | Minor Allele Frequency

	25%	50%	75%
Minor Allele Frequency	0.06	0.18	0.33
Distance to TSS	9,553	39,530	154,279

Table II-5 | PR Curve Results

Peak Space	Signal	Hindbrain	Limb	Midbrain	Neural Tube	Average
DNase	DNase	0.3761	0.393	0.3671	0.2884	**0.3562**
H3K27ac	H3K27ac	0.3266	0.3264	0.4072	0.2639	0.3310
H3K4me3	H3K4me3	0.2034	0.1239	0.2406	0.1316	0.1749
H3K4me1	H3K4me1	0.2036	0.2481	0.3044	0.1559	0.2280
		N=20,000	N=20,000	N=20,000	N=20,000	N=20,000
Peak Space	**Signal**	**Hindbrain**	**Limb**	**Midbrain**	**Neural Tube**	**Average**
DNase	DNase	0.3788	0.4159	0.3797	0.2951	0.3673
DNase	H3K27ac	0.3113	0.3265	0.3959	0.2526	0.3216
DNase	Average Rank DNase-H3K27ac	0.3764	0.3948	0.4148	0.3050	**0.3727**
DNase	H3K4me3	0.2276	0.1828	0.2602	0.1615	0.2080
	Average Rank DNase-H3K4me3	0.2584	0.2392	0.2933	0.1751	0.2415
DNase	H3K4me1	0.2442	0.2799	0.3122	0.1762	0.2531
	Average Rank DNase-H3K4me1	0.2527	0.2647	0.2901	0.1740	0.2454
DNase	H3K9ac	0.2367	0.1977	0.2756	0.1721	0.2205
DNase	Average Rank DNase-H3K9ac	0.2831	0.2574	0.3250	0.2147	0.2700
DNase	H3K36me3	0.1910	0.1776	0.1911	0.1265	0.1715
DNase	Average Rank DNase-H3K36me3	0.2280	0.2262	0.2212	0.1548	0.2075
DNase	WGBS methylation**	0.2470	0.2151	0.2663	0.1550	0.2208
DNase	Average Rank DNase-WGBS	0.3127	0.3031	0.3278	0.1981	0.2854
DNase	H3K27me3**	0.2187	0.1964	0.1853	0.1285	0.1822
DNase	Average Rank DNase-H3K27me3	0.2700	0.2750	0.2325	0.1664	0.2360
	** inverse of signal	N=130,754	N=151,790	N=268,062	N=162,801	

Table II-6 | Promoter Prediction

Peak Space	Signal	Hindbrain	Limb	Midbrain	Neural Tube	Average	GM12878	K562	HepG2	Average
DNase	DNase	0.3454	0.3643	0.3973	0.4714	0.3946	0.4904	0.3848	0.4024	0.4258
DNase	H3K4me3	0.7332	0.7472	0.7507	0.7488	**0.7450**	0.7152	0.7310	0.7084	**0.7182**
H3K4me3	DNase	0.2364	0.2603	0.2239	0.1055	0.2065	0.4122	0.3016	0.2469	0.3202
H3K4me3	H3K4me3	0.5551	0.6112	0.5691	0.5555	0.5727	0.5833	0.6012	0.5484	0.5777

Peak Space	Signal	GM12878	K562	HepG2	Average
DNase	DNase	0.4904	0.3848	0.4024	0.4258
DNase	H3K4me3	0.7152	0.7310	0.7084	**0.7182**
H3K4me3	DNase	0.4122	0.3016	0.2469	0.3202
H3K4me3	H3K4me3	0.5833	0.6012	0.5484	0.5777

II.10Figures

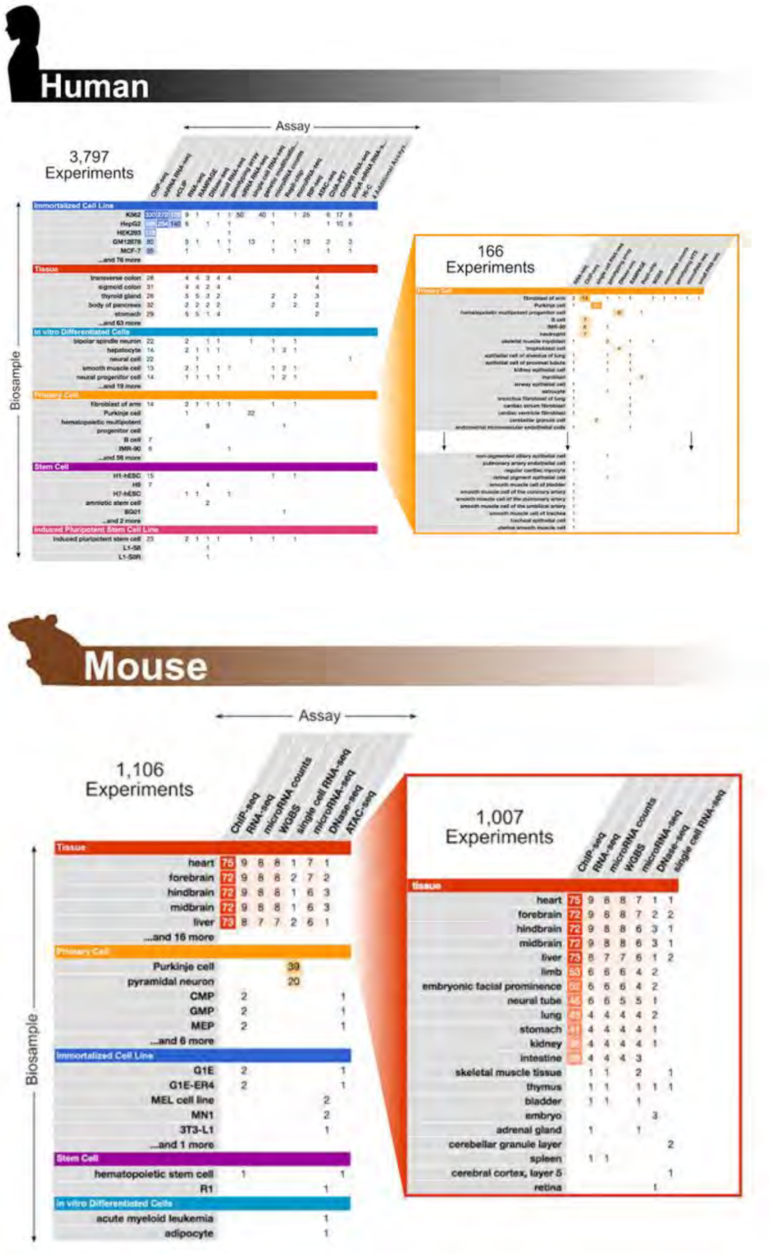


Figure II-1 | ENCODE Phase III data production as of February 1, 2017
Human and Mouse ENCODE Phase III experiments available on the ENCODE Portal. Experiments are categorized by the following assay and biosample types: blue for immortalized cell lines, red for tissues, teal for in vitro differentiated cells, orange for primary cells, purple for stem cells, and pink for iPSCs.

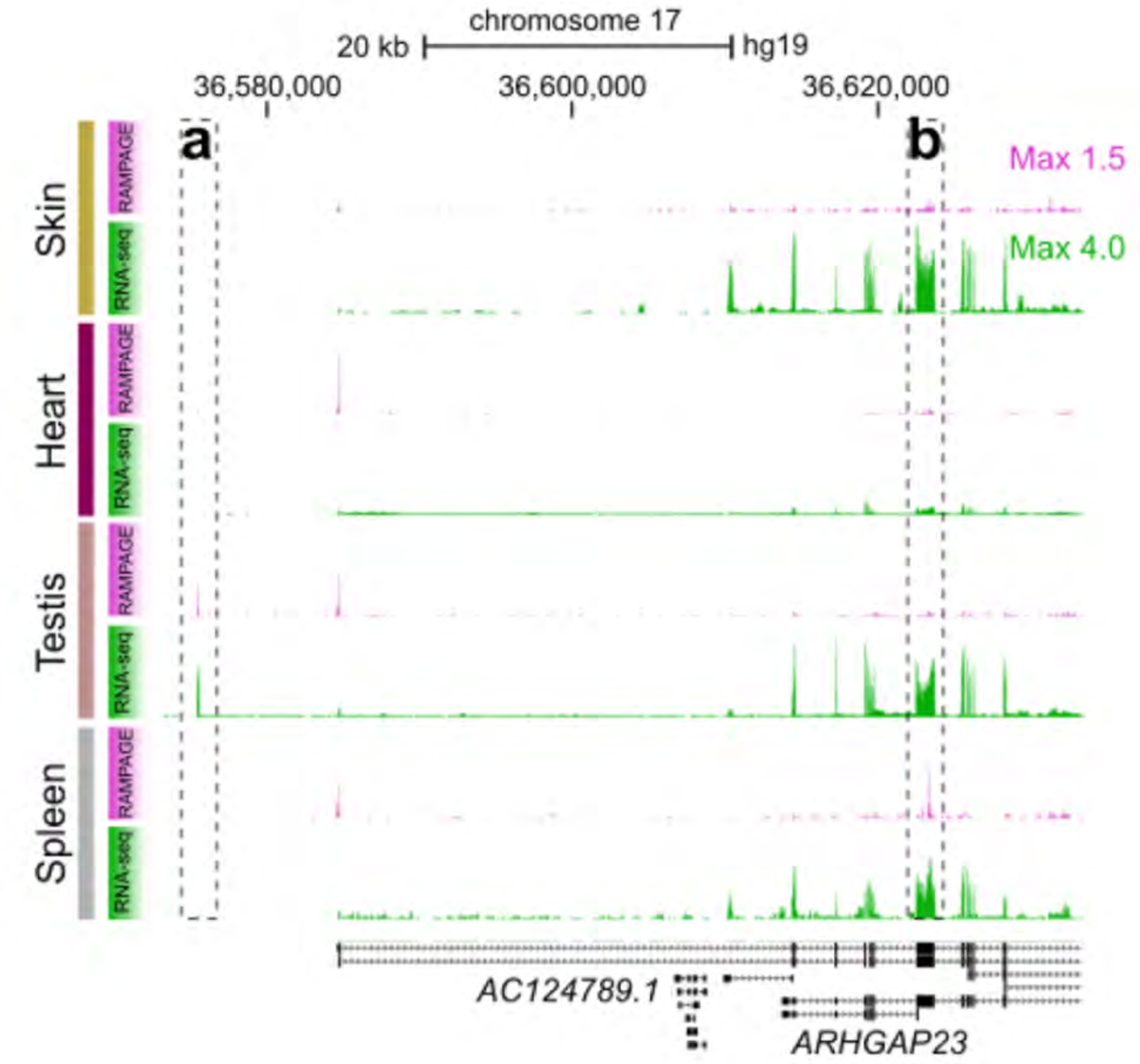


Figure II-2 | New assays used in ENCODE Phase III
Using the 5' ends of RAMPAGE reads, we can identify TSSs and quantify tissue- and transcript-specific transcription. **a,** In testis, we identified a novel, tissue-specific TSS for *ARHGAP23* upstream of previous annotated TSSs. **b,** In spleen, we identified a novel TSS within exon 7 of *ARHGAP23*.

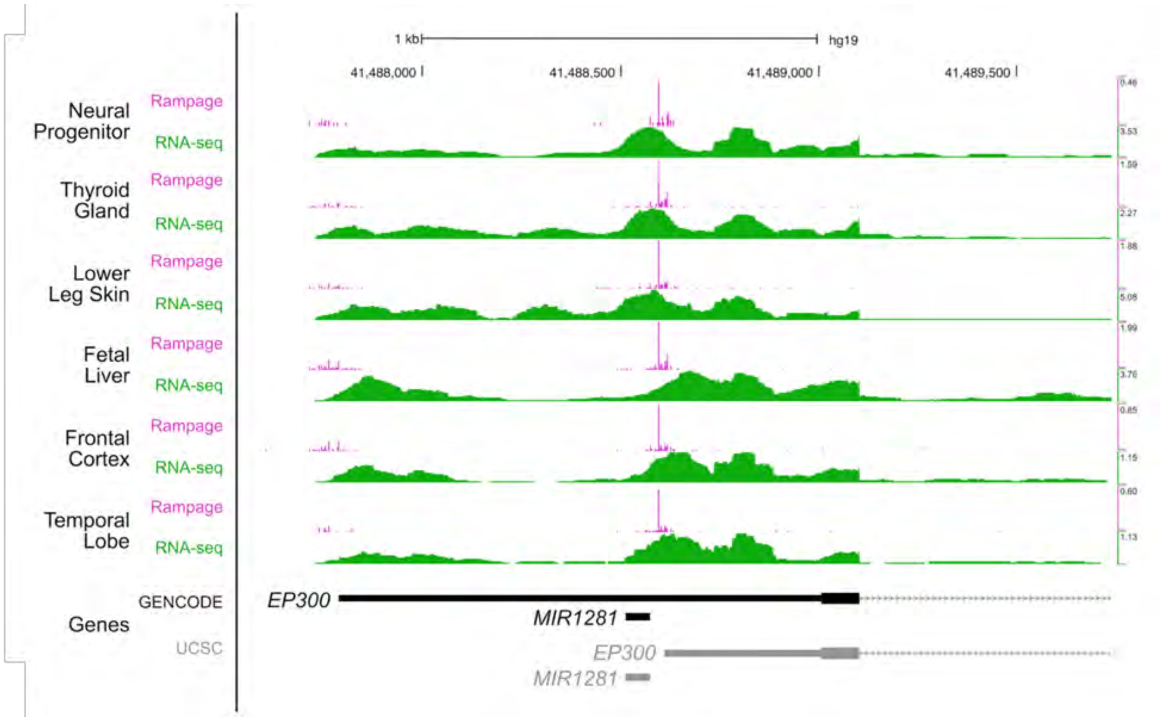


Figure II-3 | RAMPAGE data signal at EP300
RAMPAGE signals across six human tissues at *EP300* demonstrate that both the GENCODE- and UCSC-annotated TSSs for *EP300* are active.

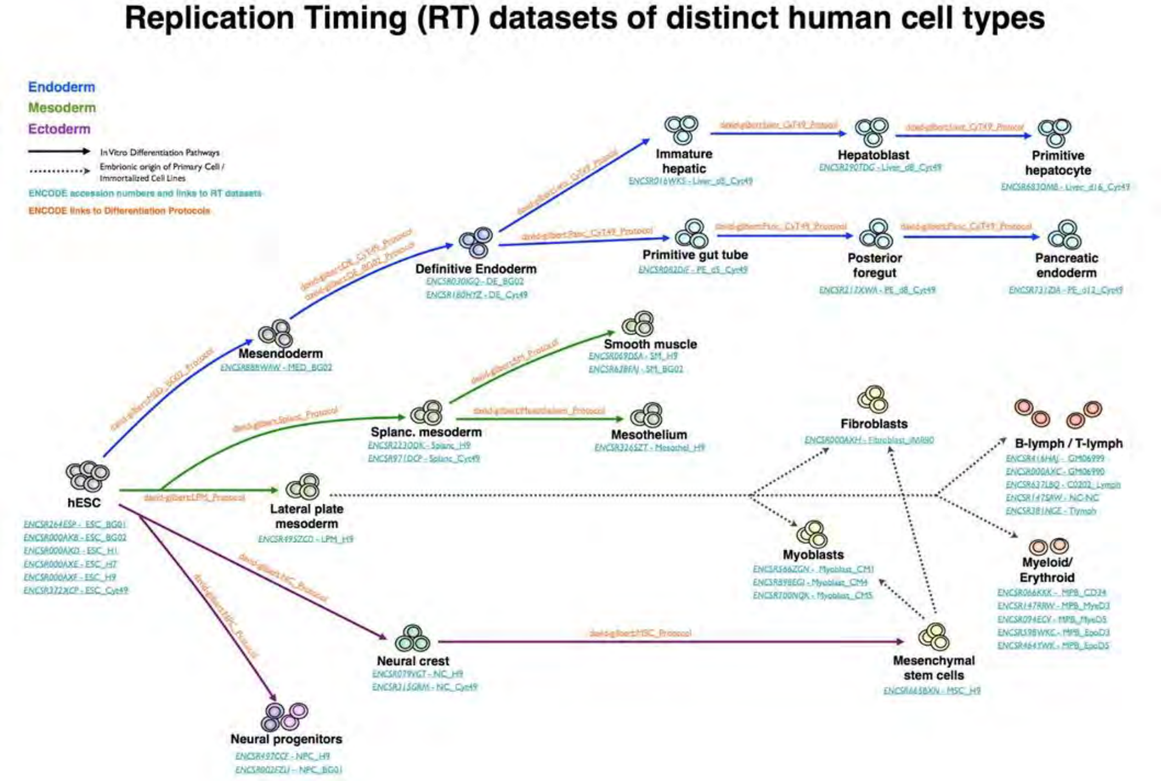

Replication Timing (RT) datasets of distinct human cell types
Endoderm
Mesoderm
Ectoderm
In Vitro Differentiation Pathways
Embrionic origin of Primary Cell / Immortalized Cell Lines
ENCODE accession numbers and links to RT datasets
ENCODE links to Differentiation Protocols
hESC
ENCSR264ESP - ESC_BG01
ENCSR000AKB - ESC_BG02
ENCSR000AKD - ESC_H1
ENCSR000AXE - ESC_H7
ENCSR000AXF - ESC_H9
ENCSR372XCP - ESC_Cyt49
david-gilbert:MED_BG02_Protocol
Mesendoderm
ENCSR888WAW - MED_BG02
david-gilbert:DE_CyT49_Protocol
david-gilbert:DE_BG02_Protocol
Definitive Endoderm
ENCSR030IGQ - DE_BG02
ENCSR180HYZ - DE_Cyt49
david-gilbert:Liver_CyT49_Protocol
Immature hepatic
ENCSR016WKS - Liver_d5_Cyt49
david-gilbert:Liver_CyT49_Protocol
Hepatoblast
ENCSR290TDG - Liver_d8_Cyt49
david-gilbert:Liver_CyT49_Protocol
Primitive hepatocyte
ENCSR683OMB - Liver_d16_Cyt49
david-gilbert:Panc_CyT49_Protocol
Primitive gut tube
ENCSR082DIF - PE_d5_Cyt49
david-gilbert:Panc_CyT49_Protocol
Posterior foregut
ENCSR217XWA - PE_d8_Cyt49
david-gilbert:Panc_CyT49_Protocol
Pancreatic endoderm
ENCSR731ZIA - PE_d12_Cyt49
david-gilbert:Splanc_Protocol
Splanc. mesoderm
ENCSR223ODK - Splanc_H9
ENCSR971DCP - Splanc_Cyt49
david-gilbert:SM_Protocol
Smooth muscle
ENCSR069DSA - SM_H9
ENCSR638FAJ - SM_BG02
david-gilbert:Mesothelium_Protocol
Mesothelium
ENCSR326SZT - Mesothel_H9
david-gilbert:LPM_Protocol
Lateral plate mesoderm
ENCSR495ZCD - LPM_H9
Fibroblasts
ENCSR000AXH - Fibroblast_IMR90
B-lymph / T-lymph
ENCSR416HAJ - GM06999
ENCSR000AXC - GM06990
ENCSR637LBQ - C0202_Lymph
ENCSR147SAW - NC-NC
ENCSR381NGE - Tlymph
Myoblasts
ENCSR566ZGN - Myoblast_CM1
ENCSR898EGI - Myoblast_CM4
ENCSR700NQK - Myoblast_CM5
Myeloid/ Erythroid
ENCSR066KKK - MPB_CD34
ENCSR147RRW - MPB_MyeD3
ENCSR094ECV - MPB_MyeD5
ENCSR598WKC - MPB_EpoD3
ENCSR464YWK - MPB_EpoD5
david-gilbert:NC_Protocol
Neural crest
ENCSR079VGT - NC_H9
ENCSR315GRM - NC_Cyt49
david-gilbert:MSC_Protocol
Mesenchymal stem cells
ENCSR658XN - MSC_H9
david-gilbert:NPC_Protocol
Neural progenitors
ENCSR497CCF - NPC_H9
ENCSR002FZU - NPC_BG01

Figure II-4 | DNA replication timing (RT) programs
Genome-wide RT programs were obtained for distinct human cell types, including embryonic stem cell (hESC)-derived, primary cells and established cell lines representing intermediate stages of endoderm, mesoderm, ectoderm, and neural crest 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). Dataset and protocol ENCODE IDs are shown in blue and brown for each cell type.

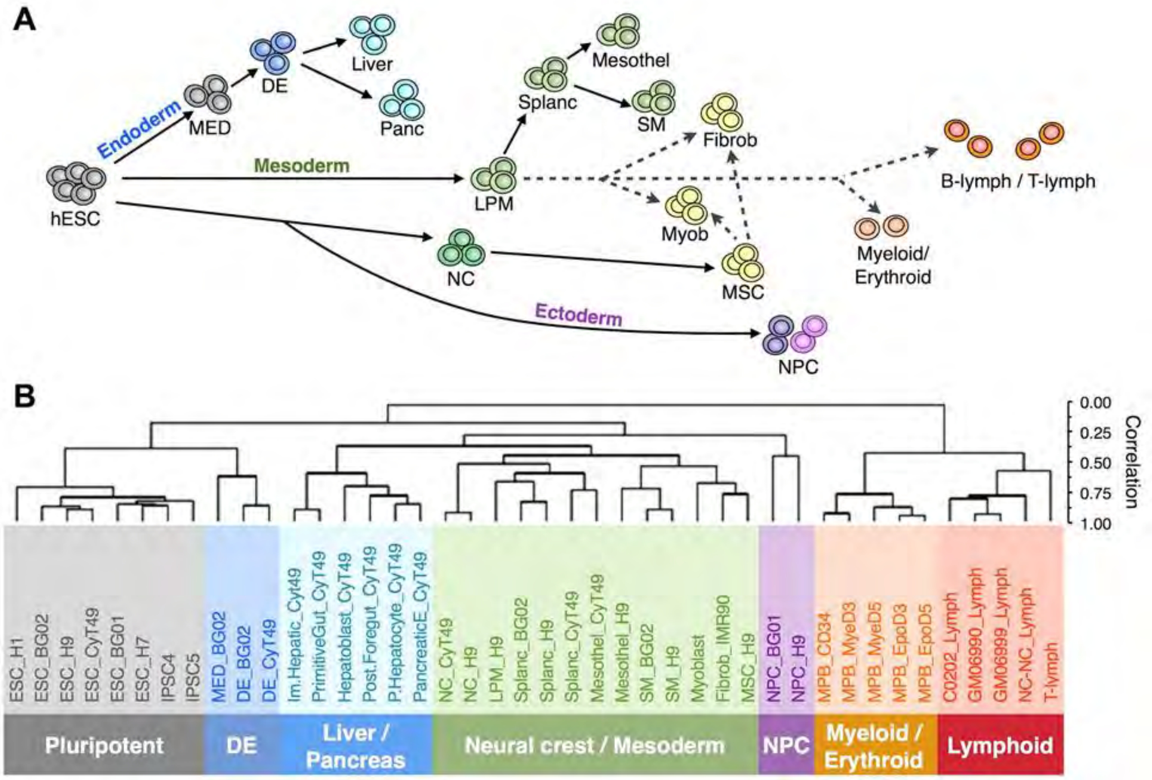


Figure II-5 | DNA replication timing (RT) programs are cell type-specific
a: Schematic diagram showing the three germ layers and the neural crest during the early stages of human development and differentiation pathways of the distinct cell types analyzed. **b:** Hierarchical clustering of RT programs from the distinct human cell types. Branches of the dendrogram were constructed based on the Pearson correlation coefficients between cell types (distance = 1 – correlation value). 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. (NC) neural crest; (MED) mesendoderm; (DE) definitive endoderm; (LPM) lateral plate mesoderm; (Splanc) splanchnic mesoderm; (Mesothel) mesothelium; (SM) smooth muscle; (Myob) myoblasts; (Fibrob) fibroblasts; (MSC) mesenchymal stem cells; (NPC) neural progenitor cells.

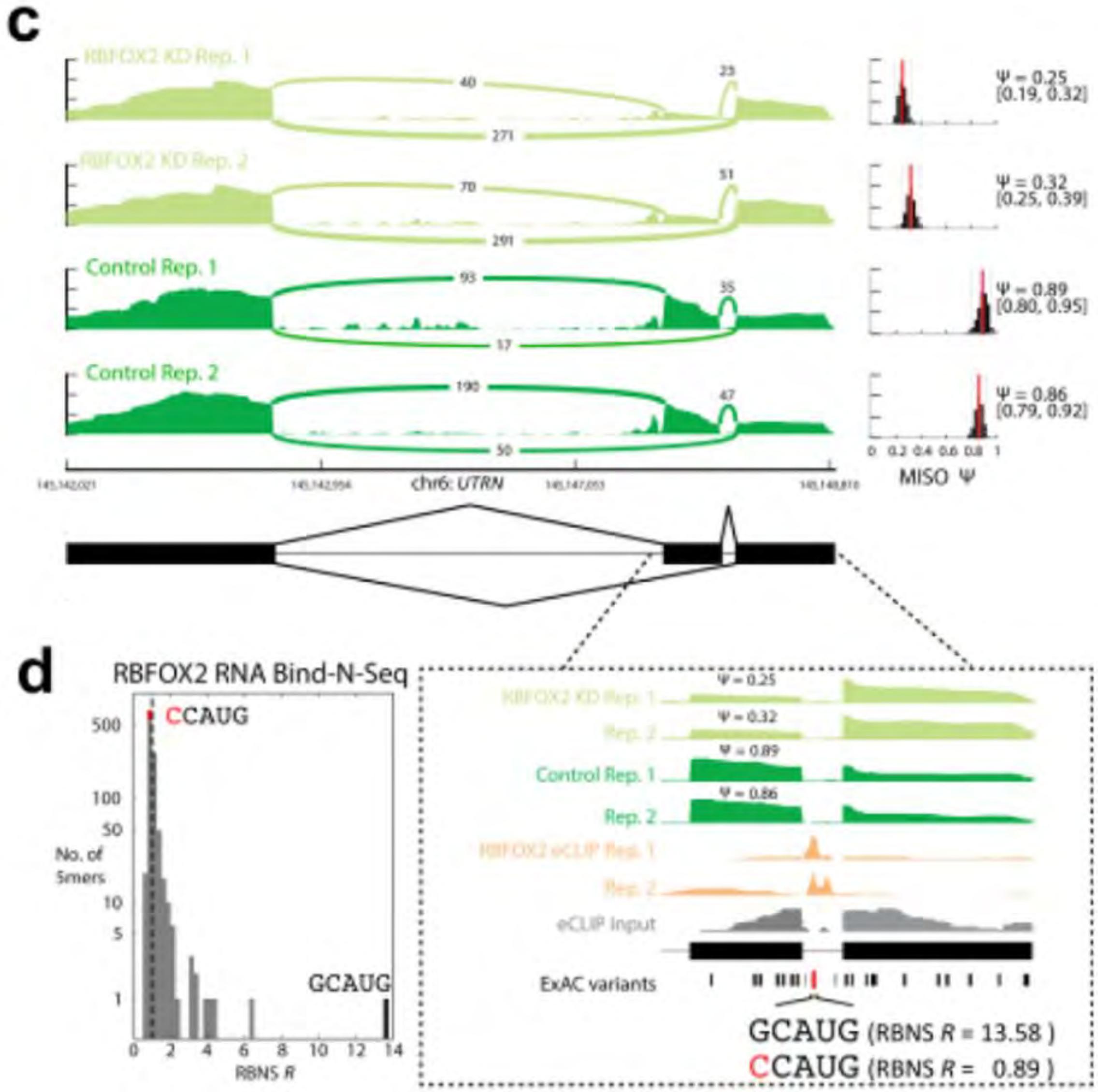


Figure II-6 | New assays used in ENCODE Phase III
c–d, Integrative analyses of RBP data can identify genetic variants that may impact RBP regulation. **c,** Control and RBFOX2 knockdown RNA-seq of exons 65–67 of the *UTRN* gene in HepG2 cells. Inclusion of the alternatively spliced exon 66 is reduced 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). **d,** (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).

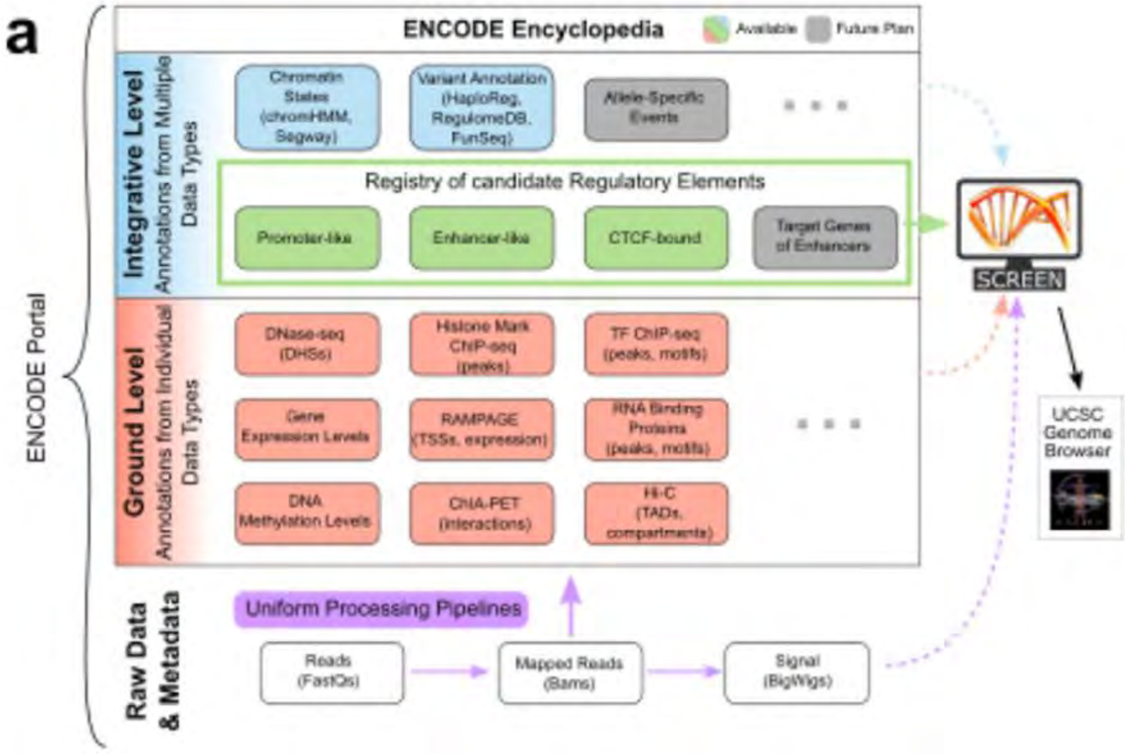

a
ENCODE Encyclopedia
Available
Future Plan
ENCODE Portal
Integrative Level
Annotations from Multiple Data Types
Chromatin States (chromHMM, Segway)
Variant Annotation (HaploReg, RegulomeDB, FunSeq)
Allele-Specific Events
Registry of candidate Regulatory Elements
Promoter-like
Enhancer-like
CTCF-bound
Target Genes of Enhancers
SCREEN
Ground Level
Annotations from Individual Data Types
DNase-seq (DHSs)
Histone Mark ChIP-seq (peaks)
TF ChIP-seq (peaks, motifs)
Gene Expression Levels
RAMPAGE (TSSs, expression)
RNA Binding Proteins (peaks, motifs)
DNA Methylation Levels
ChIA-PET (interactions)
Hi-C (TADs, compartments)
UCSC Genome Browser
Raw Data & Metadata
Uniform Processing Pipelines
Reads (FastQs)
Mapped Reads (Bams)
Signal (BigWigs)

Figure II-7 | Overview of the ENCODE Encyclopedia
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

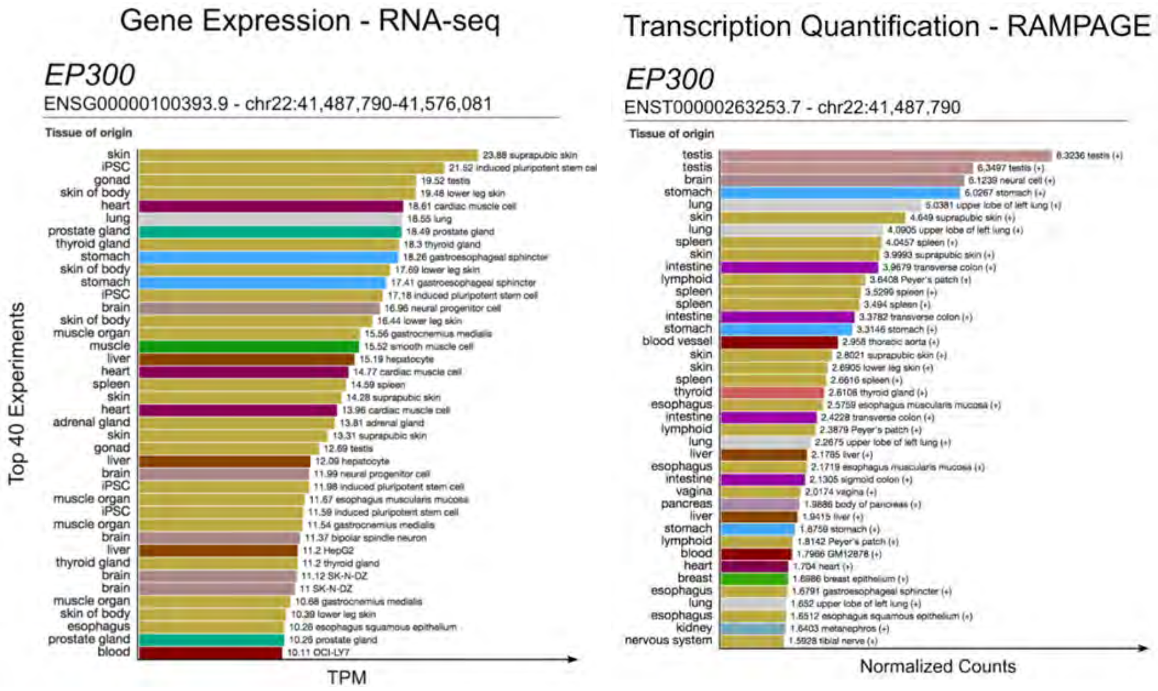


Figure II-8 | SCREEN display of gene and TSS expression levels.
Left: Gene expression of *EP300* from whole-cell RNA-seq assays shown in tags per million (TPM).

Right: RAMPAGE signal at the TSS of ENST00000263253.7 (averaged over ± 50 bp window). Bars are colored according to the tissue of origin indicated on the left.

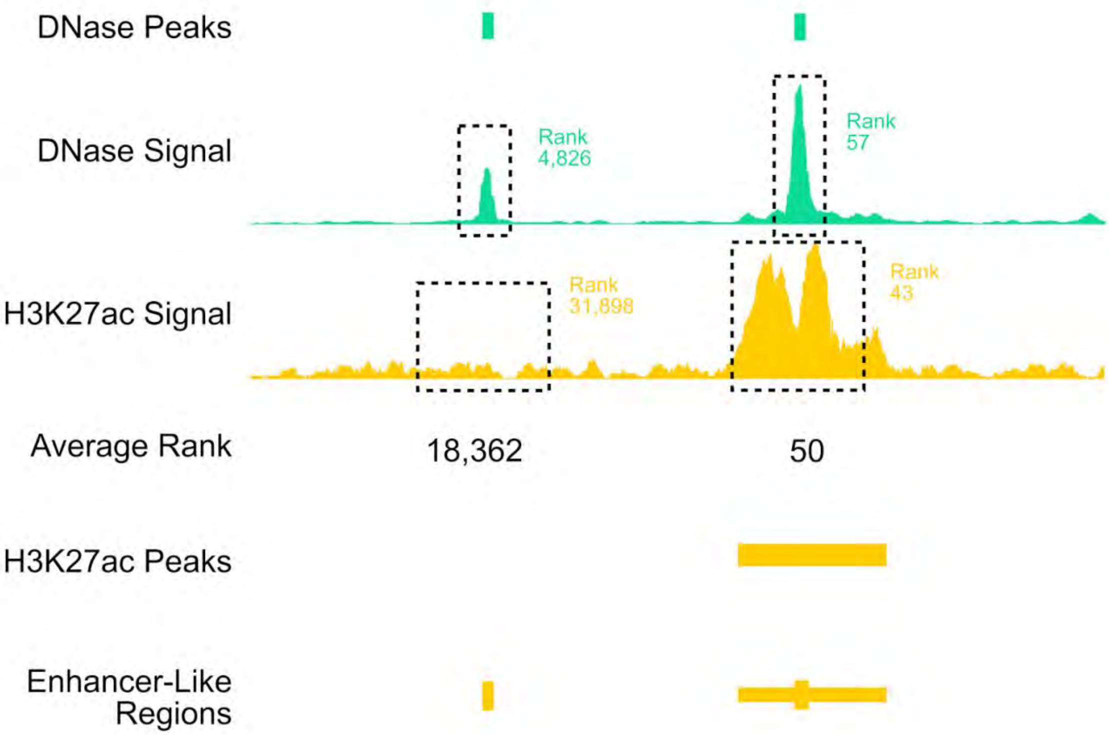


Figure II-9 | | Enhancer prediction using the average ranks
For each tissue, we sorted DNase peaks by the average rank of the DNase signal (green) and the H3K27ac signal (yellow) and estimated enhancer boundaries using the overlapping H3K27ac peaks.

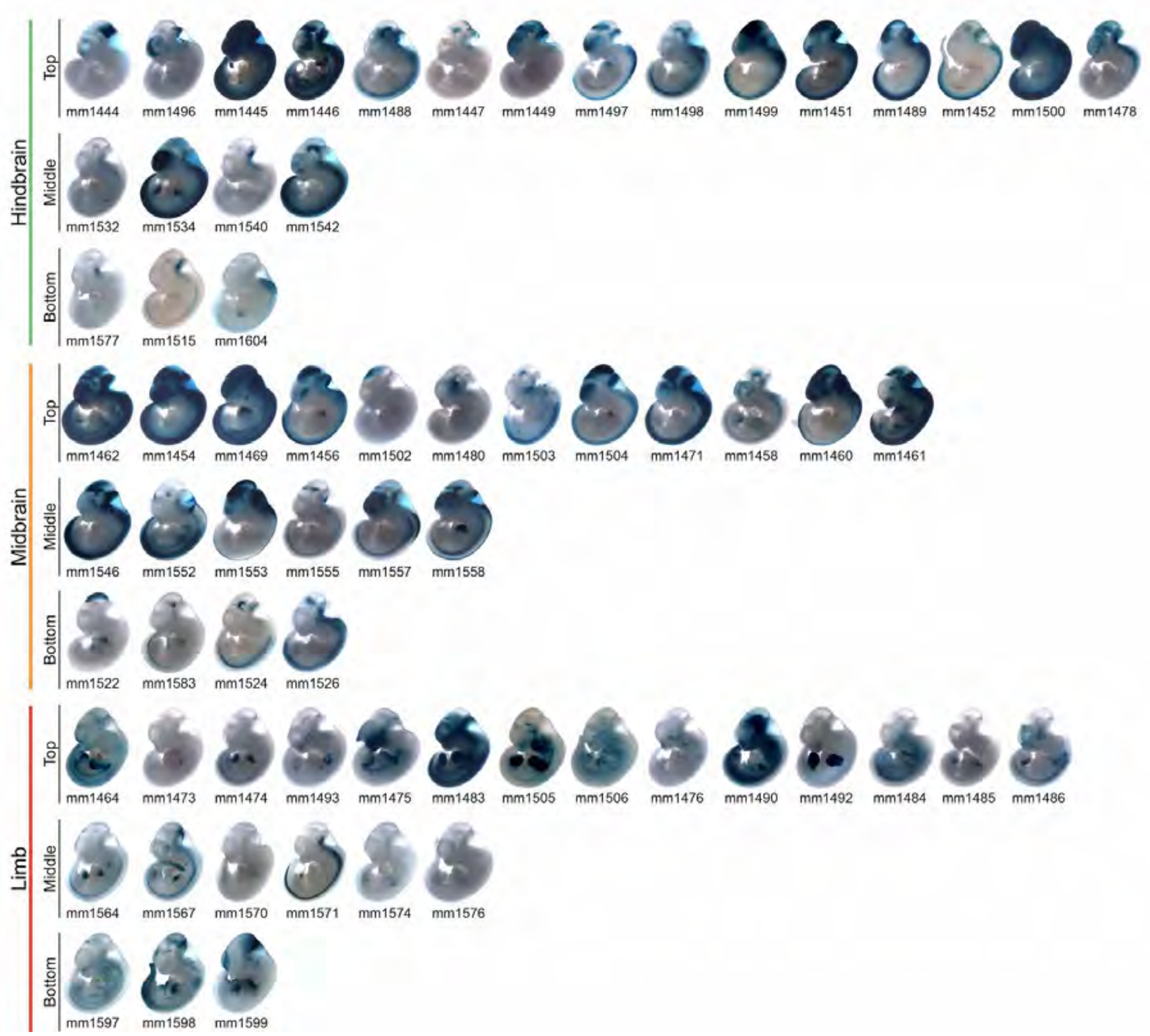

Hindbrain
Top
mm1444 mm1496 mm1445 mm1446 mm1488 mm1447 mm1449 mm1497 mm1498 mm1499 mm1451 mm1489 mm1452 mm1500 mm1478
Middle
mm1532 mm1534 mm1540 mm1542
Bottom
mm1577 mm1515 mm1604
Midbrain
Top
mm1462 mm1454 mm1469 mm1456 mm1502 mm1480 mm1503 mm1504 mm1471 mm1458 mm1460 mm1461
Middle
mm1546 mm1552 mm1553 mm1555 mm1557 mm1558
Bottom
mm1522 mm1583 mm1524 mm1526
Limb
Top
mm1464 mm1473 mm1474 mm1493 mm1475 mm1483 mm1505 mm1506 mm1476 mm1490 mm1492 mm1484 mm1485 mm1486
Middle
mm1564 mm1567 mm1570 mm1571 mm1574 mm1576
Bottom
mm1597 mm1598 mm1599

Figure II-10 | In vivo validation of ENCODE-predicted enhancers
Shown are representative transgenic embryonic day 11.5 (e11.5) mouse images for all predicted enhancers that displayed reproducible activity in the expected tissue type. Enhancer predictions were performed using a combination of H3K27ac and DHS profiling for E11.5 mouse hindbrain, midbrain, and limb tissue. Predicted enhancers were selected for validation from three different rank classes (Top, Middle, Bottom) and tested for activity using transgenic mouse assays (see Methods for further details). Blue staining indicates enhancer activity, and the unique identifier below each embryo (mm number) corresponds to the name of the enhancer in the VISTA Enhancer Browser (www.enhancer.lbl.gov).

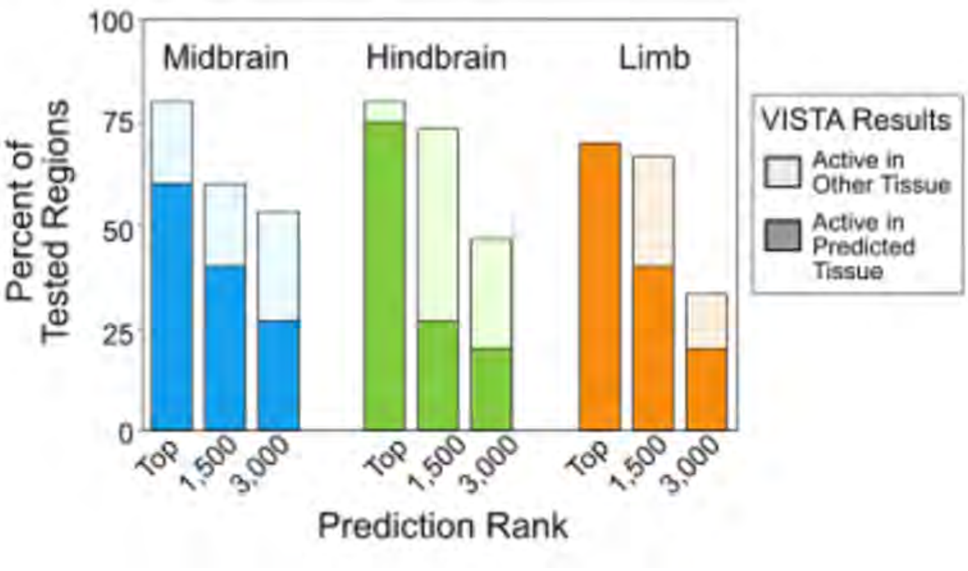


Figure II-11 | Validation rates
Validation rates of 151 enhancer-like regions tested using transgenic mouse assays. Dark color indicates the region was active in the predicted tissue while light color indicates a lack of activity in the predicted tissue but with activity in other tissues.

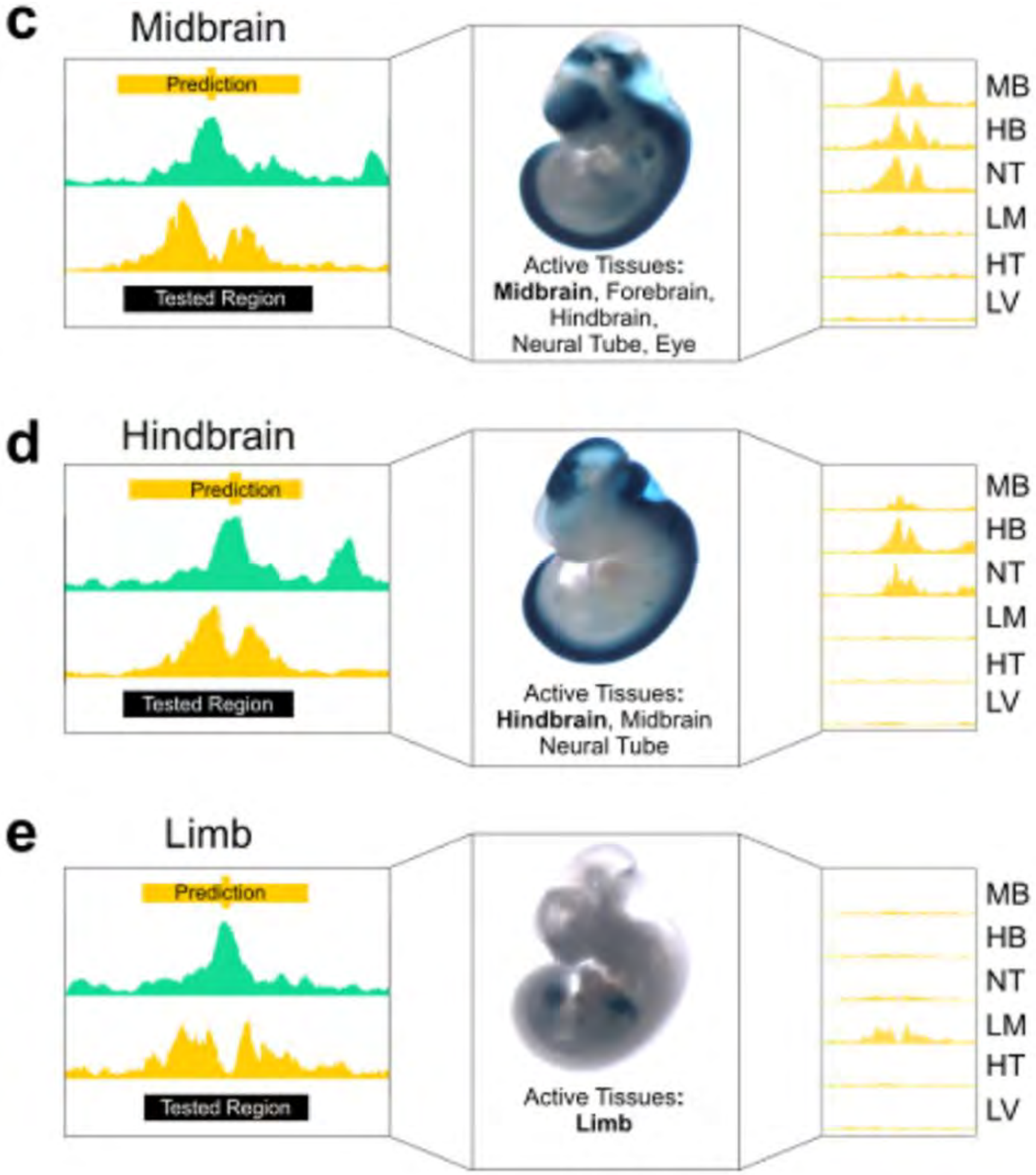


Figure II-12 | Prediction of mouse embryonic enhancers
c–e, Examples of enhancers (orange boxes) that were predicted based on DNase signal (green) and H3K27ac signal (orange) and validated in **c,** midbrain, **d,** hindbrain and **e,** limb. H3K27ac signal (yellow) in across tissues accurately predicts additional observed activity.

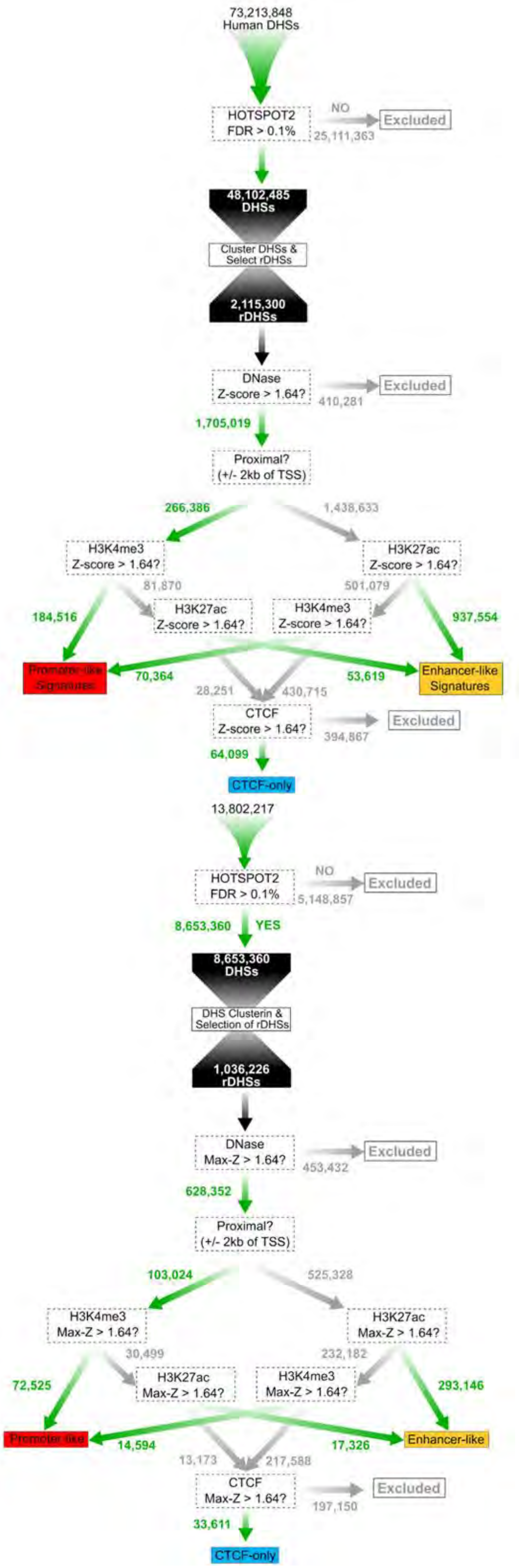


Figure II-13 | Selection of cREs

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 (n = 254,880), cREs with enhancer-like signatures (n = 991,173), and cREs bound by CTCF only (64,099). The total number of cREs is the sum of the three groups: 1,310,152.

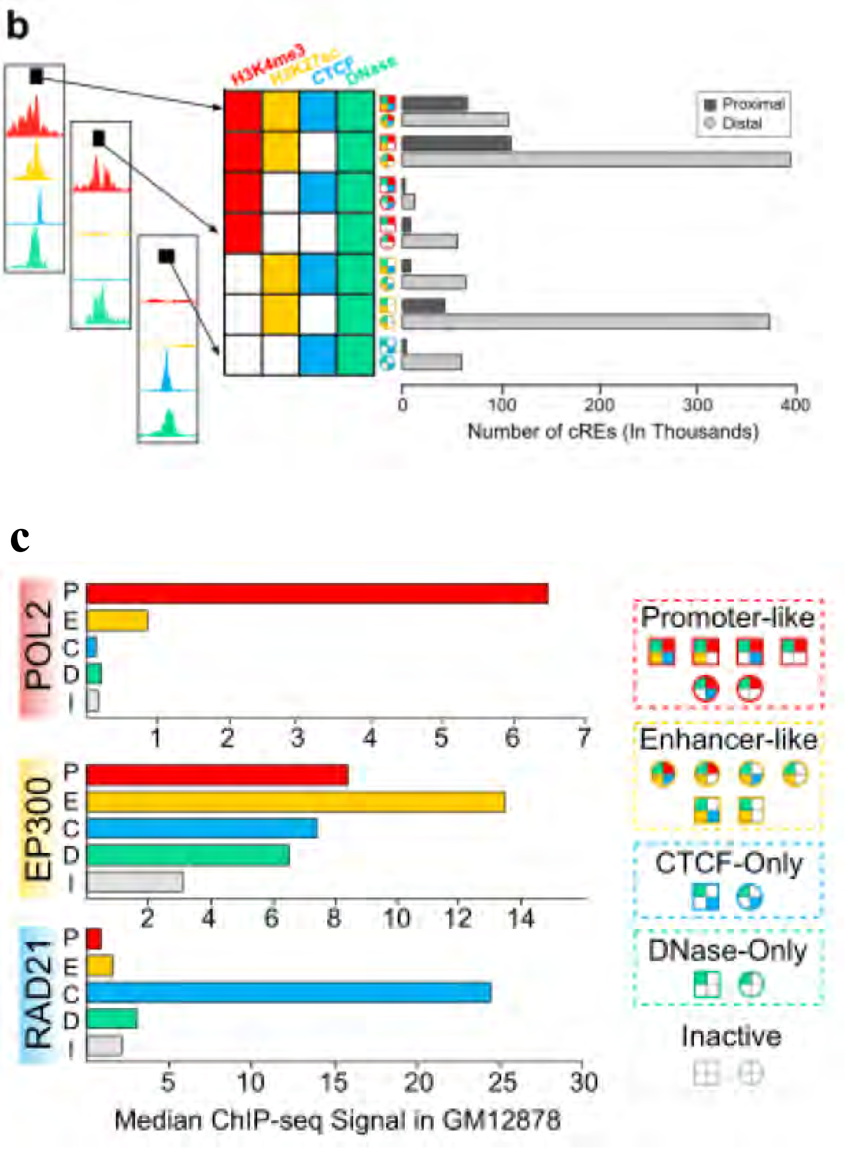


Figure II-14 | Assignment of cREs to cell type-specific 9 state and 5 group
b, Given a cell type (shown for GM12878), we assign cREs into nine states based on whether they have high Z-scores (> 1.64) for H3K4me3, H3K27ac, CTCF, and DNase in that cell type. Each cRE is either proximal (≤ 2 kb) or distal (> 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.

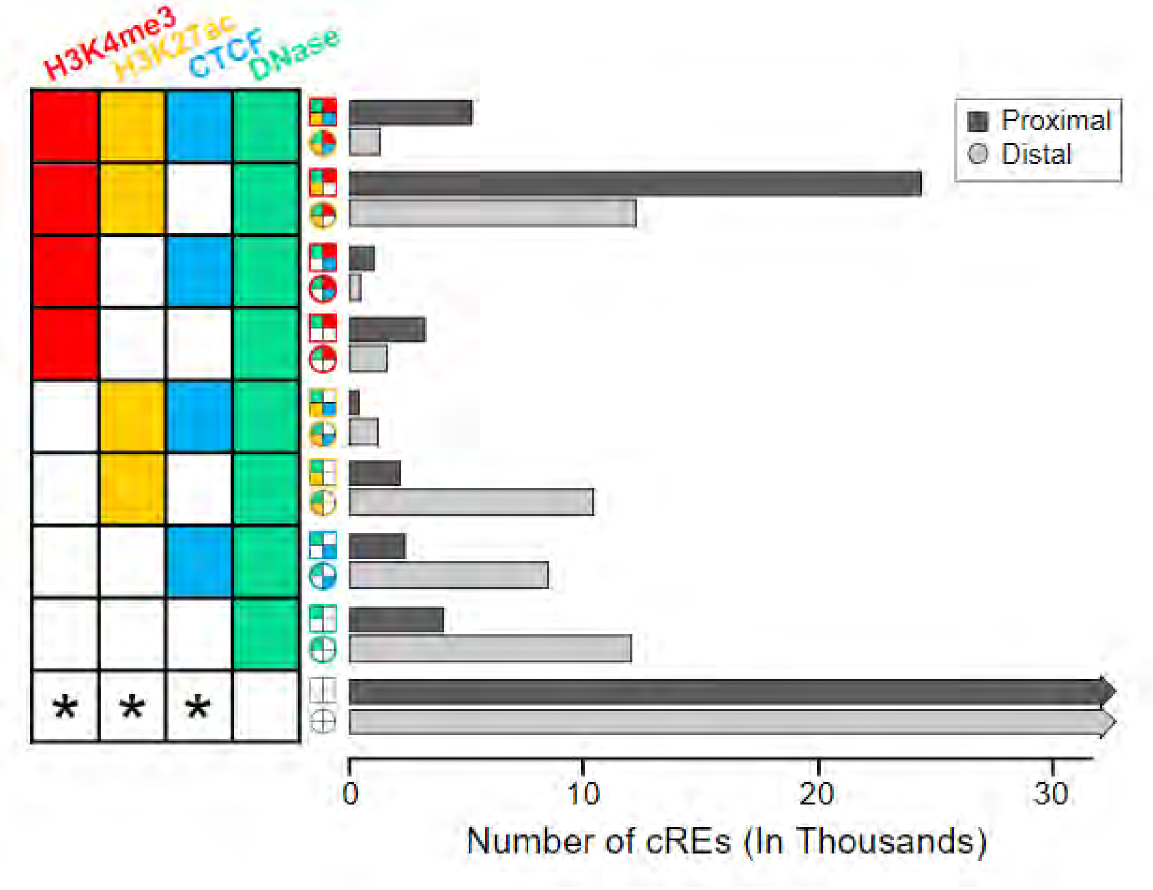
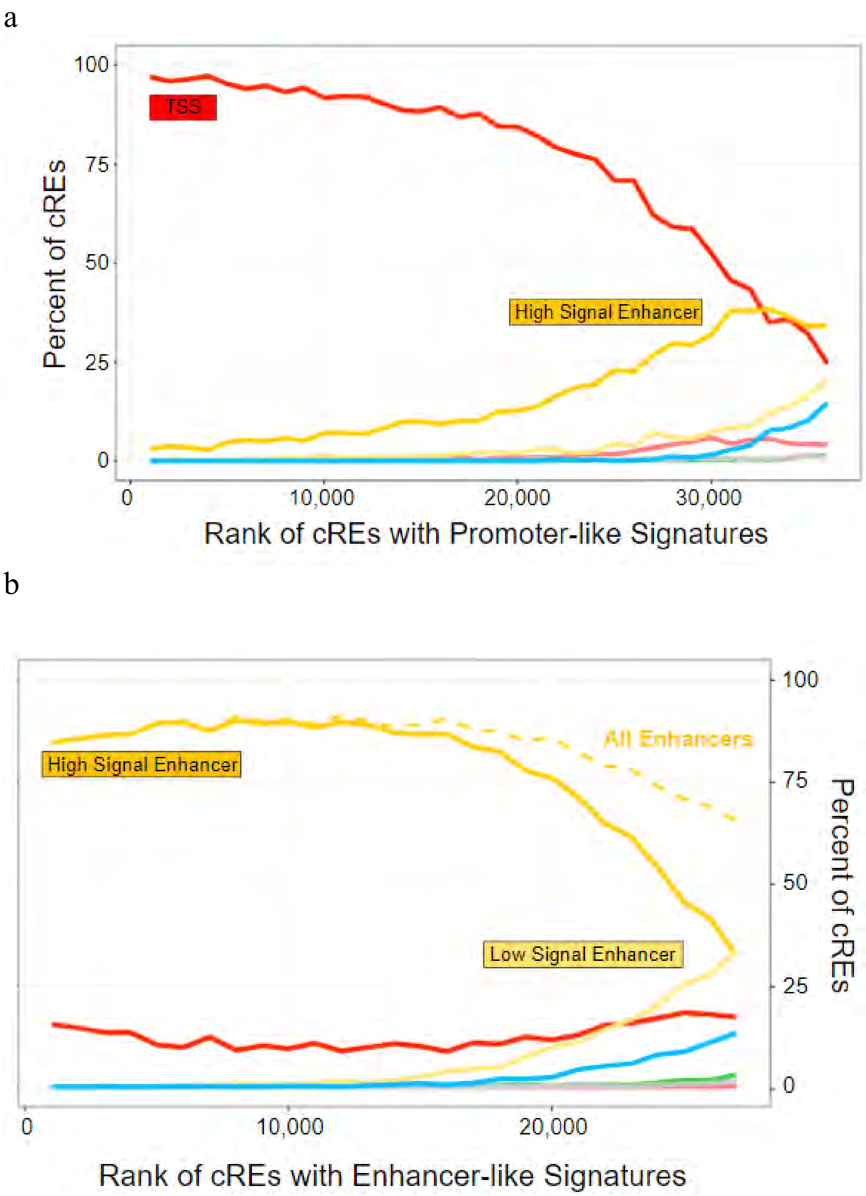


Figure II-15 | GM12878 cRE states
Summary the cREs states, with each state further stratified by TSS proximity. The bar graph reveals that cREs with high H3K4me3 are mostly TSS-proximal, regardless of whether or not they have high H3K27ac or CTCF, while cREs with low H3K4me3 are mostly TSS-distal.



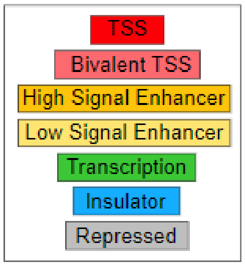


Figure II-16 | Overlap of cREs with chromHMM states
In GM12878, we ranked cREs with a, promoter-like signatures and b, enhancer-signatures by H3K4me3 and H3K27ac Z-scores respectively. For each bin of 1 k cREs, we calculated the percent of cREs overlapping each chromHMM state.

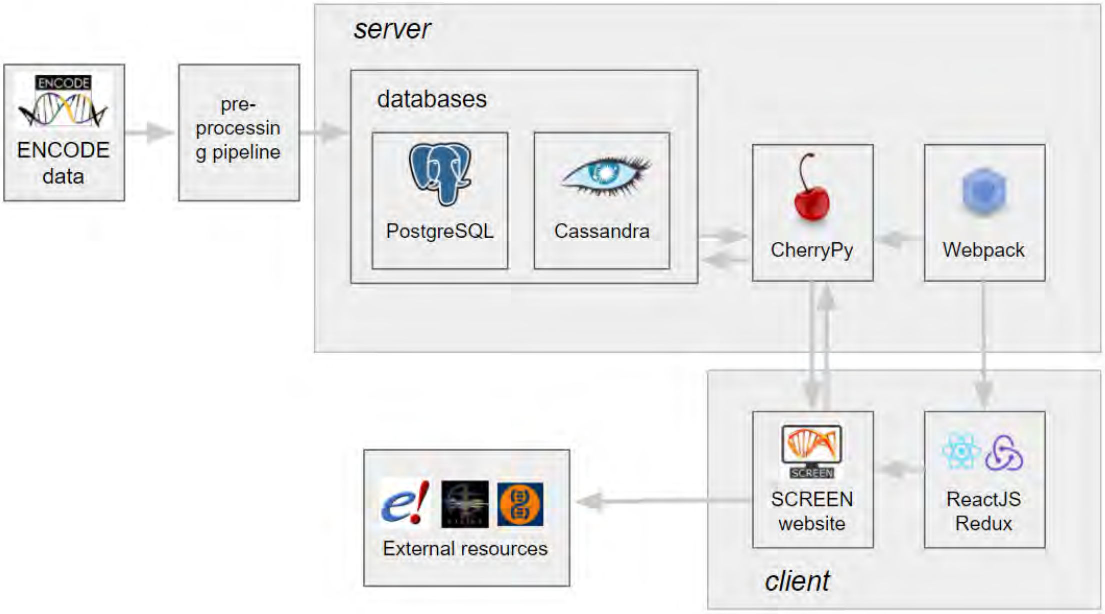
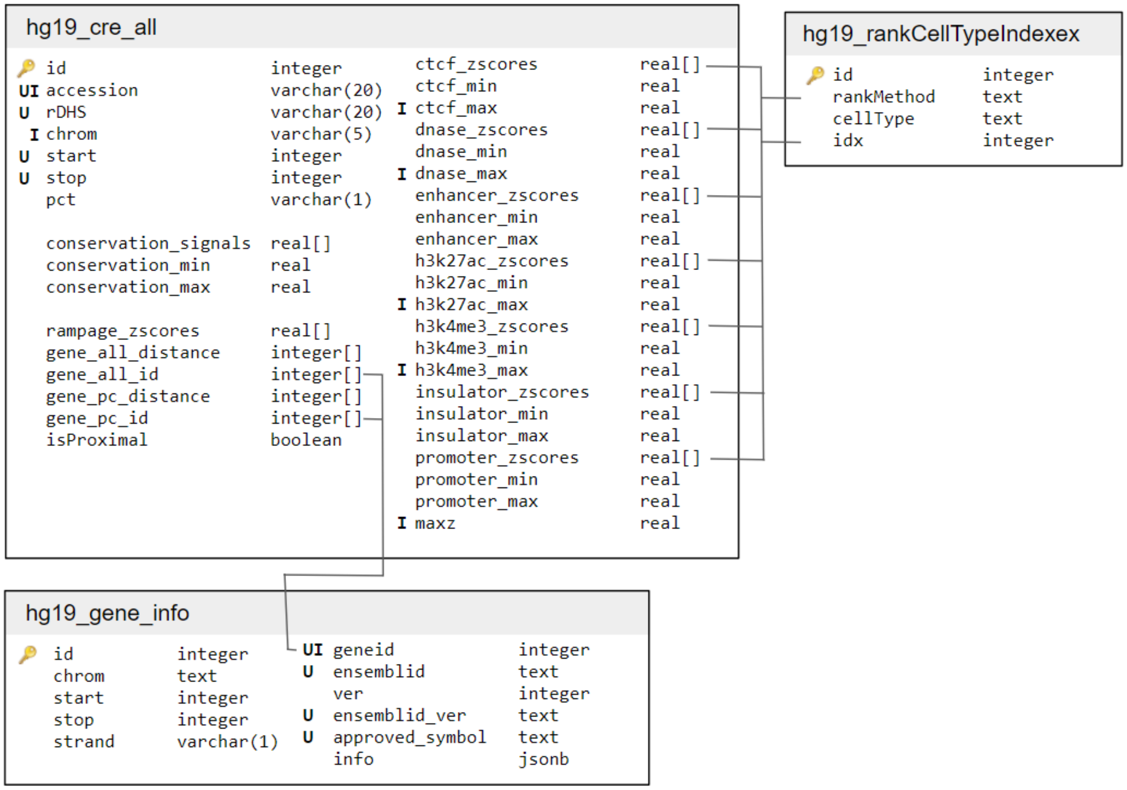


Figure II-17 | General architecture of SCREEN
ENCODE data is pre-processed by a C++ and Python pipeline before import into the PostgreSQL and Cassandra databases. SCREEN's webserver is CherryPy-based and serves ReactJS and Redux UI code compiled by a Webpack server. The client renders the majority of the UI and communicates with CherryPy via AJAX requests. In addition to SCREEN's own visualizations, links are provided to external resources including Ensembl, the UCSC Genome Browser, and GeneCards.



hg19_cre_all

🔑	id	integer		ctcf_zscores	real[]
UI	accession	varchar(20)		ctcf_min	real
U	rDHS	varchar(20)	I	ctcf_max	real
I	chrom	varchar(5)		dnase_zscores	real[]
U	start	integer		dnase_min	real
U	stop	integer	I	dnase_max	real
	pct	varchar(1)		enhancer_zscores	real[]
				enhancer_min	real
	conservation_signals	real[]		enhancer_max	real
	conservation_min	real		h3k27ac_zscores	real[]
	conservation_max	real		h3k27ac_min	real
			I	h3k27ac_max	real
	rampage_zscores	real[]		h3k4me3_zscores	real[]
	gene_all_distance	integer[]		h3k4me3_min	real
	gene_all_id	integer[]	I	h3k4me3_max	real
	gene_pc_distance	integer[]		insulator_zscores	real[]
	gene_pc_id	integer[]		insulator_min	real
	isProximal	boolean		insulator_max	real
				promoter_zscores	real[]
				promoter_min	real
				promoter_max	real
			I	maxz	real

hg19_rankCellTypeIndexex

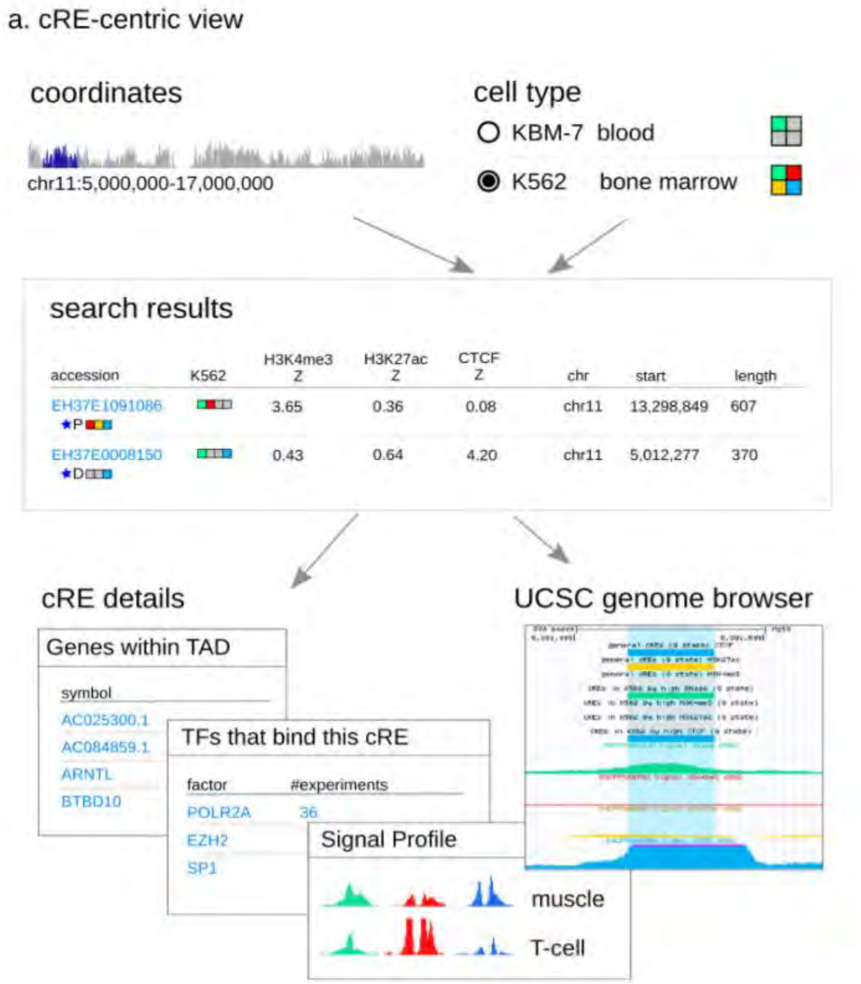
🔑	id	integer
	rankMethod	text
	cellType	text
	idx	integer

hg19_gene_info

🔑	id	integer	UI	geneid	integer
	chrom	text	U	ensemblid	text
	start	integer		ver	integer
	stop	integer	U	ensemblid_ver	text
	strand	varchar(1)	U	approved_symbol	text
				info	jsonb

Figure II-18 | Core of SCREEN's PostgreSQL schema

Key icons indicate primary keys; U indicates a field guaranteed to be unique for a row; I indicates an indexed field; square brackets denote array types. The cREs themselves are stored in the hg19_cre_all table; identifying information (accession, coordinates) are in the top of the left column, with information related to conservation and nearby genes in the center and bottom left, respectively, and Z-scores for the core four marks in the right column. Indices for values in the Z-score arrays are stored in the hg19_rankCellTypeIndexex table's idx field by assay (rankMethod field). Nearest gene_all (coding and non-coding) and gene_pc (coding only) IDs correspond to the hg19_gene_info table's geneid field.



cell type

○	KBM-7	blood	▦
●	K562	bone marrow	▦

search results

accession	K562	H3K4me3 Z	H3K27ac Z	CTCF Z	chr	start	length
EH37E1091086 ★P	▦	3.65	0.36	0.08	chr11	13,298,849	607
EH37E0008150 ★D	▦	0.43	0.64	4.20	chr11	5,012,277	370

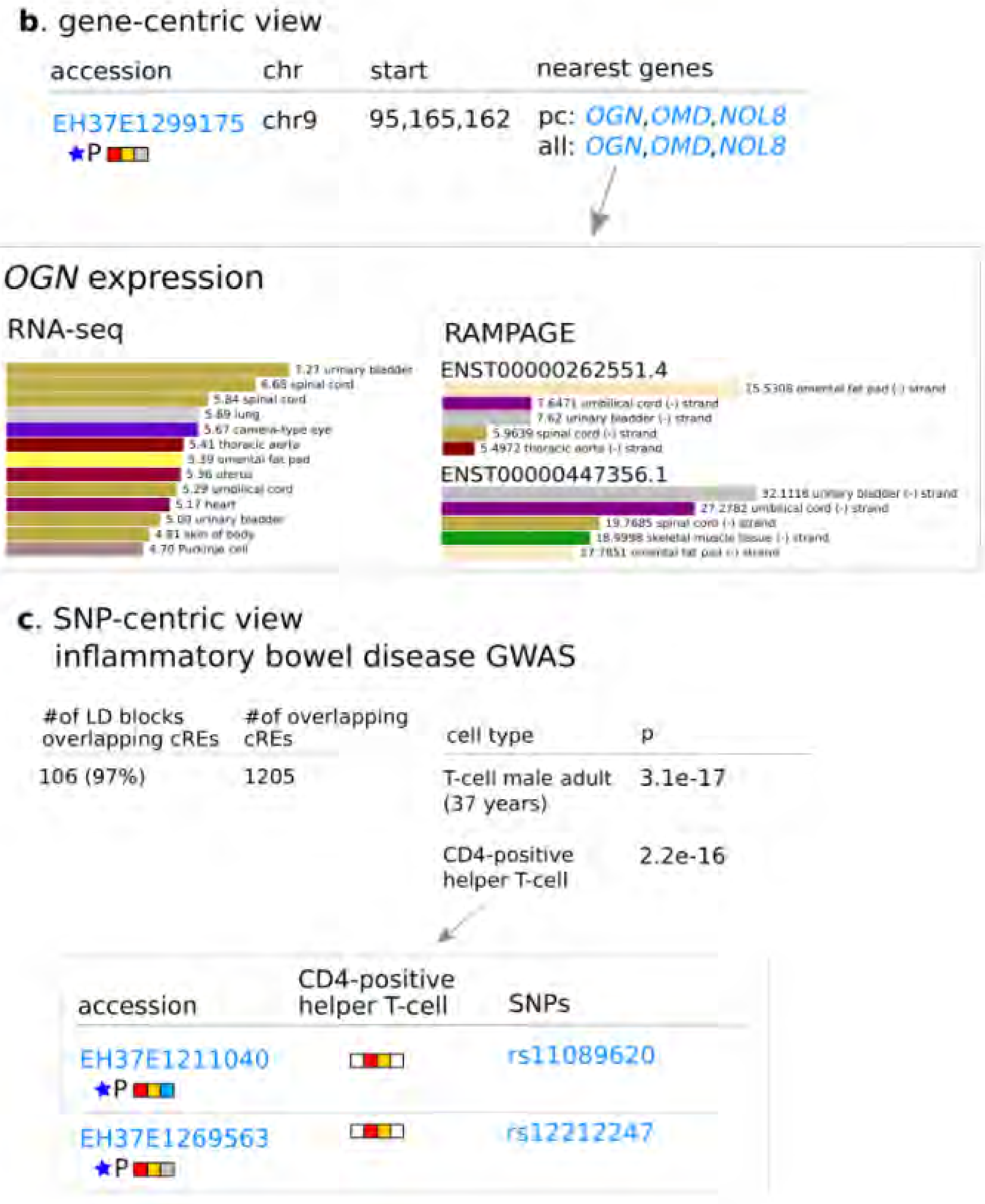
Genes within TAD

symbol
AC025300.1
AC084859.1
ARNTL
BTBD10

TFs that bind this cRE

factor	#experiments
POLR2A	36
EZH2	
SP1	

Figure II-19 | Overview of SCREEN cRE-centric search view
Using the facets on the main search page (top), the user can retrieve cREs (center) by genomic coordinates and activity profiles in a particular cell type; here, two cREs active in K562 are shown on chromosome 11. Both cREs 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 cRE is marked with a "P", indicating that it is promoter-proximal (within 2 kb of an annotated promoter); the bottom cRE is marked with a “D” for promoter-distal. Four colors correspond to high values (>1.64) for the four epigenetic signals: DNase (green), H3K4me3 (red), H3K27ac (yellow), CTCF (blue). Gray indicates a Z-score below 1.64 for the given mark. The cRE details view shows neighboring genes, bound transcription factors, 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. gene-centric view

accession	chr	start	nearest genes
EH37E1299175 ★P	chr9	95,165,162	pc: *OGN*,*OMD*,*NOL8* all: *OGN*,*OMD*,*NOL8*

OGN expression

RNA-seq
7.27 urinary bladder
6.65 spinal cord
5.84 spinal cord
5.69 lung
5.67 camera-type eye
5.41 thoracic aorta
5.39 omental fat pad
5.36 uterus
5.29 umbilical cord
5.17 heart
5.09 urinary bladder
4.81 skin of body
4.70 Purkinje cell
RAMPAGE
ENST00000262551.4
15.5308 omental fat pad (-) strand
7.6471 umbilical cord (-) strand
7.62 urinary bladder (-) strand
5.9639 spinal cord (-) strand
5.4972 thoracic aorta (-) strand
ENST00000447356.1
32.1116 urinary bladder (-) strand
27.2782 umbilical cord (-) strand
19.7685 spinal cord (-) strand
18.9998 skeletal muscle tissue (-) strand
17.7851 omental fat pad (-) strand

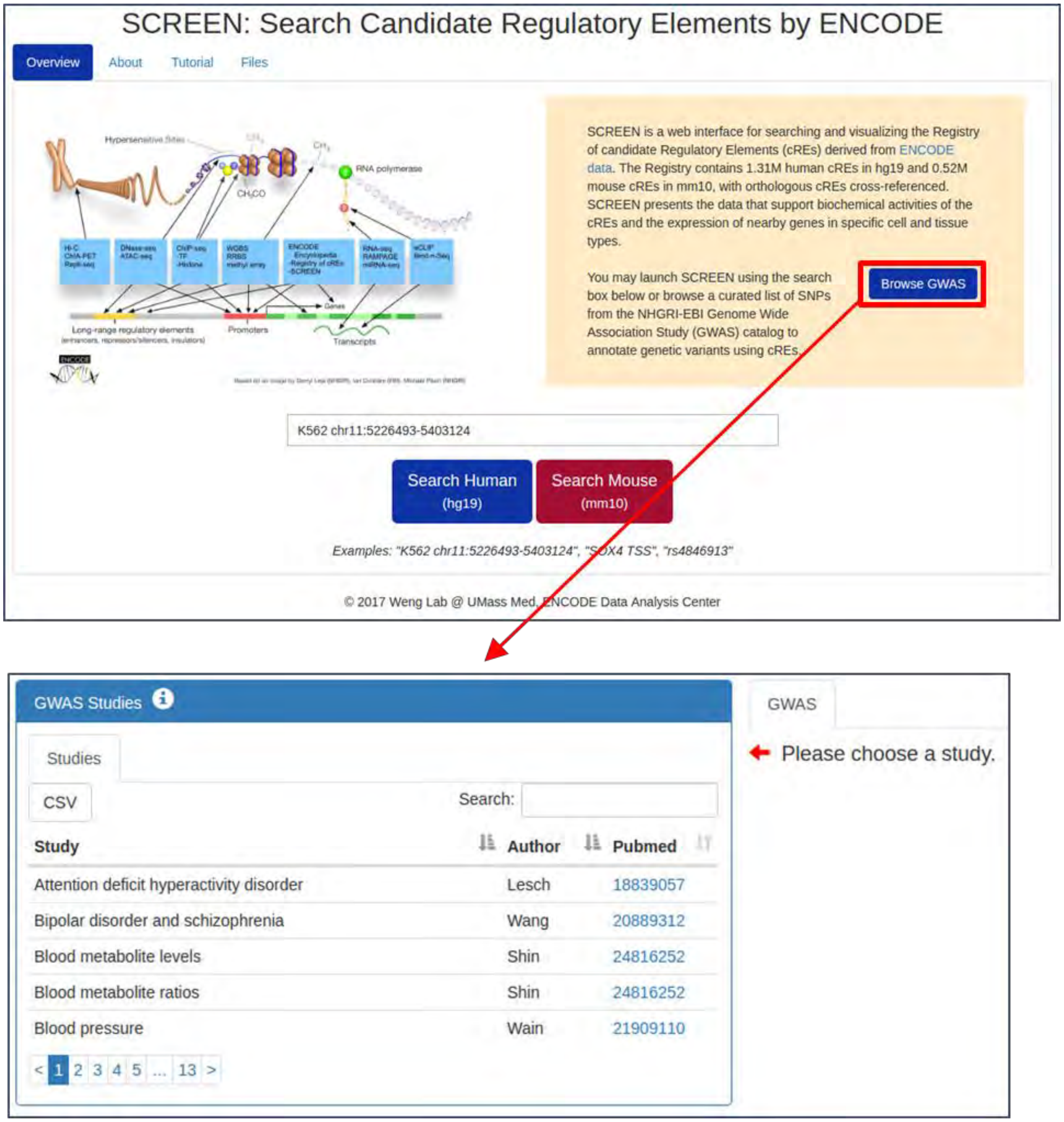
c. SNP-centric view
inflammatory bowel disease GWAS

#of LD blocks overlapping cREs	#of overlapping cREs
106 (97%)	1205

cell type	p
T-cell male adult (37 years)	3.1e-17
CD4-positive helper T-cell	2.2e-16

accession	CD4-positive helper T-cell	SNPs
EH37E1211040 ★P		rs11089620
EH37E1269563 ★P		rs12212247

Figure II-20 | Overview of SCREEN gene-centric view
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.



SCREEN: Search Candidate Regulatory Elements by ENCODE

SCREEN is a web interface for searching and visualizing the Registry of candidate Regulatory Elements (cREs) derived from ENCODE data. The Registry contains 1.31M human cREs in hg19 and 0.52M mouse cREs in mm10, with orthologous cREs cross-referenced. SCREEN presents the data that support biochemical activities of the cREs and the expression of nearby genes in specific cell and tissue types.

You may launch SCREEN using the search box below or browse a curated list of SNPs from the NHGRI-EBI Genome Wide Association Study (GWAS) catalog to annotate genetic variants using cREs.

K562 chr11:5226493-5403124

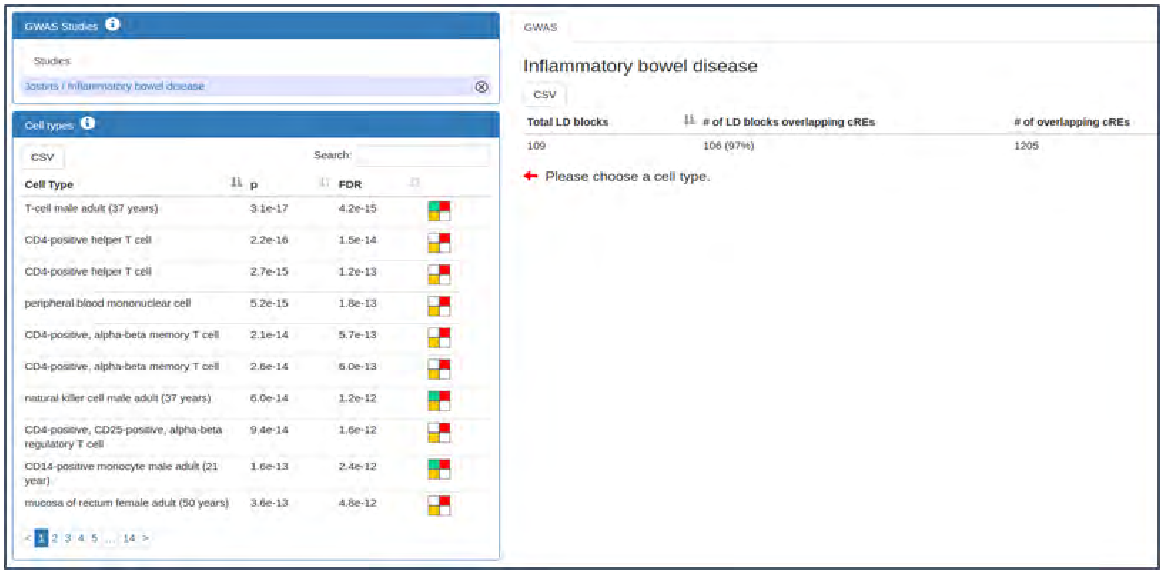
Examples: "K562 chr11:5226493-5403124", "SOX4 TSS", "rs4846913"

© 2017 Weng Lab @ UMass Med, ENCODE Data Analysis Center

Please choose a study.

Study	Author	Pubmed
Attention deficit hyperactivity disorder	Lesch	18839057
Bipolar disorder and schizophrenia	Wang	20889312
Blood metabolite levels	Shin	24816252
Blood metabolite ratios	Shin	24816252
Blood pressure	Wain	21909110

Figure II-21 | GWAS App
The "Browse GWAS" button on SCREEN's homepage (left) produces the GWAS app (right). Selecting a study displays the number of linkage disequilibrium blocks from the study which overlap a cRE, as well as the cell types with the most active cREs overlapping LD blocks.



Jostins / Inflammatory bowel disease

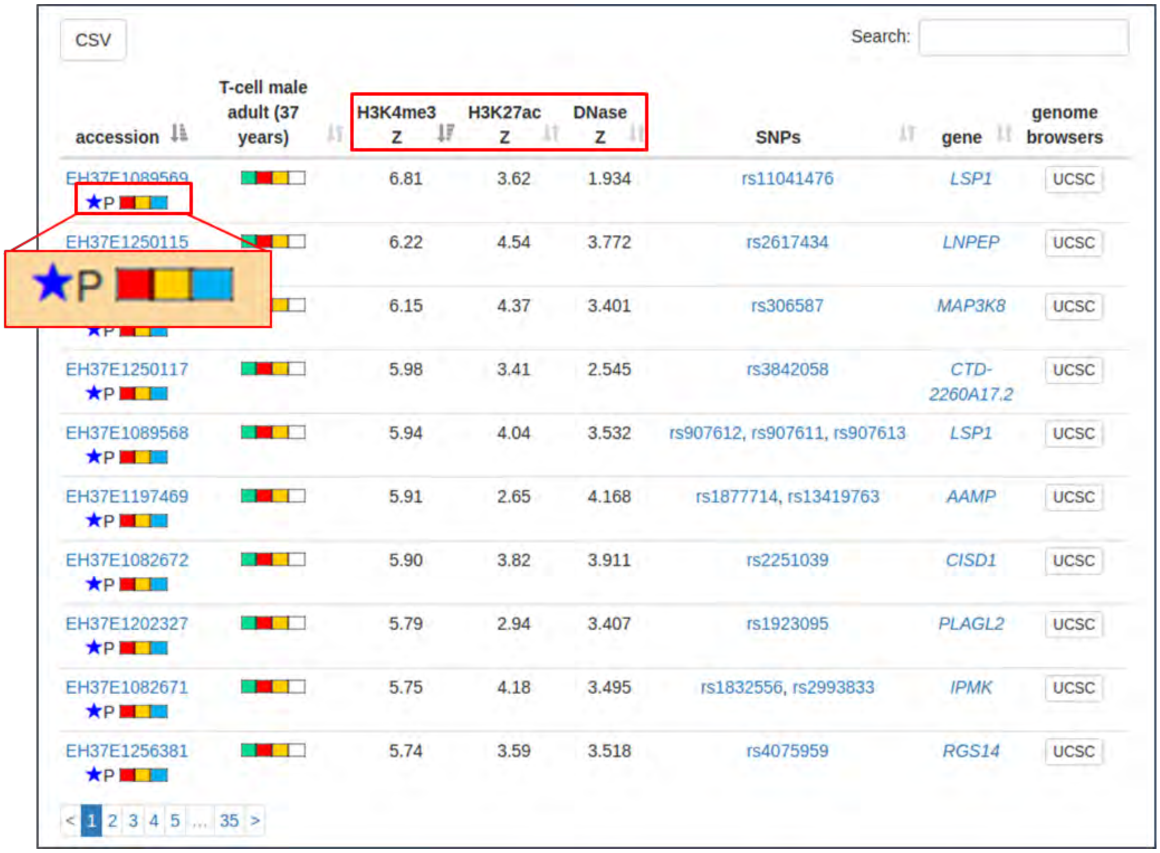
Cell Type	p	FDR	
T-cell male adult (37 years)	3.1e-17	4.2e-15	🟩🟥🟨⬜
CD4-positive helper T cell	2.2e-16	1.5e-14	⬜🟥🟨⬜
CD4-positive helper T cell	2.7e-15	1.2e-13	⬜🟥🟨⬜
peripheral blood mononuclear cell	5.2e-15	1.8e-13	⬜🟥🟨⬜
CD4-positive, alpha-beta memory T cell	2.1e-14	5.7e-13	⬜🟥🟨⬜
CD4-positive, alpha-beta memory T cell	2.6e-14	6.0e-13	⬜🟥🟨⬜
natural killer cell male adult (37 years)	6.0e-14	1.2e-12	🟩🟥🟨⬜
CD4-positive, CD25-positive, alpha-beta regulatory T cell	9.4e-14	1.6e-12	⬜🟥🟨⬜
CD14-positive monocyte male adult (21 year)	1.6e-13	2.4e-12	🟩🟥🟨⬜
mucosa of rectum female adult (50 years)	3.6e-13	4.8e-12	⬜🟥🟨⬜

Inflammatory bowel disease

Total LD blocks	# of LD blocks overlapping cREs	# of overlapping cREs
109	106 (97%)	1205

Please choose a cell type.

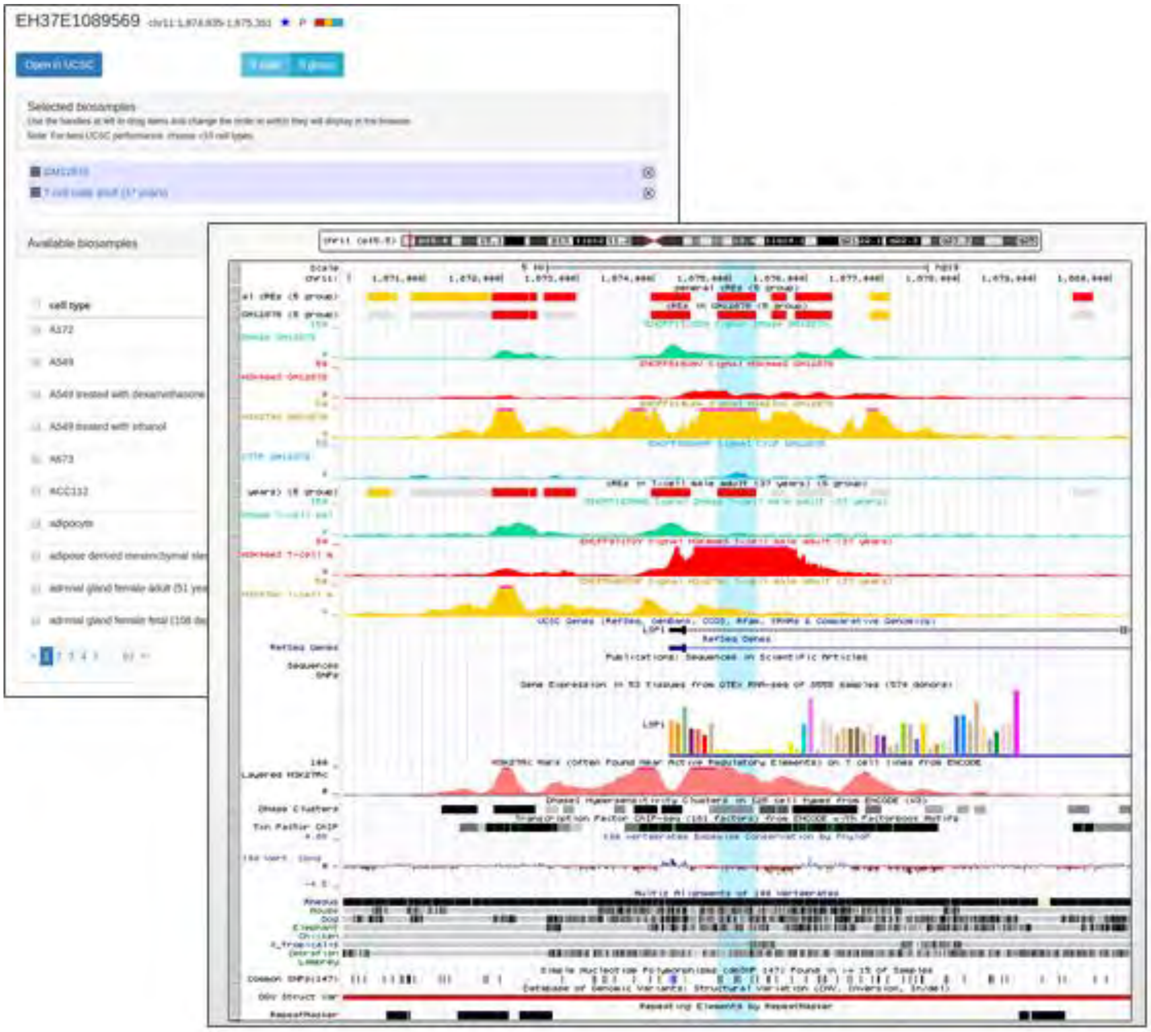
Figure II-22 | Results are shown here for a 2012 IBD GWAS study
Selecting a cell type produces a list of cREs active in the selected cell type which overlap SNPs from the study



accession	T-cell male adult (37 years)	H3K4me3 Z	H3K27ac Z	DNase Z	SNPs	gene	genome browsers
EH37E1089569 ★P 🟥🟨🟦	🟩🟥🟨⬜	6.81	3.62	1.934	rs11041476	*LSP1*	UCSC
EH37E1250115	🟩🟥🟨⬜	6.22	4.54	3.772	rs2617434	*LNPEP*	UCSC
★P 🟥🟨🟦	🟨⬜	6.15	4.37	3.401	rs306587	*MAP3K8*	UCSC
EH37E1250117 ★P 🟥🟨🟦	🟩🟥🟨⬜	5.98	3.41	2.545	rs3842058	*CTD-2260A17.2*	UCSC
EH37E1089568 ★P 🟥🟨🟦	🟩🟥🟨⬜	5.94	4.04	3.532	rs907612, rs907611, rs907613	*LSP1*	UCSC
EH37E1197469 ★P 🟥🟨🟦	🟩🟥🟨⬜	5.91	2.65	4.168	rs1877714, rs13419763	*AAMP*	UCSC
EH37E1082672 ★P 🟥🟨🟦	🟩🟥🟨⬜	5.90	3.82	3.911	rs2251039	*CISD1*	UCSC
EH37E1202327 ★P 🟥🟨🟦	🟩🟥🟨⬜	5.79	2.94	3.407	rs1923095	*PLAGL2*	UCSC
EH37E1082671 ★P 🟥🟨🟦	🟩🟥🟨⬜	5.75	4.18	3.495	rs1832556, rs2993833	*IPMK*	UCSC
EH37E1256381 ★P 🟥🟨🟦	🟩🟥🟨⬜	5.74	3.59	3.518	rs4075959	*RGS14*	UCSC

Figure II-23 | SCREEN Search Results

Results are shown here for a T-cell line from a 37 year-old male donor. The symbols in the left columns indicate that the top result, EH37E1089569, is proximal to a TSS ("P") and is enriched for H3K4me3 (red), H3K27ac (yellow), and CTCF (blue) ChIP-seq signal. Clicking the SNP and gene links lead to the corresponding Ensembl and GeneCards pages, respectively; clicking the cRE link performs a search for that cRE.



EH37E1089569 chr11:1,874,835-1,875,351

Selected biosamples

Use the handles at left to drag items and change the order in which they will display in the browser.

Note: For best UCSC performance, choose <10 cell types.

GM12878

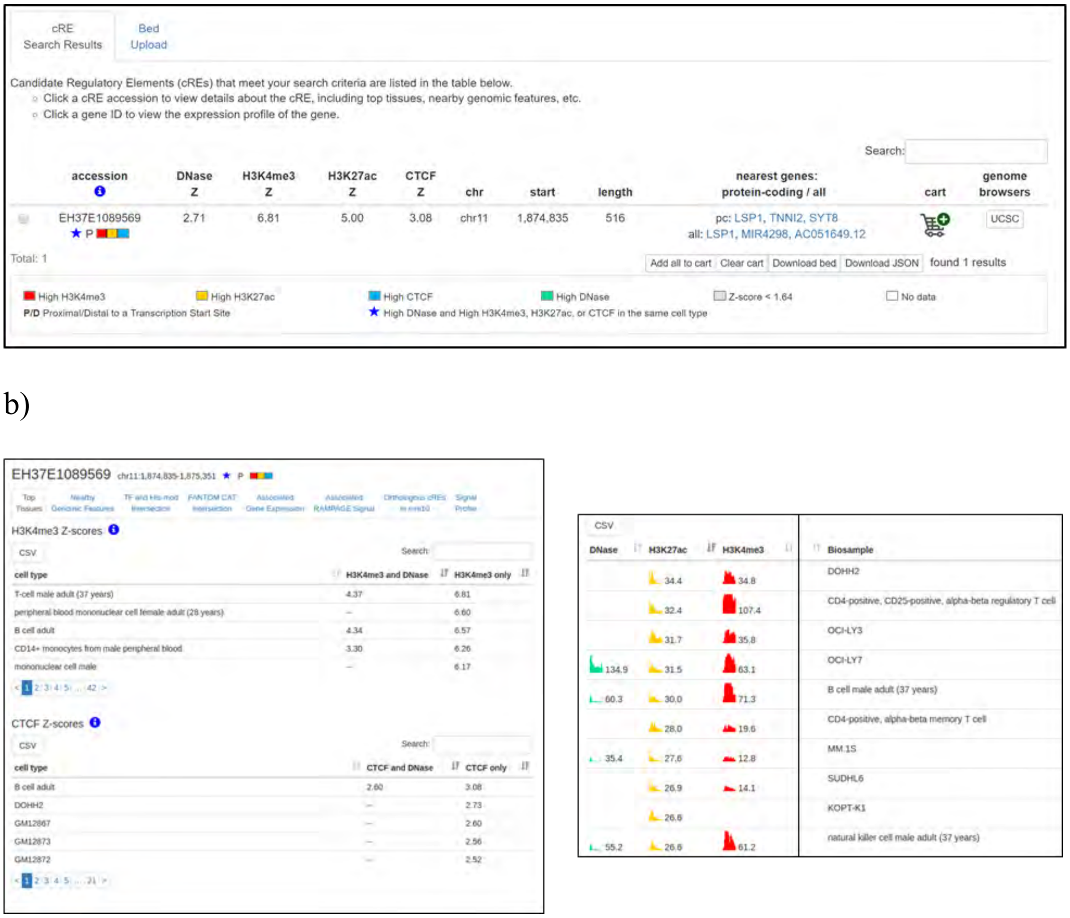
T-cell male adult (37 years)

Available biosamples

cell type
A172
A549
A549 treated with dexamethasone
A549 treated with ethanol
A673
ACC112
adipocyte
adipose derived mesenchymal ste
adrenal gland female adult (51 yea
adrenal gland female fetal (108 da

Figure II-24 | SCREEN and UCSC Genome Browser
The configure genome browser view (left), accessed via the UCSC buttons in the search results table, provides access to SCREEN's custom UCSC Genome Browser trackhubs (right). Shown here are the condensed 5-group tracks for GM12878 and a T-cell line. EH37E1089569 is highlighted in blue within the browser, and shows strong H3K4me3 and H3K27ac signal in the two selected immune cell types, with the 5-group tracks coloring EH47E1089569 red in these cell types to indicate that it is active and promoter-like. The RefSeq tracks show that this cRE falls less than 1kb downstream of the first transcription start site and exon of LSP1.

a)



Candidate Regulatory Elements (cREs) that meet your search criteria are listed in the table below.

- Click a cRE accession to view details about the cRE, including top tissues, nearby genomic features, etc.
- Click a gene ID to view the expression profile of the gene.

Search:

accession	DNase Z	H3K4me3 Z	H3K27ac Z	CTCF Z	chr	start	length	nearest genes: protein-coding / all	cart	genome browsers
EH37E1089569 ★ P 🟥🟨🟦	2.71	6.81	5.00	3.08	chr11	1,874,835	516	pc: LSP1, TNNI2, SYT8 all: LSP1, MIR4298, AC051649.12		UCSC

Total: 1

found 1 results

🟥 High H3K4me3 🟨 High H3K27ac 🟦 High CTCF 🟩 High DNase ▨ Z-score < 1.64 □ No data

P/D Proximal/Distal to a Transcription Start Site ★ High DNase and High H3K4me3, H3K27ac, or CTCF in the same cell type

b)

EH37E1089569 chr11:1,874,835-1,875,351 ★ P 🟥🟨🟦

H3K4me3 Z-scores

cell type	H3K4me3 and DNase	H3K4me3 only
T-cell male adult (37 years)	4.37	6.81
peripheral blood mononuclear cell female adult (28 years)	–	6.60
B cell adult	4.34	6.57
CD14+ monocytes from male peripheral blood	3.30	6.26
mononuclear cell male	–	6.17

< 1 2 3 4 5 ... 42 >

CTCF Z-scores

cell type	CTCF and DNase	CTCF only
B cell adult	2.60	3.08
DOHH2	–	2.73
GM12867	–	2.60
GM12873	–	2.56
GM12872	–	2.52

< 1 2 3 4 5 ... 21 >

DNase	H3K27ac	H3K4me3	Biosample
	34.4	34.8	DOHH2
	32.4	107.4	CD4-positive, CD25-positive, alpha-beta regulatory T cell
	31.7	35.8	OCI-LY3
134.9	31.5	63.1	OCI-LY7
60.3	30.0	71.3	B cell male adult (37 years)
	28.0	19.6	CD4-positive, alpha-beta memory T cell
35.4	27.6	12.8	MM.1S
	26.9	14.1	SUDHL6
	26.6		KOPT-K1
55.2	26.6	61.2	natural killer cell male adult (37 years)

Figure II-25 | SCREEN cRE Details View
a. EH37E1089569 displayed in the locus-centric search app after following the link from the GWAS app. b. cRE details view for EH37E1089569 showing Z-scores for H3K4me3 and CTCF, high in immune cells (left) and signal profile view revealing the underlying strong H3K4me3 and H3K27ac signals in various immune cell types (right).

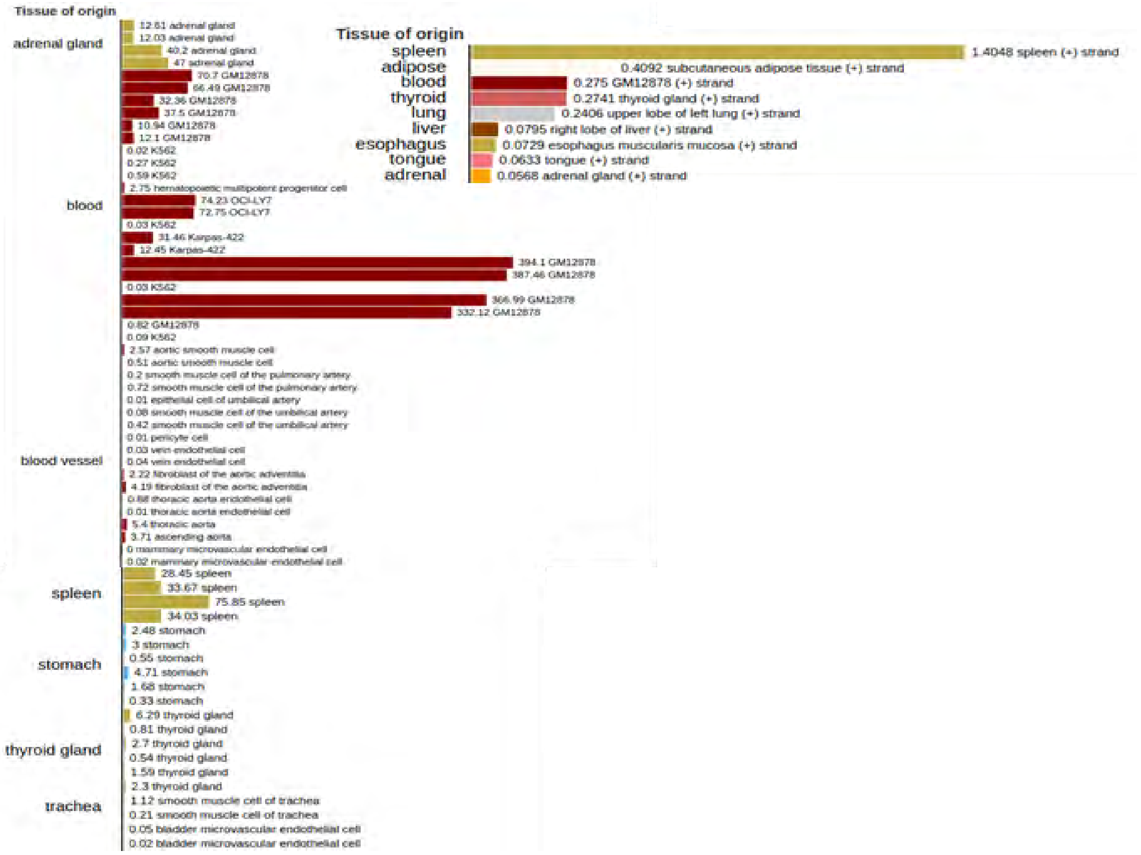


Figure II-26 | Gene expression view for LSP1
RNA-seq (left) reveals strong expression in immune cells (blood, red bars), spleen, and adrenal gland (gold bars); RAMPAGE (right) reveals that the TSS nearest EH37E1089569, corresponding to the first exon of LSP1, is most active in spleen.

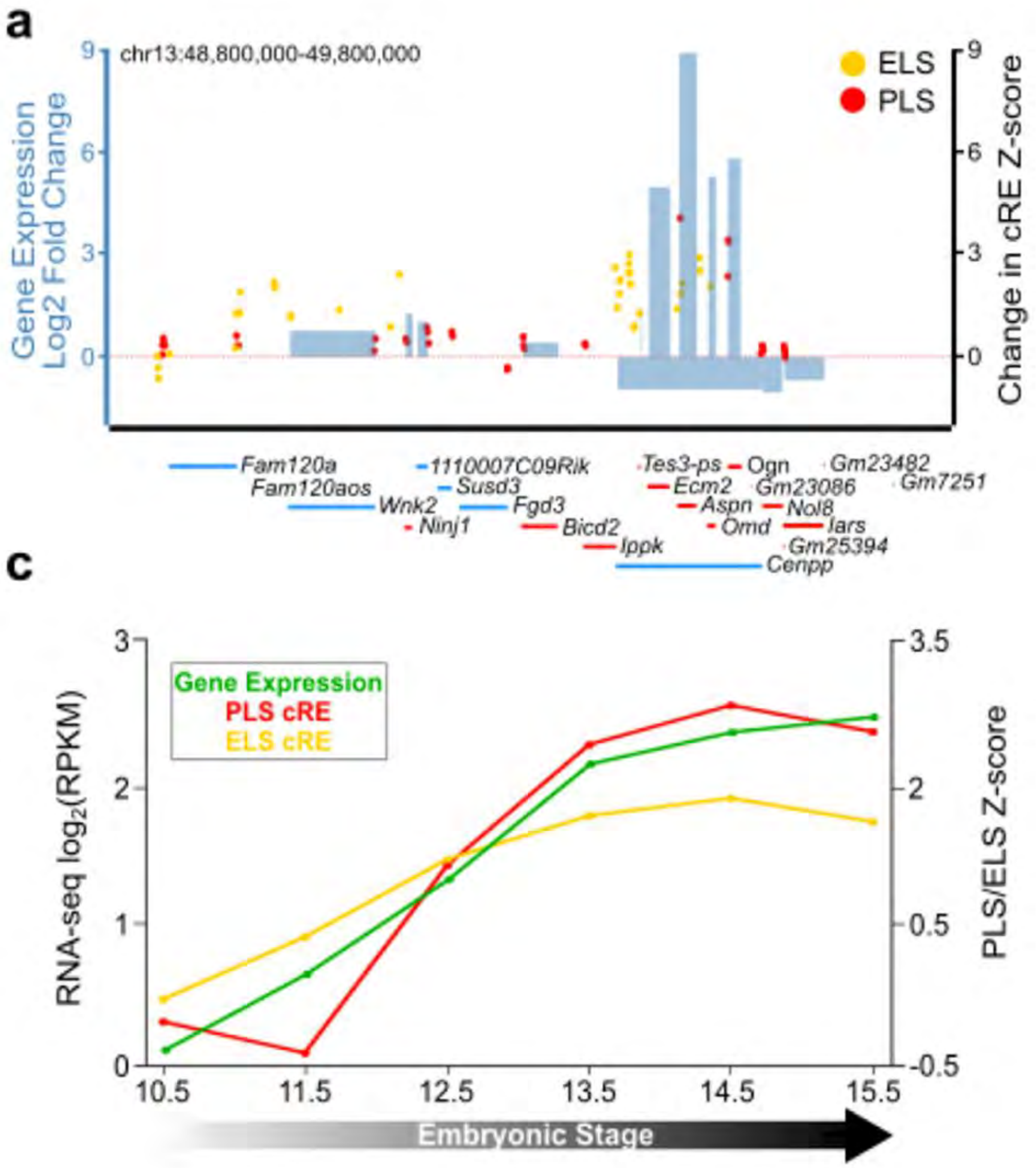


Figure II-27 | SCREEN Differential Gene Expression
Analyzing differential gene expression and cRE activity across developmental time points. a, Comparison between Limb e11.5 and e15.5 gene expression and cRE activity. Blue bars indicate differentially expressed genes while red and yellow dots indicate 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. 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.

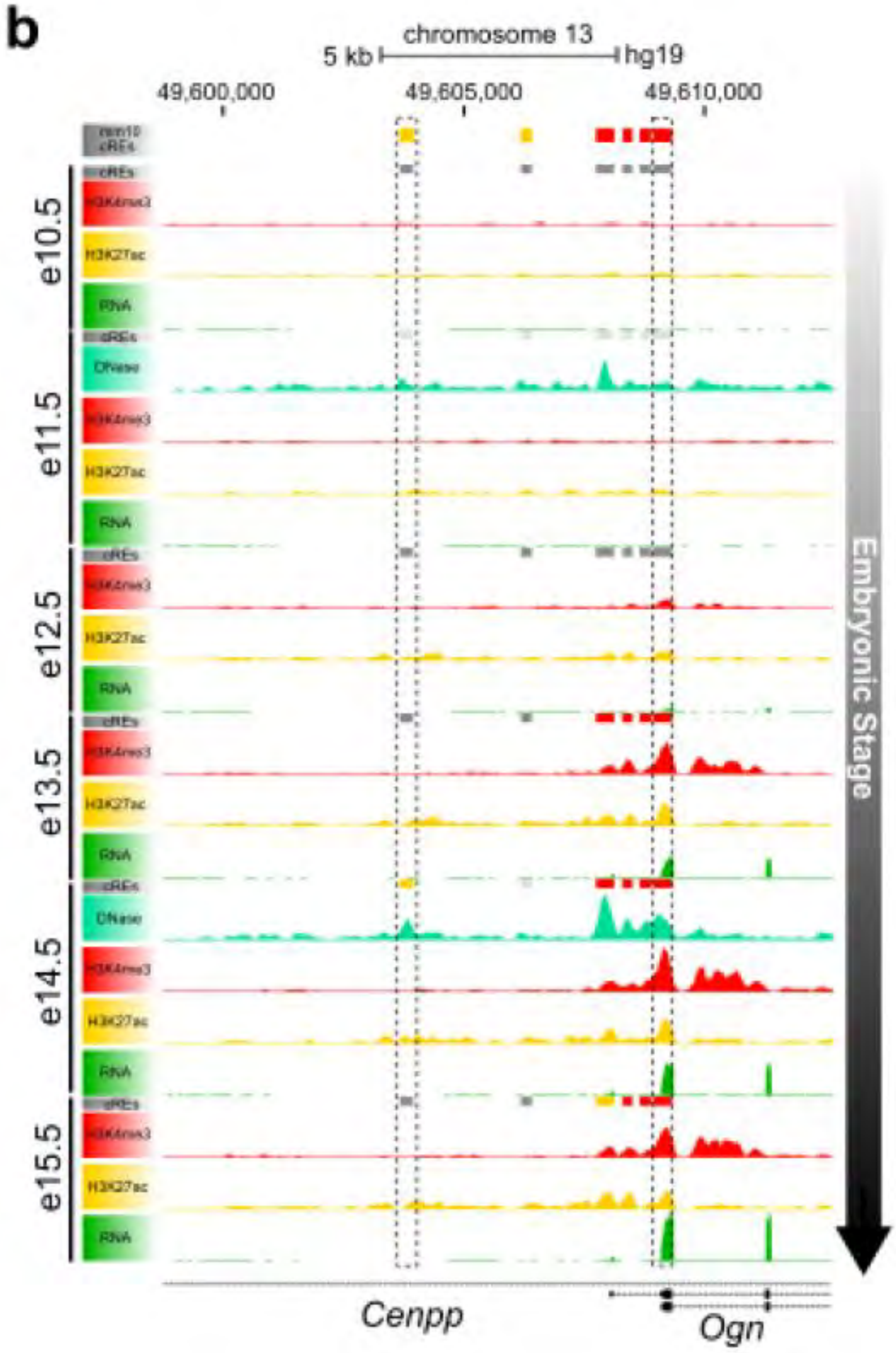


Figure II-28 | Signals around Ogn locus
Genome browser view of the Ogn locus with H3K27ac, H3K4me4, DNase, and RNA-seq signals for the limb across all surveyed time points. Promoter-like cREs are designated by red bars and enhancer-like cREs are designated by orange bars.

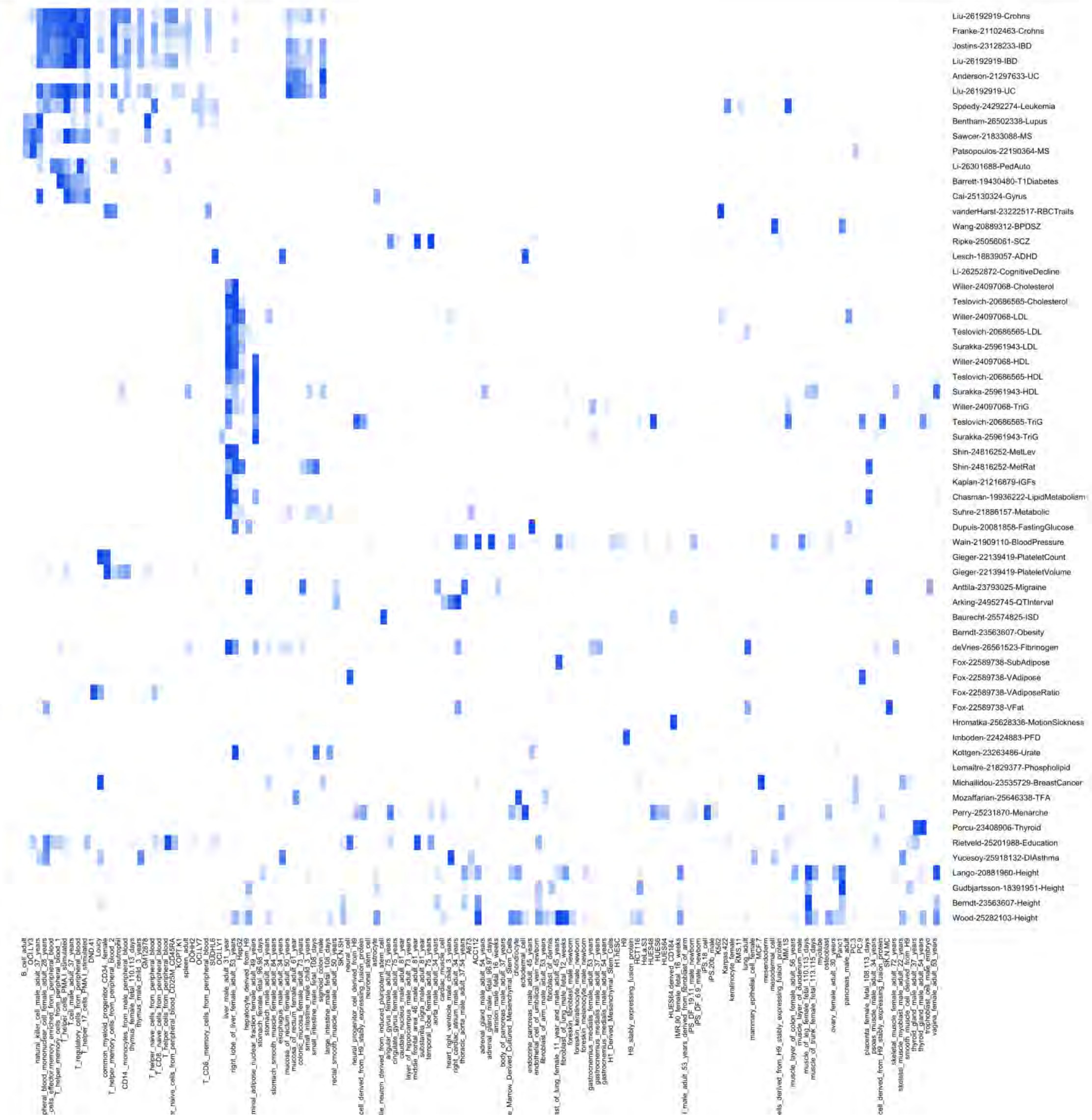


Figure II-29 | Overall cell type enrichments for variants reported by GWAS
Heatmap indicates enrichment a -log(p-value) of the variants associated with each disease (rows) in cREs active in each cell type (columns). Activity is defined as H3K27ac Z-score > 1.64. Color values in each row are scaled per study.

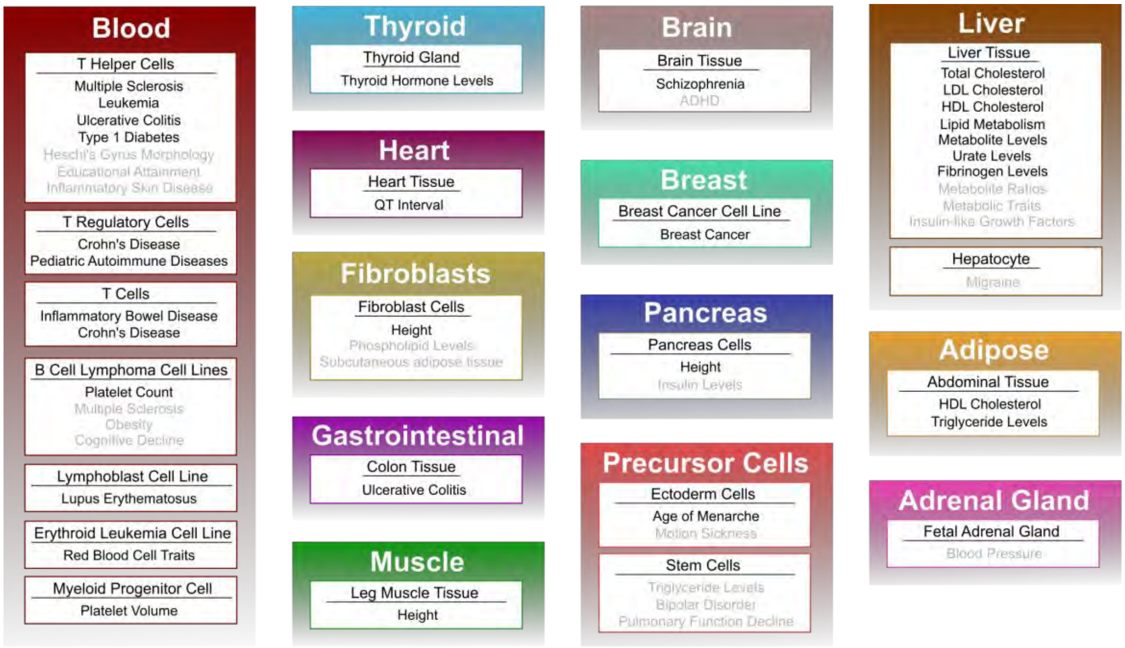
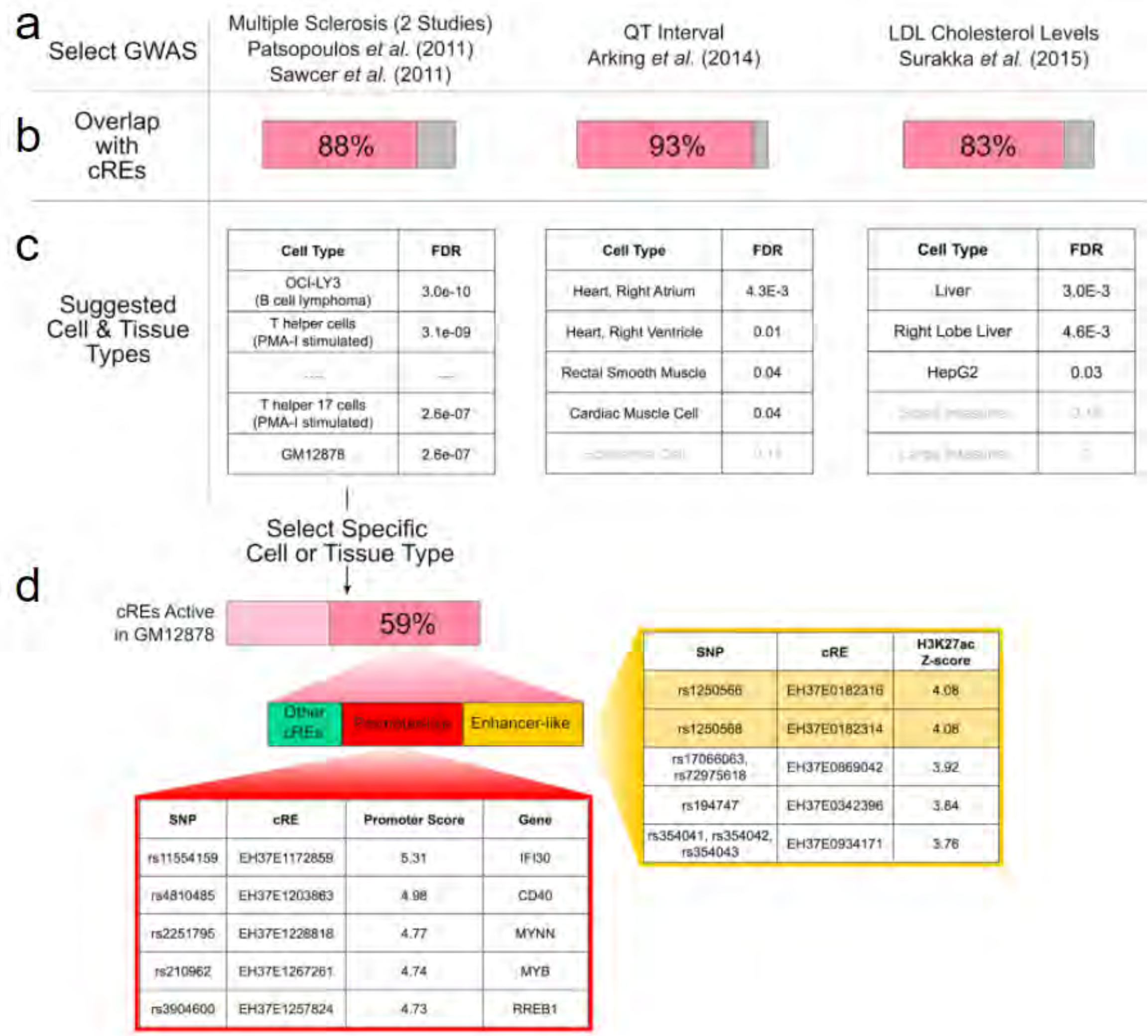


Figure II-30 | Top cell type enrichments for variants reported by GWAS
For each GWAS included in SCREEN, we report the cell or tissue type of which active cREs are significantly enriched in the disease variants. Cell types that do not meet FDR threshold of 0.05 are in gray. The majority of studies have multiple significantly enriched cell types but only the top hit is reported here. Traits listed multiple times are from different studies.



Cell Type	FDR
OCI-LY3 (B cell lymphoma)	3.0e-10
T helper cells (PMA-I stimulated)	3.1e-09
...	...
T helper 17 cells (PMA-I stimulated)	2.6e-07
GM12878	2.6e-07

Cell Type	FDR
Heart, Right Atrium	4.3E-3
Heart, Right Ventricle	0.01
Rectal Smooth Muscle	0.04
Cardiac Muscle Cell	0.04
[illegible]	[illegible]

Cell Type	FDR
Liver	3.0E-3
Right Lobe Liver	4.6E-3
HepG2	0.03
[illegible]	[illegible]
[illegible]	[illegible]

SNP	cRE	H3K27ac Z-score
rs1250566	EH37E0182316	4.08
rs1250568	EH37E0182314	4.08
rs17066063, rs72975618	EH37E0869042	3.92
rs194747	EH37E0342396	3.84
rs354041, rs354042, rs354043	EH37E0934171	3.76

SNP	cRE	Promoter Score	Gene
rs11554159	EH37E1172859	5.31	IFI30
rs4810485	EH37E1203863	4.98	CD40
rs2251795	EH37E1228818	4.77	MYNN
rs210962	EH37E1267261	4.74	MYB
rs3904600	EH37E1257824	4.73	RREB1

Figure II-31 | 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.

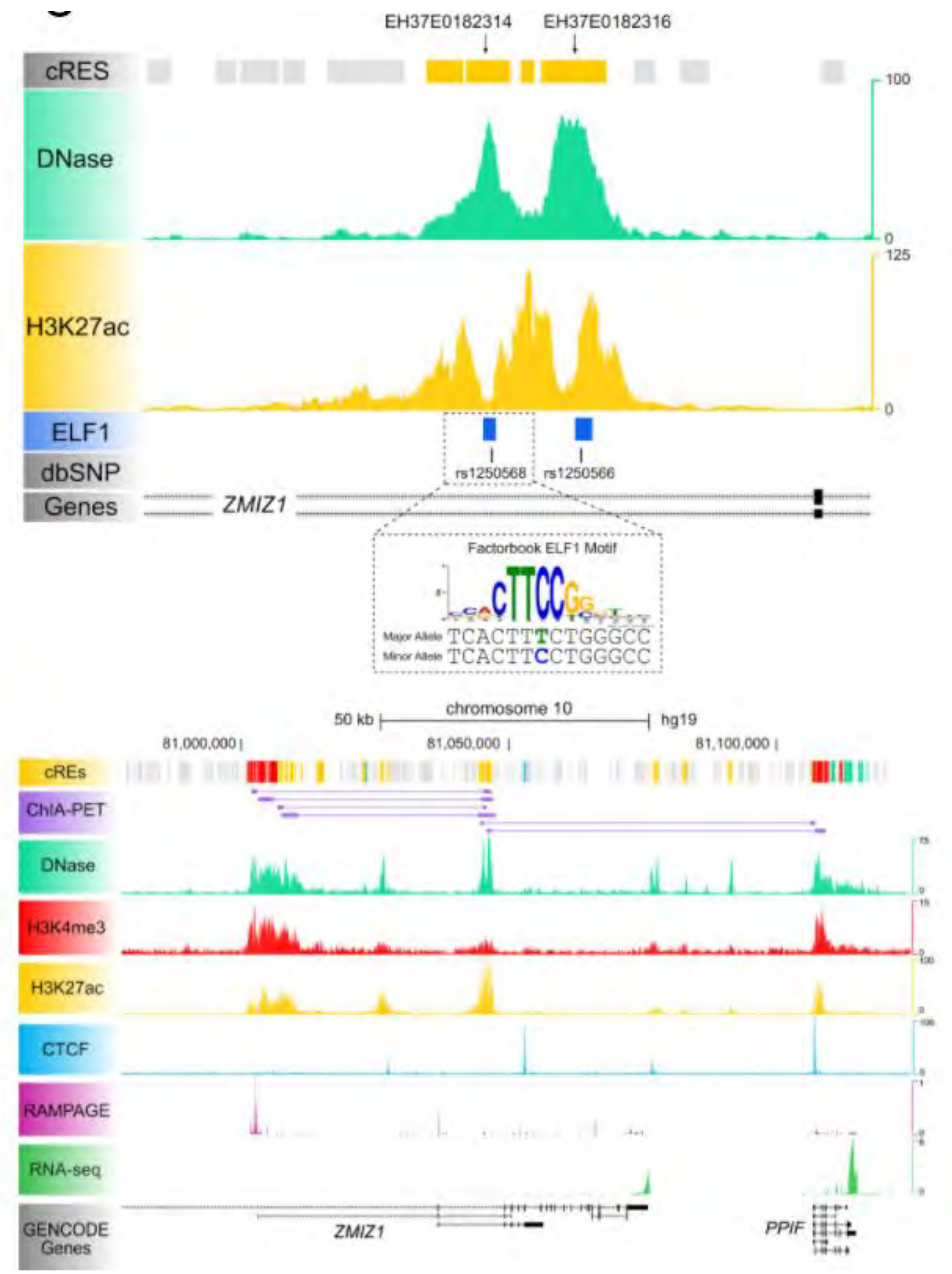


Figure II-32 | Annotating GWAS variants using SCREEN
e, Zoomed in genome browser view of MS-associated SNP rs1250568, which overlaps an ELF1 ChIP-seq peak (blue box) and an ELF1 motif. **f,** Zoomed out genome browser view of the locus showing POL2 ChIA-PET links between rs1250568 and two genes *ZMIZ1* and *PPIF*.

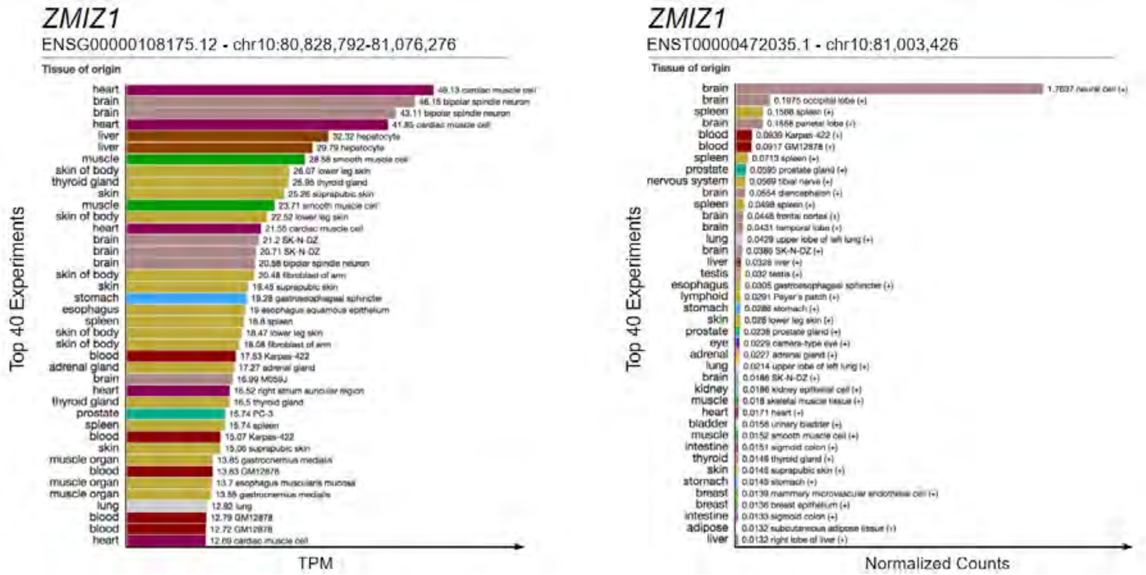


Figure II-33 | SCREEN display of the ZMIZ1 gene and TSS levels
a, Gene expression of *ZMIZ1* from whole-cell RNA-seq assays shown in tags per million (TPM). **b**, RAMPAGE signal at the TSS of ENST00000472035.1 (averaged over ± 50 bp window). Bars are colored by tissue of origin indicated on the left

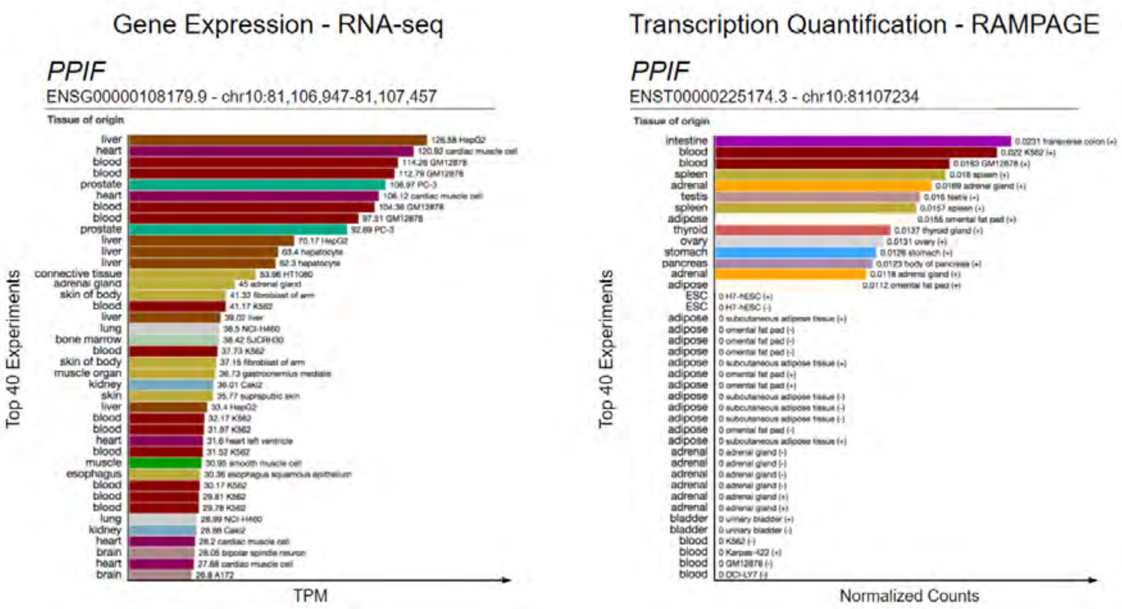


Figure II-34 | SCREEN display of PPIF gene and its TSS expression levels
a, Gene expression of *PPIF* from whole-cell RNA-seq assays shown in tags per million (TPM). **b**, RAMPAGE signal at the TSS of ENST00000225174.3 (averaged over ± 50 bp window). Bars are colored by tissue of origin indicated on the left.

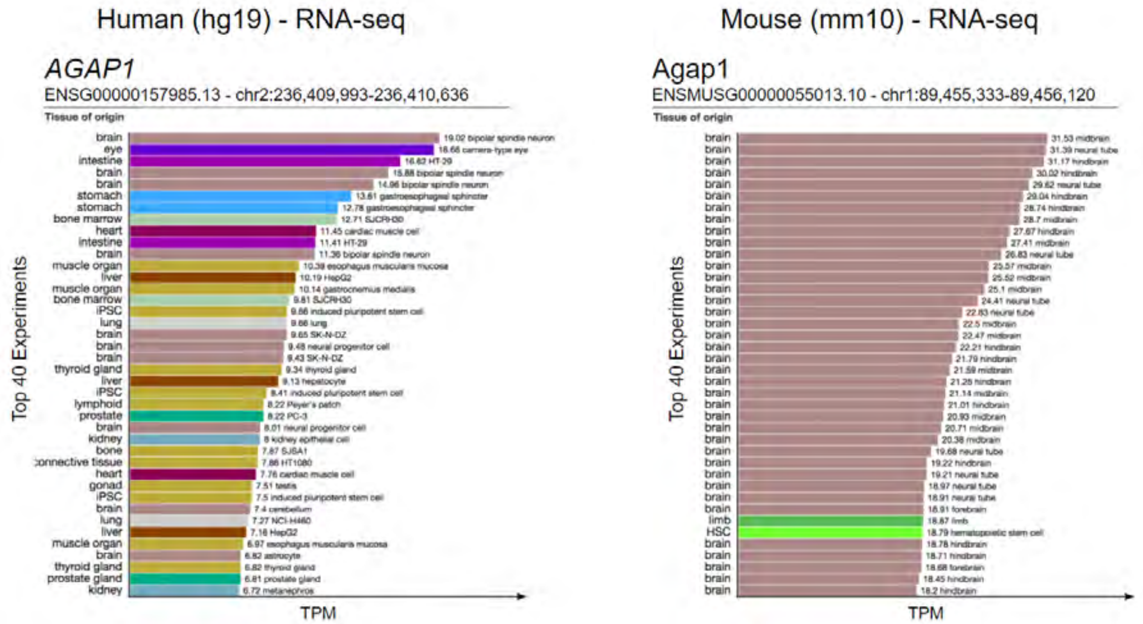


Figure II-35 | SCREEN display of AGAP1 expression levels

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 displayed in tags per million (TPM).

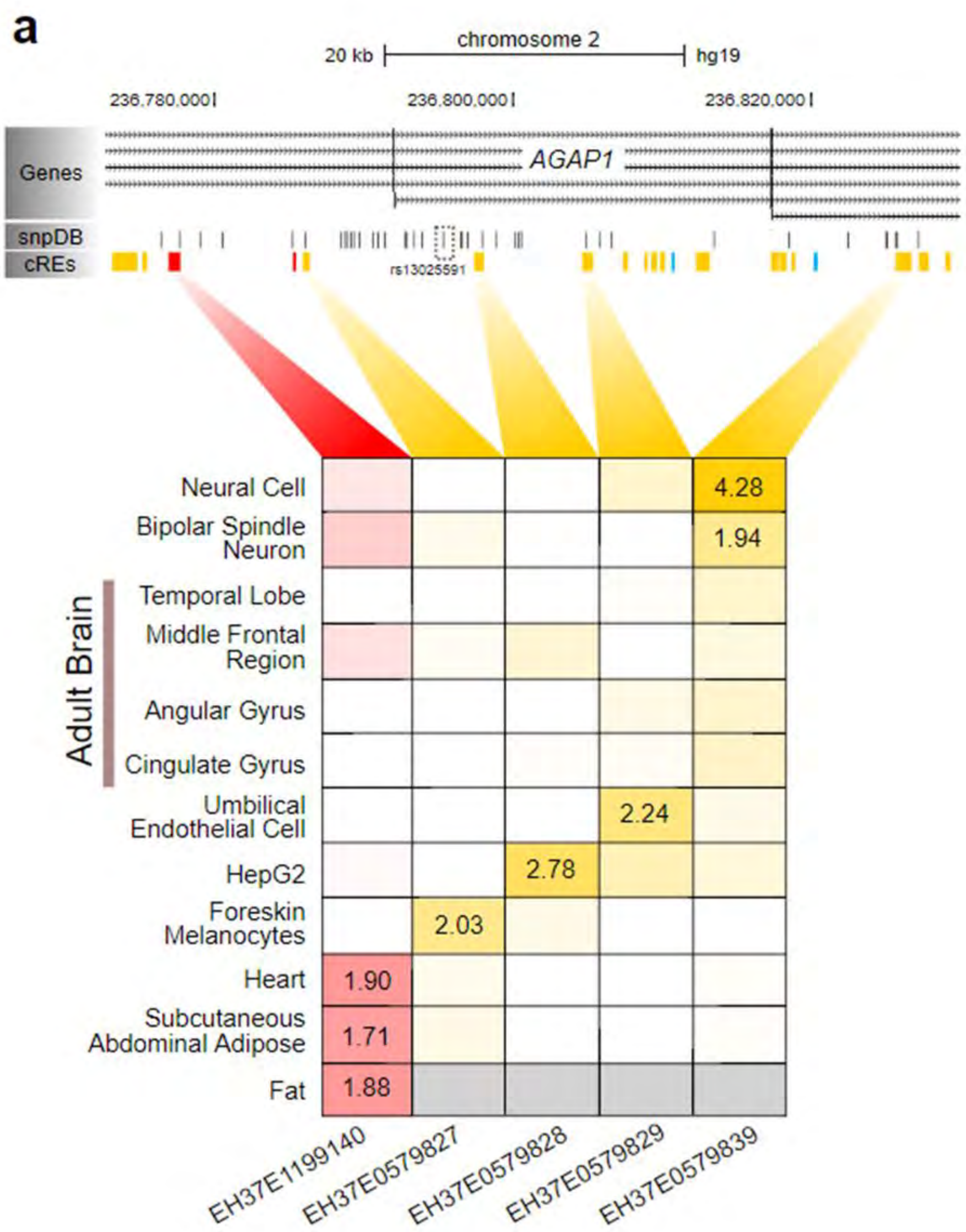


Figure II-36 | Fine mapping GWAS variants using SCREEN
a, H3K4me3 and H3K27ac Z-scores for cREs containing SNPs in LD with the schizophrenia-associated SNP rs13025591. H3K4me3 Z-scores and H3K27ac Z-scores are displayed in red and yellow, for cREs with promoter-like and enhancer-like signatures respectively.

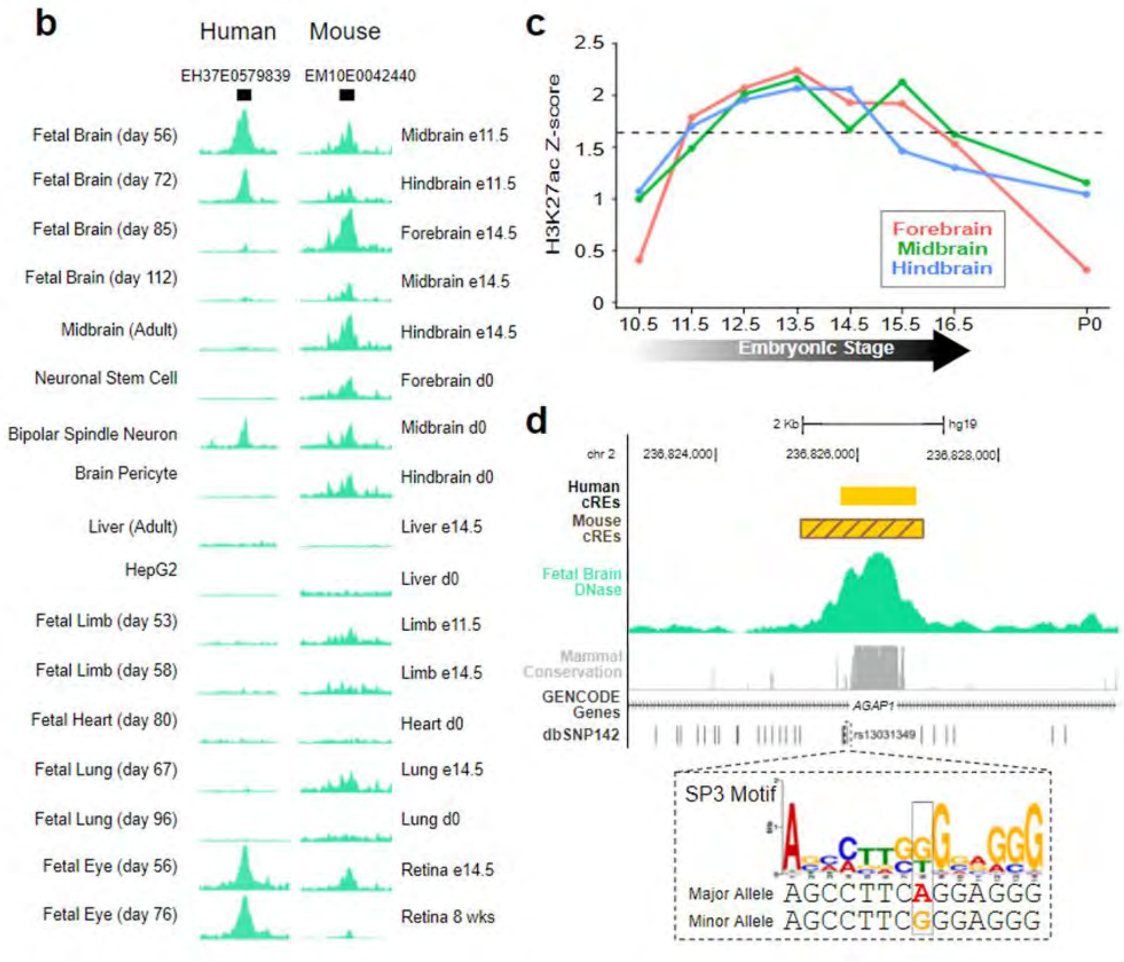


Figure II-37 | Fine mapping GWAS variants using SCREEN
b, SCREEN's Activity Profile tool allows the user to view DNase peaks at cREs across all cell types. Both the human cRE EH37E0579839 and its orthologous mouse cRE EM10E0042440 show high DNase signals in developing brain and eye tissues. **c**, H3K27ac signal at EM10E0061453 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.

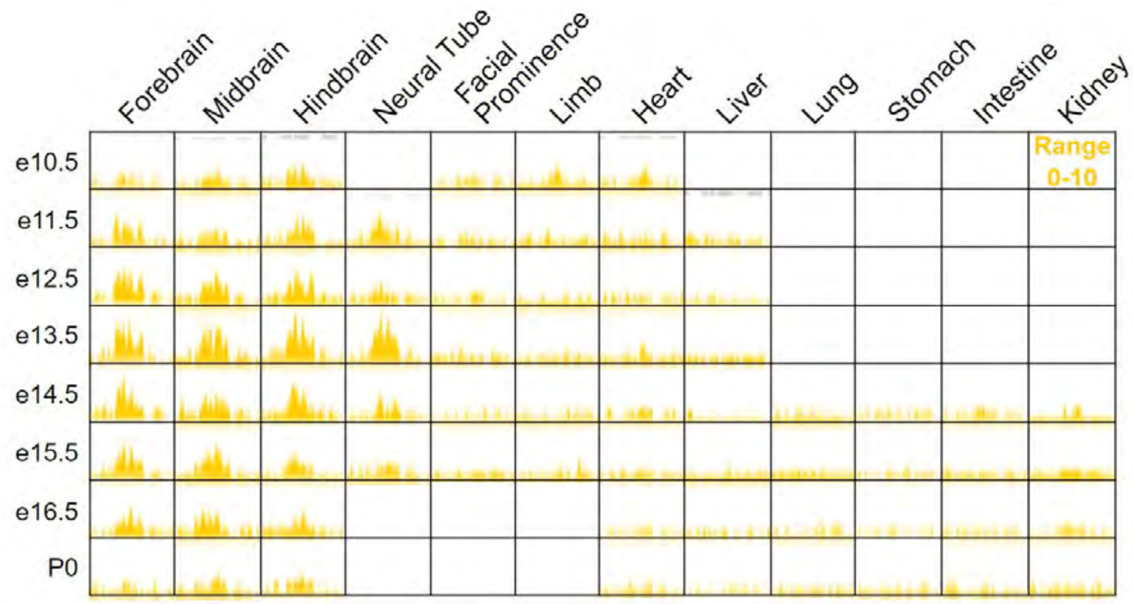


Figure II-38 | EM10E0042440 H3K27ac signal across mouse tissues
H3K27ac signal measured as fold-change between ChIP and input is displayed across 12 tissues and 8 time-points. Tissues without H3K27ac ChIP-seq data are left blank. The maximal height of signal is 10.

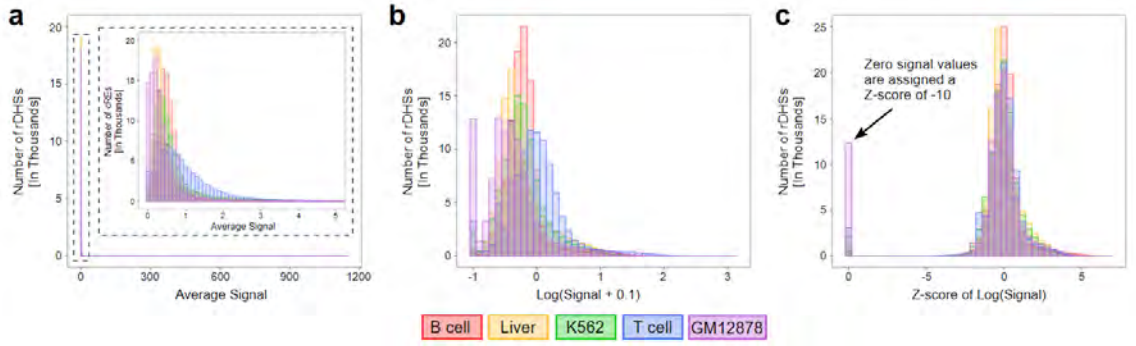


Figure II-39 | Method for normalizing epigenomics signals
a, distribution of the H3K27ac signals at rDHSs from five cell types (B cell, Liver, K562, T cell, and GM12878; shown in different colors). **b**, Distributions of the Log of the H3K27ac signals in **a**. Individually, log(signal) values of the rDHSs in each cell type roughly follow a normal distribution. **c**, Distribution of the Z-scores corresponding to the log(signal) values in **b**. Zero signal values are assigned a Z-score of -10.

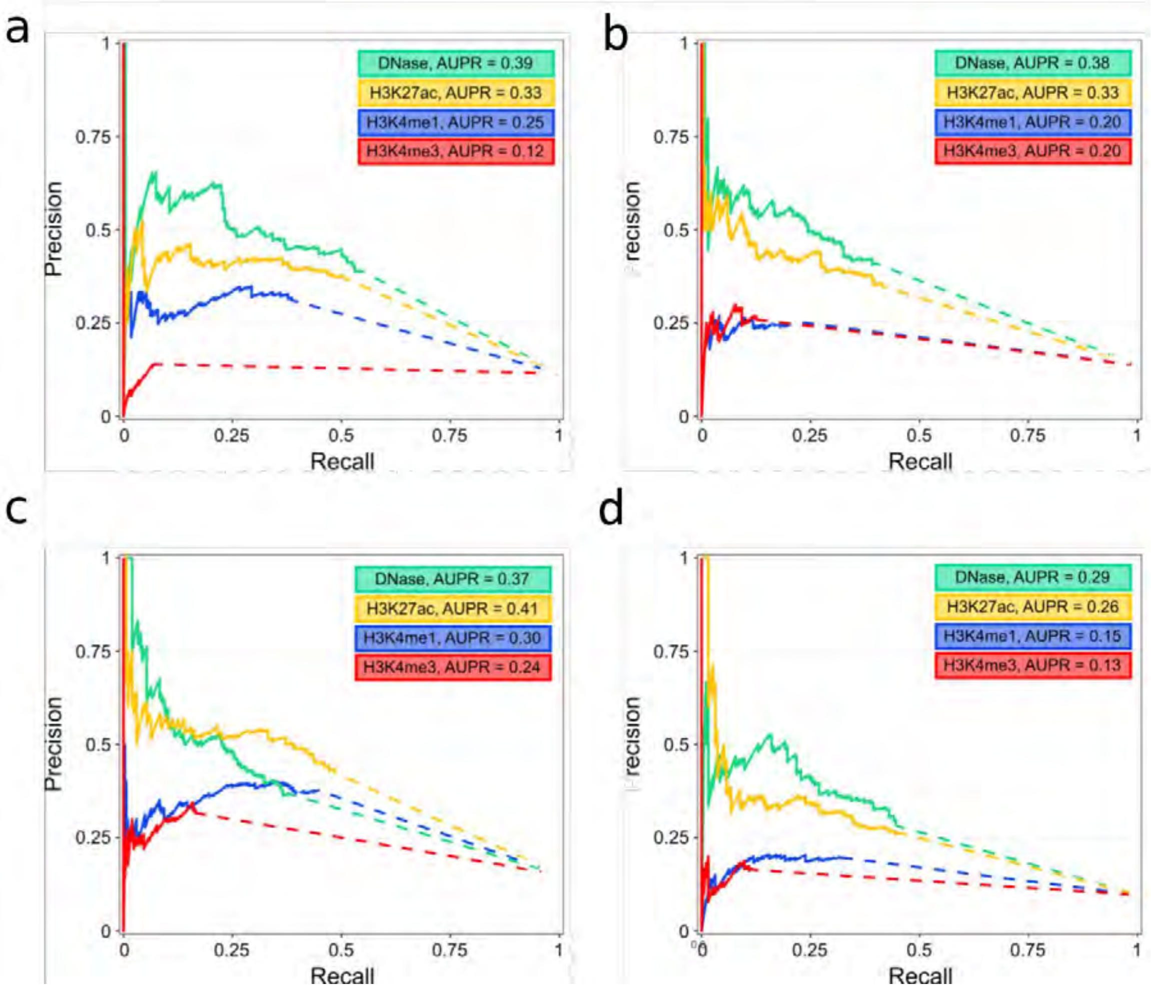


Figure II-40 | Precision-Recall (PR) curves for VISTA Enhancer prediction
PR curves for **a,** limb, **b,** hindbrain, **c,** neural tube, and **d,** midbrain enhancers at e11.5. Colours indicate peaks and signals used for anchoring and ranking the enhancer predictions. All peaks were set to 300 bp centred on their summits and the 20k top-ranked peaks were used for each tissue to ensure consistent genome coverage.

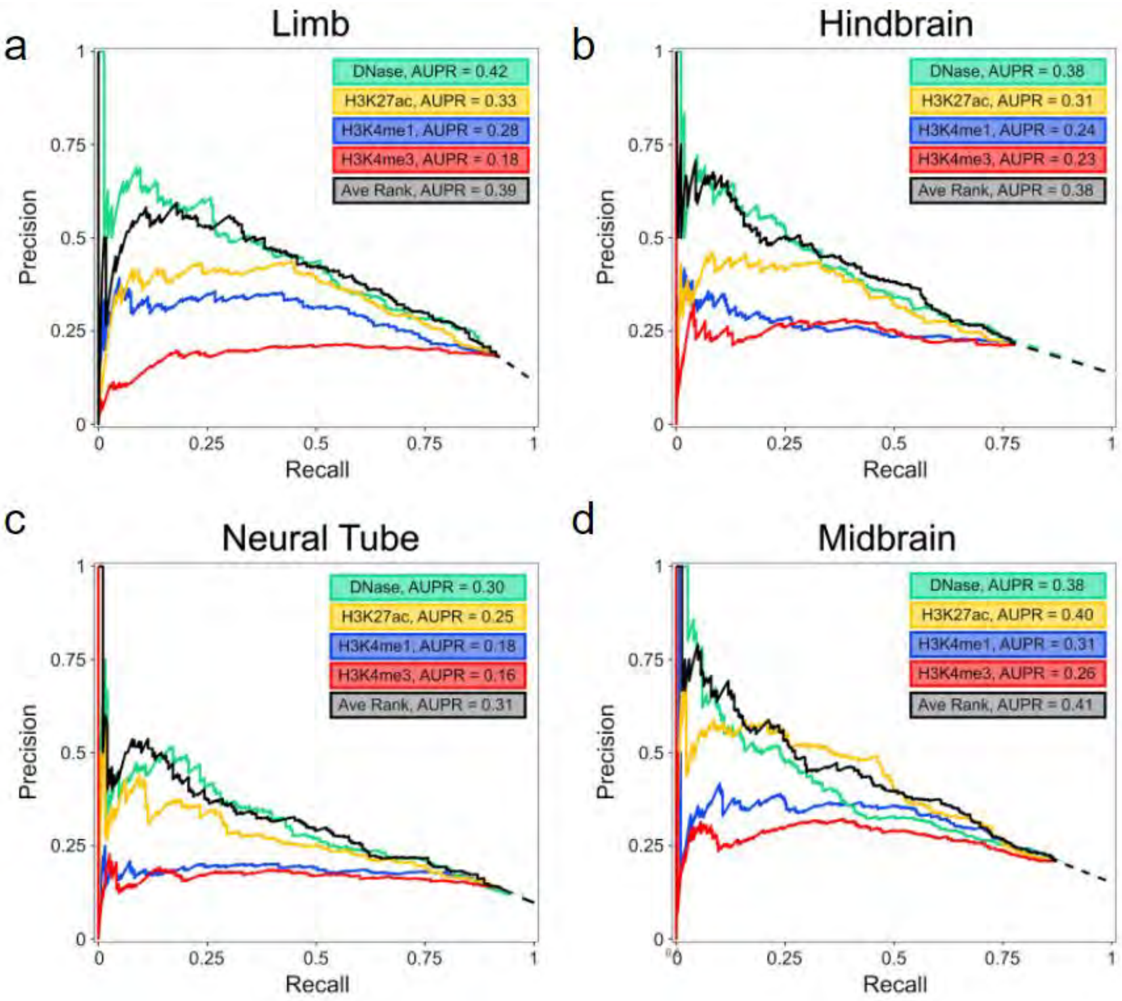


Figure II-41 | PR curves for VISTA Enhancer prediction anchored on DHSs
PR curves for **a,** limb, **b,** hindbrain, **c,** neural tube, and **d,** midbrain enhancers at e11.5. All predictions were anchored on DHSs in the respective tissue. Colours indicate signals used for ranking predictions; black indicates the average of DNase and H3K27ac signals.

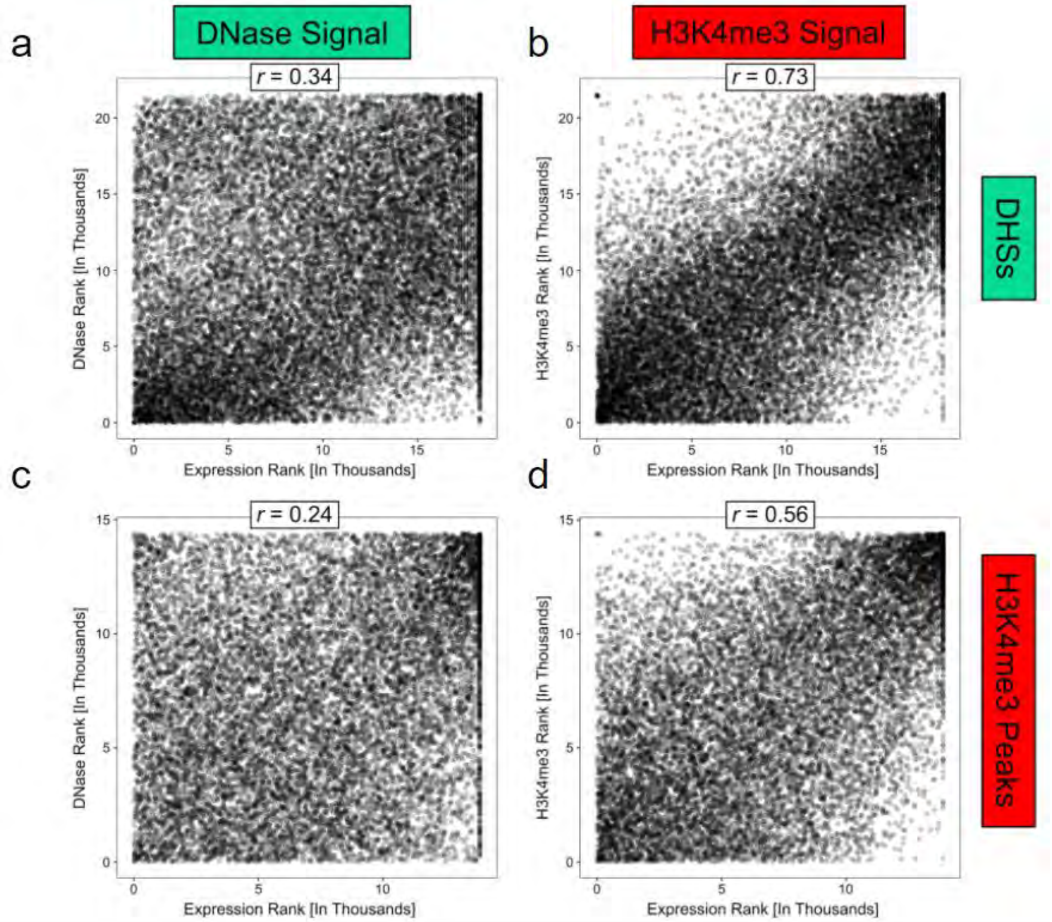


Figure II-42 | Correlation of gene expression with epigenomic signals
Scatterplots demonstrating correlation of expression with **a)** DHSs ranked according to the DNase signal (*r* = 0.34), **b)** DHSs ranked according to the H3K4me3 signal (*r* = 0.73), **c)** H3K4me3 peaks ranked according to the DNase signal (*r* = 0.24), and **d)** H3K4me3 peaks ranked according to the H3K4me3 signal (*r* = 0.56).

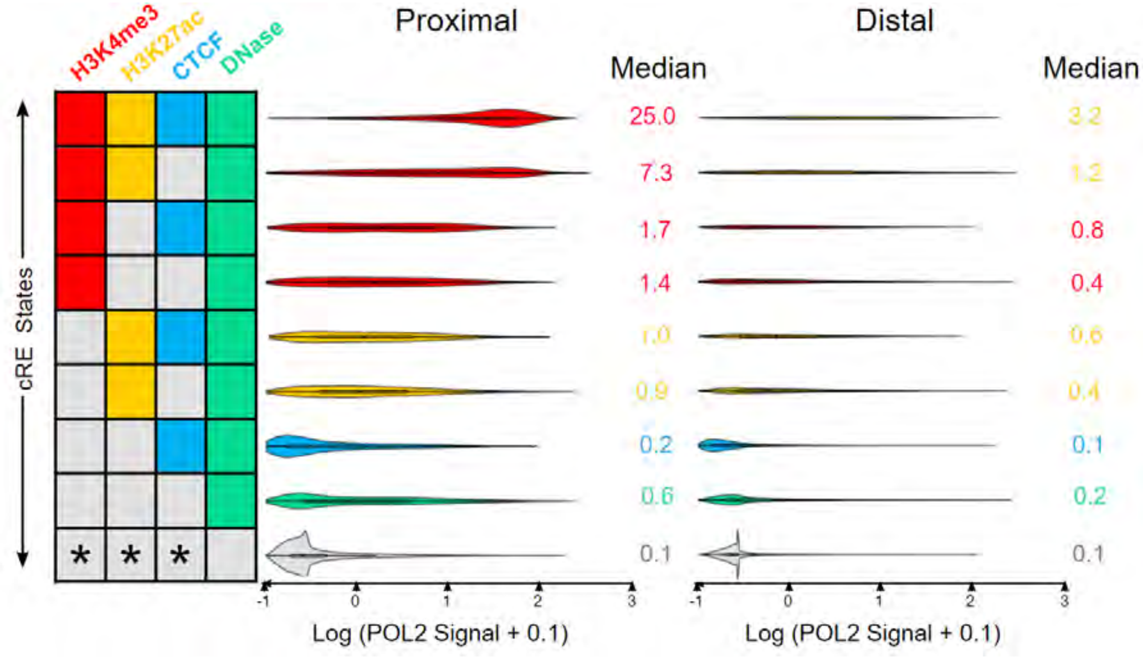


Figure II-43 | POL2 signals for GM12878 cREs
Violin plots show the average POL2 signal for cREs belonging to each of the nine cRE states. cREs proximal and distal to the nearest TSSs are displayed separately. Median values are displayed along with the number of cREs in each state.

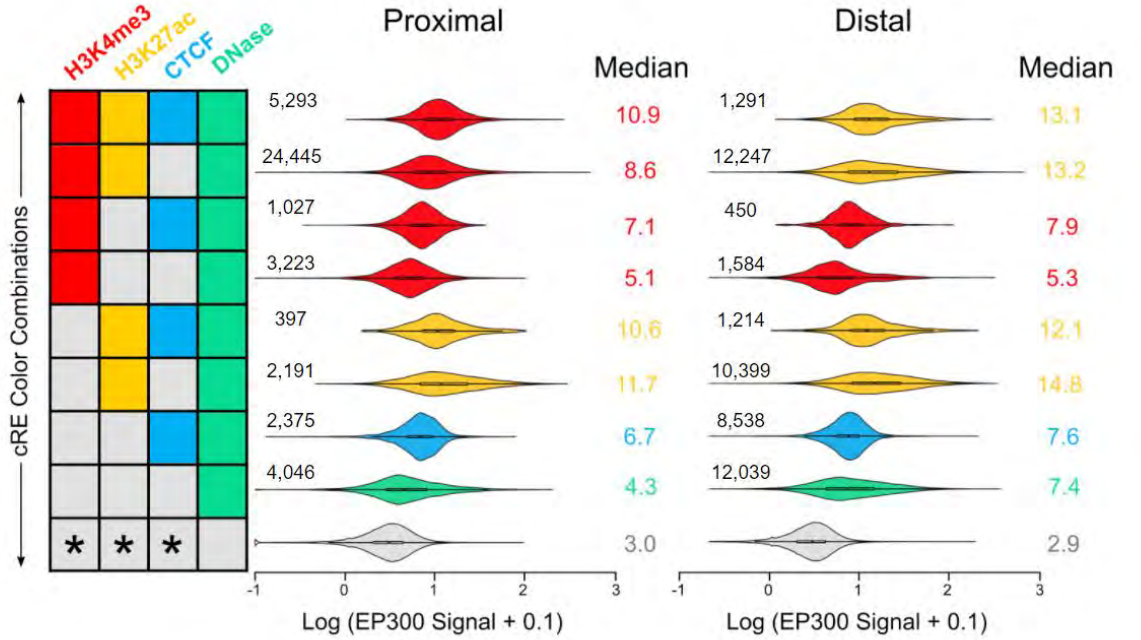


Figure II-44 | EP300 signals for GM12878 cREs

Violin plots show the average EP300 signal for cREs belonging to each of the nine cRE states. cREs proximal and distal to the nearest TSSs are displayed separately. Median values are displayed along with the number of cREs in each state

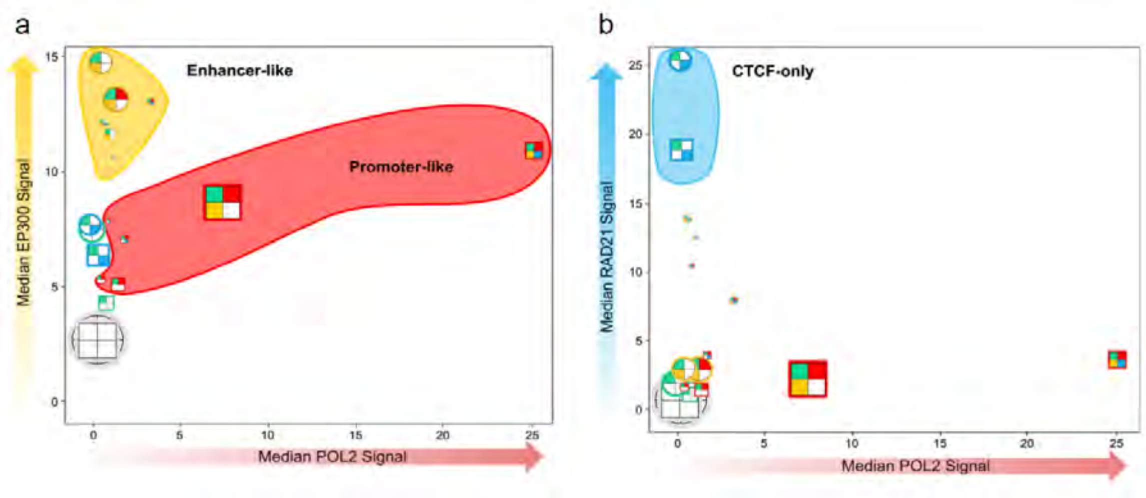


Figure II-45 | cRE states cluster into groups
Scatterplots of a, the median EP300 signal or b, the median RAD21 signal vs. the median POL2 signal for each cRE state in GM12878. The size of an icon is proportional to the number of cREs in that state except for the inactive state. Proximal cREs are represented by square icons. Distal cREs are represented by circular icons.

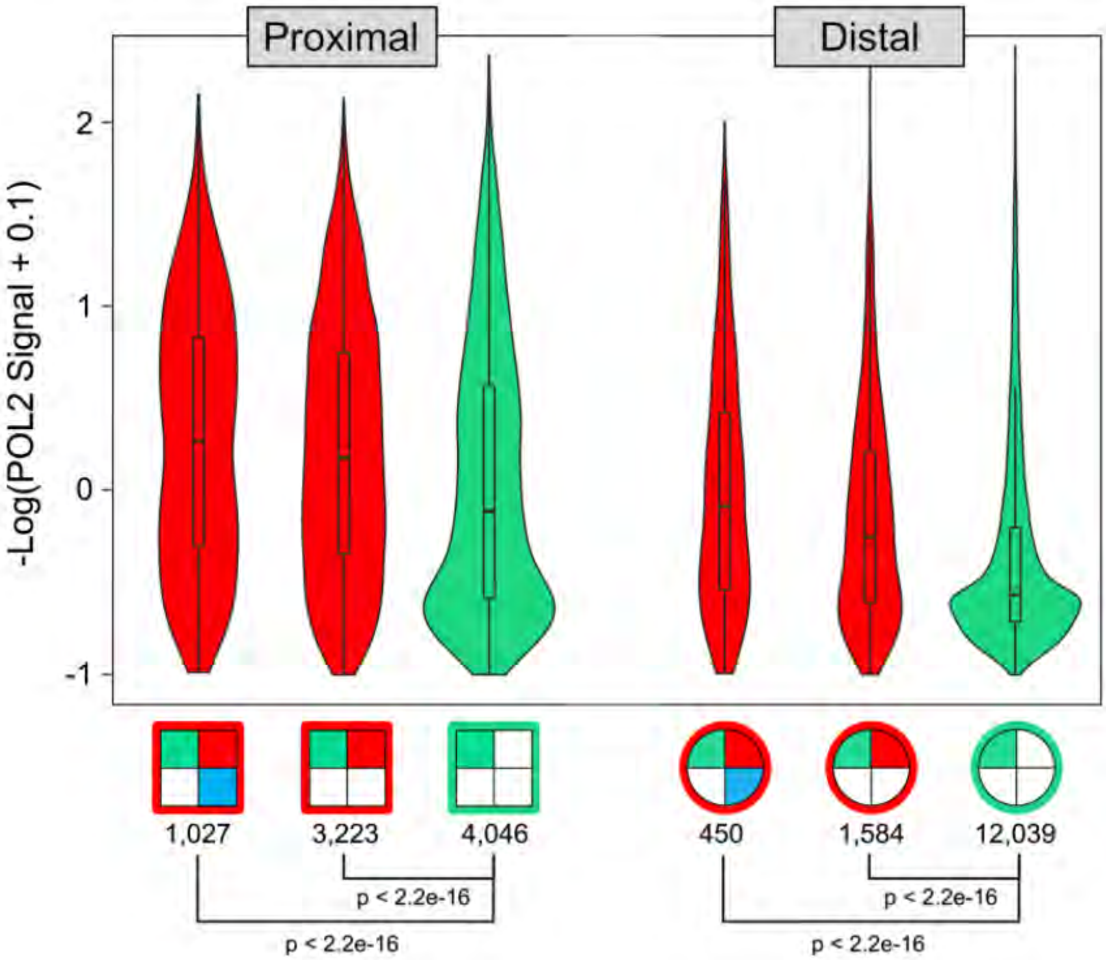


Figure II-46 | POL2 signals at cREs
Violin plots of POL2 signals for cREs with promoter-like and DNase-only signatures; these cREs belong to three states, and are stratified on the basis of whether the cREs are proximal (±2 kb) or distal to a GENCODE V19 TSS. *p*-values were calculated using a Wilcoxon test.

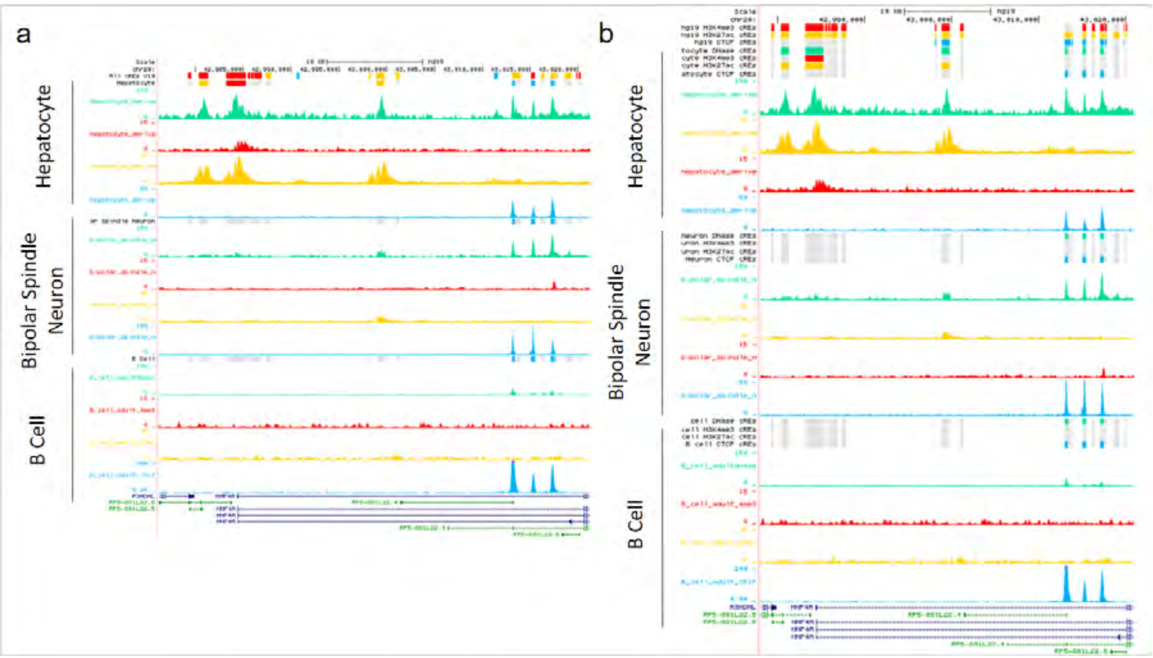


Figure II-47 | UCSC Genome Browser views of cREs around the HNF4a TSS
Browser views of hepatocyte, bipolar spindle neuron, and B cell cREs in a, five-group and b, nine-state classifications, revealing that the promoter region of HNF4a is active in hepatocytes but not in neurons or B cells.

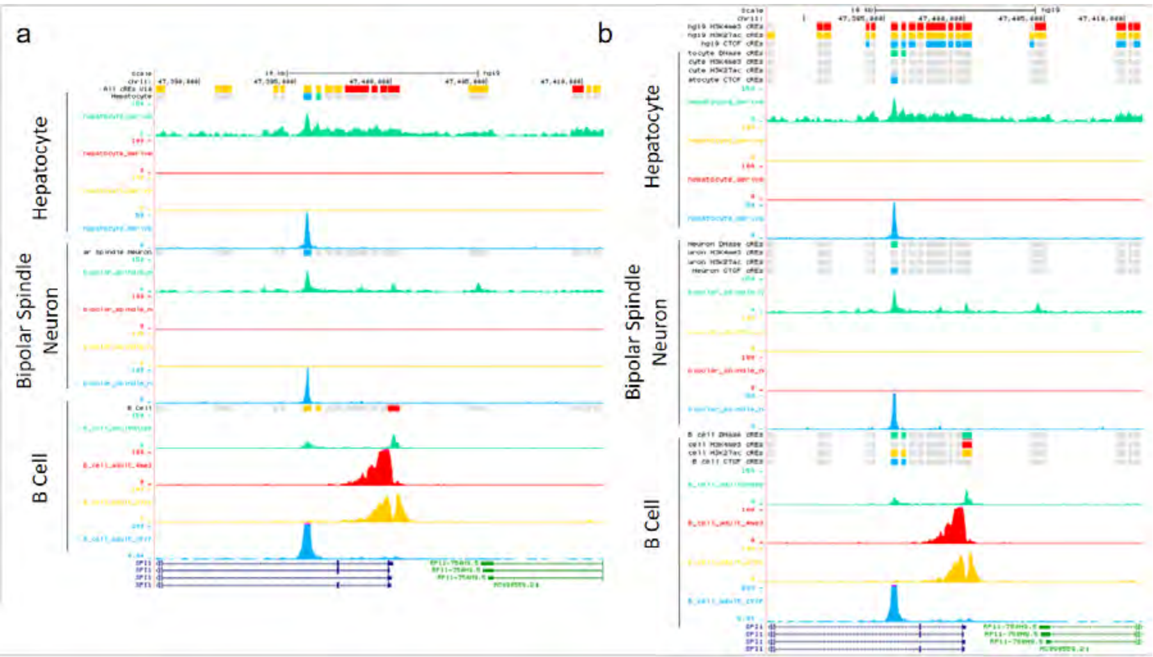


Figure II-48 | UCSC Genome Browser views of cREs around the SPI1 TSS
Browser views of hepatocyte, bipolar spindle neuron, and B cell cREs in **a,** five-group and **b,** nine-state classifications, revealing that the promoter region of *SPI1* is active in B cells but not in neurons or hepatocytes.

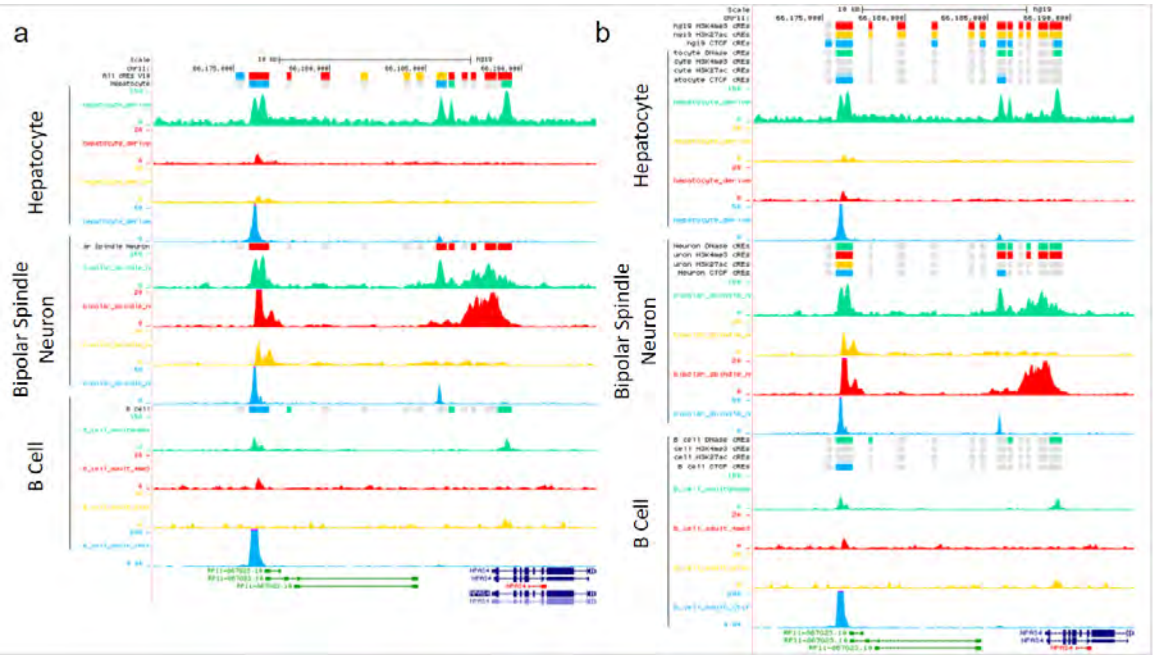


Figure II-49 | UCSC Genome Browser views of cREs around NPAS4 TSS
Browser views of hepatocyte, bipolar spindle neuron, and B cell cREs in **a,** five-group and **b,** nine-state classifications, revealing that the promoter region of *NPAS4* is active in bipolar spindle neurons but not in B cells or hepatocytes.

III. Chapter III: Factorbook V5: Peak-centric ENCODE Visualizer

III.1 Preface

This research chapter encompasses work performed in conjunction with Henry Pratt, Arjan van der Velde, Jill Moore, Eugenio Mattei, and Zhiping Weng, and is being drafted into a manuscript.

Transcription factors (TFs) are DNA-binding proteins that regulate transcription of genetic information from DNA to RNA. TFs have activating or repressive activity via many potential mechanisms, and differential TF binding dependent upon cell type specificity, phase of development, and experimental design. I found that better understanding of TF binding is critically necessary for not only elucidation of gene regulation through building transcription factor network models (Rieck and Wright 2014), but also for increased comprehension of disease mechanisms, such as TF binding changes found in cancer (Liu et al. 2017).

Factorbook originated in the Weng lab several years before I joined (Wang et al. 2013), but was years out of data with the experimental data being collected at ENCODE, and lacked many of the analysis products and data visualizations needed to better investigate TFs. I completely rewrote and re-implemented Factorbook, tripling the number of TF datasets in humans, and entirely adding the mouse component. By expanding and redesigning the heatmap comparisons component of Factorbook, I greatly added to the ability to compare binding patterns of TF in relation to other TFs and histone modifications. I am also one of the first to incorporate crowdsourcing to rate and comment on motifs. This feature will become a critical feature, as many motifs found by the common de novo motif discovery packages are invalid, and cannot, with total accuracy, be filtered by machine. Having a collection of motifs validated by experimental literature and manually curated will be invaluable for future work in deciphering how, for example, SNPs modulate TF binding.

III.2 Introduction

Factorbook is a pack-centric data visualizer for ChIP-seq and DNase-seq experiments from ENCODE. The tool, now more than 5 years old, has undergone numerous improvements and expansions since its original inception (Wang et al. 2013; Wang et al. 2012). The site now encompasses both human and mouse for more than 2,000 ChIP-seq transcription factor (TF) experiments, more than 4 times the number of experiment in Factorbook V1. The pipeline for analyzing these experiments has been drastically expanded, and produces >16TB of analysis products, taking >10 years of total compute time (>1 week on 500 CPUs) to run on a high performance compute cluster. The Factorbook user interface has been completely rewritten, incorporating the knowledge and experience gained from implementing SCREEN, our visualizer for candidate Regulatory Elements (cREs) for ENCODE (see II.5 SCREEN: A Web Engine for Searching and Visualizing cREs on page 36).

We have developed several workflows over the past few years. The Factorbook workflow for processing ChIP-seq datasets now encompasses all steps needed to take raw data, find sites of signal enrichment and sequence similarity, aggregate and analyze these data, and finally produce downstream JSON data products suitable for D3 visualizations in a user's web browser (Bostock, Ogievetsky, and Heer 2011). The pipeline produces more than 10 million separate data products for factorbook.org, and requires >60,000 compute jobs. This workflow has proven an essential tool to better elucidate epigenetic regulation of the genome, and could be utilized for other types of data by other researchers. Using our insights gleaned from developing the Factorbook workflow, we have also extensively reworked and refactored the ENCODE ChIP-seq processing pipeline to better suite our computing environment. This pipeline is now also being utilized by psychENCODE to process all ChIP-seq data. Our workflow pipelines are also open-sourced are also open-sourced on GitHub (github.com/weng-lab) using Apache, MIT, or GPL licenses. The workflows are also developed to run on Linux, FreeBSD, or Mac OS X, and only utilize open-source external tools.

III.3 Methods

To accommodate the large expansion of ChIP-seq TF and histone data from ENCODE2 and ENCODE3, we have completely redesigned our analysis pipeline. Utilizing SnoPlowPy as the metadata core, we then developed a flexible pipeline for Factorbook that would permit us to develop and test the pipeline, then run it on large multi-core machines or on a cluster. The pipeline has more than a dozen stages, with each stage spawning 100s to 1000s of jobs. For more information on our solution for managing these jobs, see Chapter V: SnoPlowPy: Advanced ENCODE Data Manipulation Tool on page 205. The Factorbook TF pipeline works as follows. First, using the SnoPlowPy metadata system, all required data files from ENCODE DCC are downloaded. These files include ChIP-seq signal files in Jim Kent's BigWig format (Kent et al. 2010) that indicate fold change signal level over control genome-wide. Narrow peak files from the ENCODE DCC processing pipeline are also downloaded. These files indicate genomic regions where fold change signal strength exceeds that of computed background, using MACS2 peak calling algorithm. In addition, these peak files have been further filtered using the Irreproducibility Discovery Rate (IDR) framework, an algorithm that compares peak ranking across files, allowing one both measure consistency across biological and/or technical replicates, as well as determine a significance threshold to keep only peaks that are reproducible (Li et al. 2011).

Next, for all ChIP-seq experiments, the pipeline performs *de novo* motif discovery using the MEME-ChIP software suite (Machanick and Bailey 2011). Discovery is performed on 100-basepair regions centered on the summits of the top 500 ChIP-seq peaks as determined by signal rank. Up to five motifs are discovered for each set of 500 regions. To accommodate for the TF peak summit being offset from the true center of the motif, as well as to potentially find multiple motifs near, if not at, the peak summit, we have expanded our pipeline to perform a sliding-window de novo MEME-ChIP motif search. This discovery is performed for regions centered from 1 to 30 basepairs to the left of the top 500 peak summits and from 1 to 30 basepairs to the right of peak summits. The resulting sets of sliding-window motif logos are displayed via drop down menus (Figure Figure III-4 on page 148).

Furthermore, we have performed motif filtering, as well as kmedoid cluster analysis, across these sliding-window motifs, as well as across all experiments for a given

TF (Wang et al. 2012; Wang et al. 2013). Motifs are assessed for quality using the FIMO tool from the MEME software suite (Bailey et al. 2009). The regions used to generate each set of five motifs constitutes a "training set" during FIMO analysis. For each training set, we generate a "testing set" consisting of 300-basepair regions centered on the summits of the top 501-1000 ChIP-seq peaks ordered by signal rank. We also generate one hundred "control sets" per training set consisting of 500 randomly-sampled GC-matched regions of the genome that do not overlap the training set regions. We then use FIMO to determine the average number of control set regions containing the five discovered motifs per control set and the number of testing set peaks containing the five motifs. The number of testing set occurrences for each motif is tested against its normal distribution for the control sets; motifs not meeting an FDR threshold of 1e-5 are considered to have failed quality assessment and are grayed out on final display.

Motifs are further assessed for quality by comparing the number of occurrences of the motifs within 300-basepair regions centered on all peaks ranked 501 and above ("testing set 2") versus 300-basepair regions directly flanking the testing set 2 regions on either side ("control set 2"). Motifs that do not occur in at least 10% of testing set 2 regions or whose ratio of testing set 2 occurrences to control set 2 occurrences is not at least 1.25 are considered to have failed quality assessment and are grayed out on final display.

The motifs from all 61 offsets are grouped using kmedoid clustering (Romer, Kayombya, and Fraenkel 2007). Distance between individual motifs is computed by first determining the optimal alignment of the two motifs and then calculating the average difference between values within the two motifs' position weight matrices, trimmed to the length of the shorter motif. Average motifs are produced for each cluster by averaging the position weight matrices of the member motifs, trimmed to the length of the shortest member motif. These average motifs are displayed as an overview of all the motifs discovered for each experiment. Averaged motifs are also indexed to allow a full-text search for motifs of interest, and clicking on an individual motif provides a list of the most similar motifs discovered in other experiments.

Motif discovery is one of the most time-consuming stages of the pipeline, with runs taking ~12 hours per experiment. During this stage of the pipeline, jobs are spawned to compute the background distributions needed for later automated filtering of motifs. A large number of small files are produced during this stage. To combat inefficiencies in storing and transferring these small files, the entire MEM-chip runs for each ChIP-seq experiment is combined into one tar file. This tar file will be utilized directly by the Factorbook website backend later. These files are also highly compressible, often being compressed to 60% of their original size. Our analysis, though, is still highly computationally expensive, and our current computational model poses a number of challenges. Moving the computational demands of these large-scale projects to the cloud would provide a unified environment which would facilitate collaboration between ENCODE groups, improve reproducibility, and allow for greater flexibility in optimizing pipeline parameters and resource usage. Cloud-storage of data would save time, improve security, and facilitate access control among groups. As such, we will soon start developing ENCLOUD, an ENCODE-related Cloud Compute Engine, that will allow us to fully move the pipeline to the Amazon and/or Google clouds. We recently were awarded a NIH Commons Credits to implement this environment.

All of the tools and pipelines we develop are open-sourced on GitHub (github.com/weng-lab) using Apache, MIT, or GPL licenses. The tools are developed to run on Linux, FreeBSD, or Mac OS X, and only utilize open-source external tools. The software languages utilized are typically Python, C++14, or BASH, and are designed in a modular fashion to promote reuse and repurposing. Input and output data formats follow those allowed by the UCSC Genome Browser[1], with PostgresQL open-source relational databases often used to store metadata. Our tools are currently designed to work on ENCODE data (~200TB of data). We are developing FAQs and user documentation for our tools using pydoc and doxygen. Most tools are designed to automatically scale when run: jobs can be tested on a single computer with 1 or many cores, and then be deployed on our cluster to run in parallel on hundreds of datasets.

III.4 Uses

Factorbook utilizes a menu-based user interface to categorize and organize the data analysis products the user can view; an example of its overall structure is presented in Figure III-1 on page 145. The first ChIP-seq TF page the user sees (Figure III-2) shows the TF grid, an alphabetized matrix of TFs (as rows) and biosamples (as columns). Cells in the grid with integers indicates the number of ChIP-seq experiments available for that TF for a given biosample. Clicking on the TF name or the cell integer brings the user to the next level of Factorbook, the TF Function page. This page shows information about

[1] https://genome.ucsc.edu/FAQ/FAQformat

the TF, including a brief overview of its molecular function, one or more 3D structures of the TF from the Protein Data Bank (Berman et al. 2000), and mined from online references, including RefSeq (Pruitt, Tatusova, and Maglott 2005) and Gene Card (Stelzer et al. 2002). Next, the user can select the Histone Profiles menu (Figure III-3); here, averaged, aggregated profiles of histone modifications are displayed on a +/- 2kb (inclusive) window around TF peak summits (when available from the MACS2 NarrowPeak pipeline) or in the center of the peak. The aggregation profiles are organized by distance to the nearest TSS, with proximal profiles have peaks within 1 kb of a TSS, and the distal profiles grouping all other peaks. Similarly, on the Nucleosome Profiles page (Figure III-7), averaged, aggregated profiles of MNase data for GM12878 and K562 are displayed.

Next, we have the Motif menu entry (Figure III-4), which displays the top 5 motifs from MEME-ChIP, their e-value scores, sequence, the number of peaks in the top 500 peaks had that motif, as well as our filtering information. Lastly, we have heatmaps for histones and TFs (Figure III-5 and Figure III-6). Here, users can compare a given TF in a specific cell type against the histone marks and other TFs in same cell type. Each column in a heatmap row indicates a ChIP-seq peak of the currently selected (“pivot”) TF. For each heatmap row, columns are sorted by descending order of ChIP-seq fold-change signal (from the BigWig file). On the page, heatmap rows themselves are sorted by decreasing Pearson correlation value. Histone heatmaps enrichment is represented in a normalized scale over a 10kb window centered on the peak summit (for NarrowPeaks) or on the center of the peak. Likewise, for TF heatmaps, binding strengths are represented in a normalized scale over a 2kb window, also centered on the peak summit (or peak center).

III.5 Discussion

Factorbook is a web-based analysis tool that integrates all public ENCODE ChIP-seq transcription factor, histone, and DNase-seq data. Factorbook is peak-centric, with peak data being summarized and displayed in several different manners. For TFs, aggregation plots display average histone mark signals in a +/-2kb regions centered on TF peak summits; peaks are separated by their distance to the nearest TSS, with proximal peaks being defined as being within 2kb of a TSS, and the distal peaks being all remaining peaks. Nucleosome profiles show the effect of TF binding on the location of nucleosomes using MNase-seq data, and peaks are similarly split into proximal and distal by TSS distance.

Factorbook has become a canonical resource for TF motif information; it has been cited by more than 127 publications[2], its motif tracks have been integrated into the UCSC Genome Browser (Rosenbloom et al. 2015), and it is even being mined for machine learning competitions (Keilwagen, Posch, and Grau 2017). Factorbook's centralization of ENCODE TF motif information has assisted finding many biological insights. Recently, Factorbook motifs assisted the search for SNPs that could disrupt TF binding in chronic lymphocytic leukemia (Law et al. 2017). Likewise, gene regulatory networks have been built incorporating RNA-seq data with Factorbook motifs to better elucidate breast cancer-related TF networks (Janky et al. 2014).

[2] https://scholar.google.com/scholar?cites=16586749045503397316&as_sdt=40000005&sciodt=0,22&hl=en

III.6 Figures

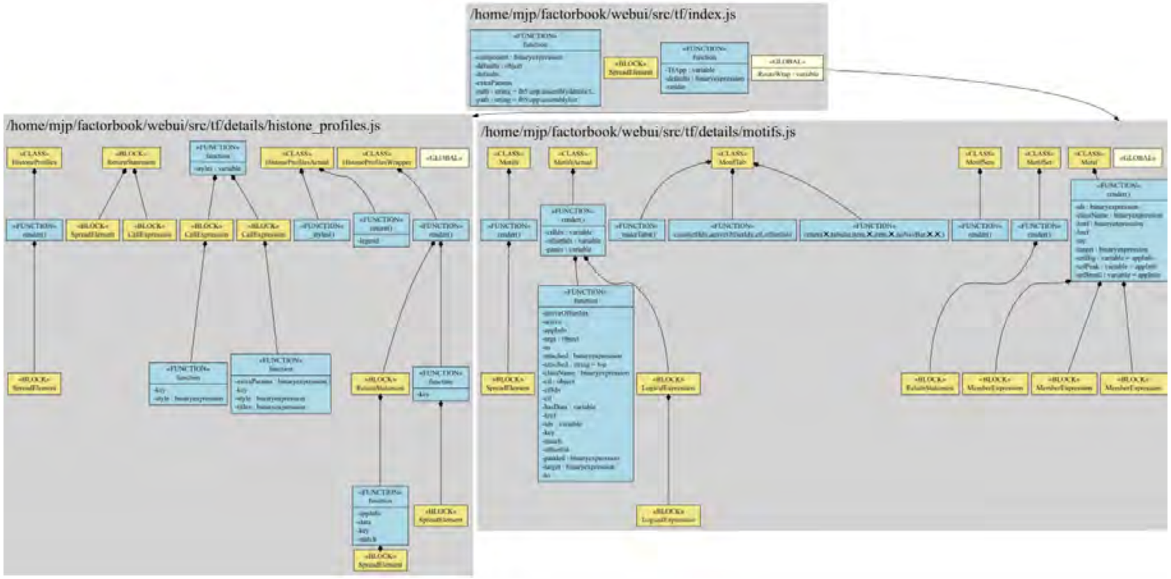
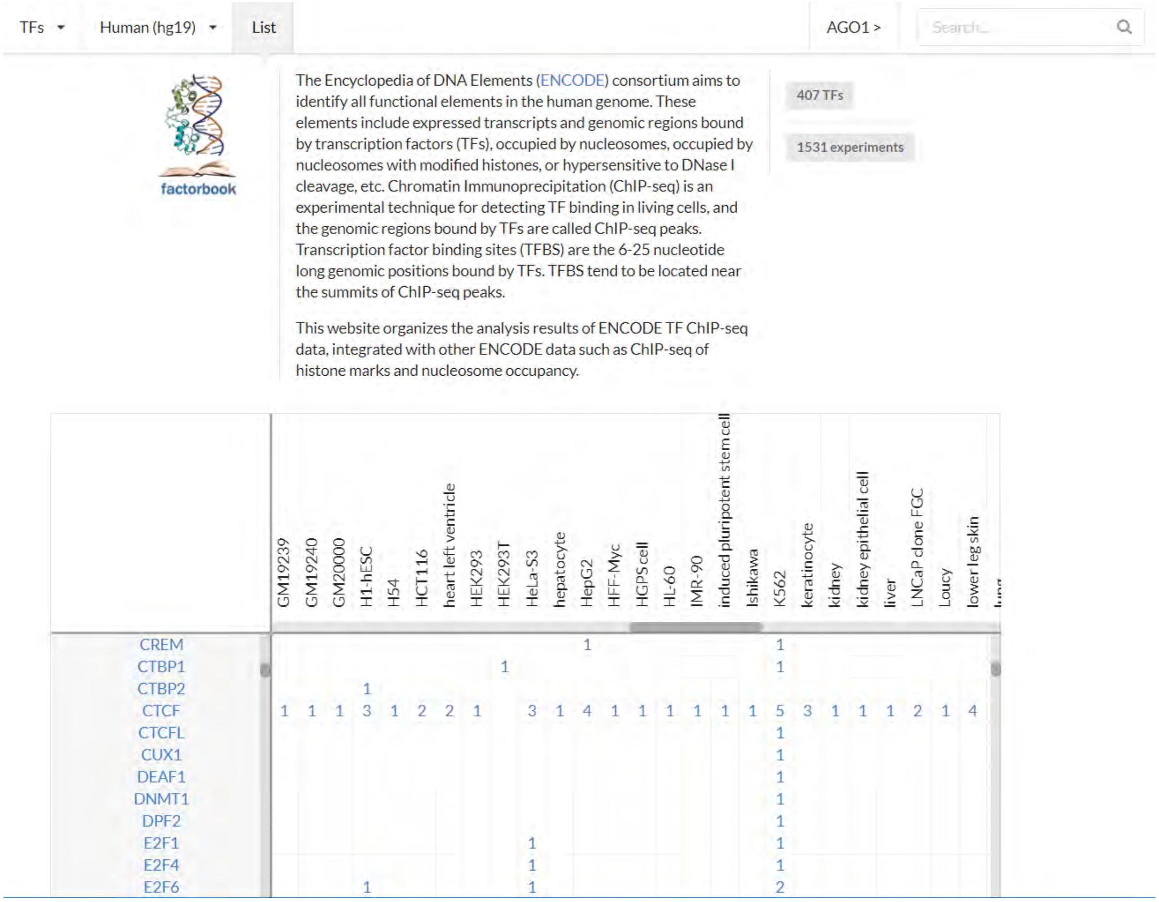

/home/mjp/factorbook/webui/src/tf/index.js
/home/mjp/factorbook/webui/src/tf/details/histone_profiles.js
/home/mjp/factorbook/webui/src/tf/details/motifs.js
«CLASS»
«BLOCK»
«FUNCTION»
«GLOBAL»
SpreadElement
CallExpression
legend

Figure III-1 | Factorbook TF UML
Overview of the TF portion of the Factorbook online database. Figure created by (Cloutier, Kpodjedo, and Boussaidi 2016)

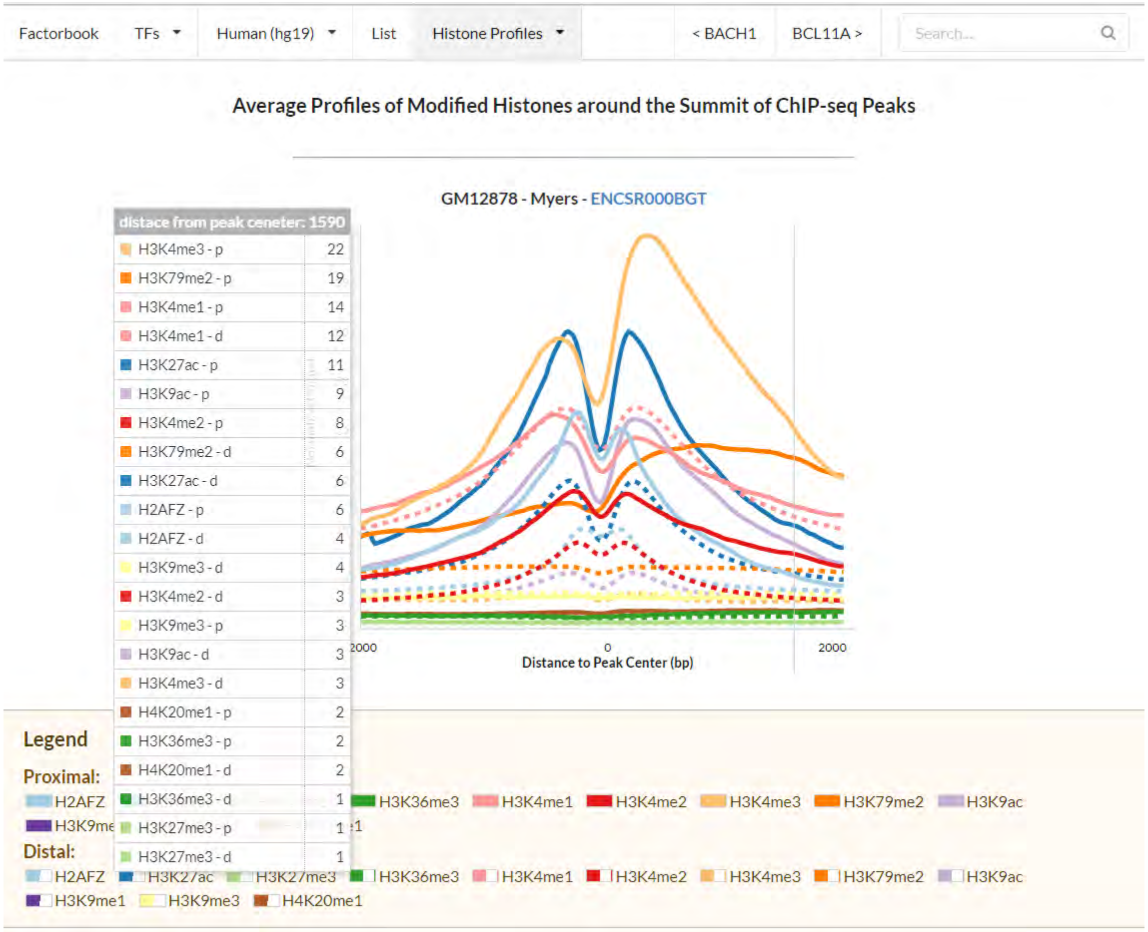


The Encyclopedia of DNA Elements (ENCODE) consortium aims to identify all functional elements in the human genome. These elements include expressed transcripts and genomic regions bound by transcription factors (TFs), occupied by nucleosomes, occupied by nucleosomes with modified histones, or hypersensitive to DNase I cleavage, etc. Chromatin Immunoprecipitation (ChIP-seq) is an experimental technique for detecting TF binding in living cells, and the genomic regions bound by TFs are called ChIP-seq peaks. Transcription factor binding sites (TFBS) are the 6-25 nucleotide long genomic positions bound by TFs. TFBS tend to be located near the summits of ChIP-seq peaks.

This website organizes the analysis results of ENCODE TF ChIP-seq data, integrated with other ENCODE data such as ChIP-seq of histone marks and nucleosome occupancy.

	GM19239	GM19240	GM20000	H1-hESC	H54	HCT116	heart left ventricle	HEK293	HEK293T	HeLa-S3	hepatocyte	HepG2	HFF-Myc	HGPS cell	HL-60	IMR-90	induced pluripotent stem cell	Ishikawa	K562	keratinocyte	kidney	kidney epithelial cell	liver	LNCaP clone FGC	Loucy	lower leg skin
CREM												1							1							
CTBP1									1										1							
CTBP2				1																						
CTCF	1	1	1	3	1	2	2	1		3	1	4	1	1	1	1	1	1	5	3	1	1	1	2	1	4
CTCFL																			1							
CUX1																			1							
DEAF1																			1							
DNMT1																			1							
DPF2																			1							
E2F1										1									1							
E2F4										1									1							
E2F6				1						1									2							

Figure III-2 | Factorbook: ChIP-seq TF Index
Home page of Factorbook, showing the sparse grid of TFs (running vertically) vs biosamples (running horizontally). The numbers in each non-zero cell indicate the number of TF experiments available for that particular TF in the given biosample.



Average Profiles of Modified Histones around the Summit of ChIP-seq Peaks

GM12878 - Myers - ENCSR000BGT

distace from peak ceneter: 1590	
H3K4me3 - p	22
H3K79me2 - p	19
H3K4me1 - p	14
H3K4me1 - d	12
H3K27ac - p	11
H3K9ac - p	9
H3K4me2 - p	8
H3K79me2 - d	6
H3K27ac - d	6
H2AFZ - p	6
H2AFZ - d	4
H3K9me3 - d	4
H3K4me2 - d	3
H3K9me3 - p	3
H3K9ac - d	3
H3K4me3 - d	3
H4K20me1 - p	2
H3K36me3 - p	2
H4K20me1 - d	2
H3K36me3 - d	1
H3K27me3 - p	1
H3K27me3 - d	1

Legend

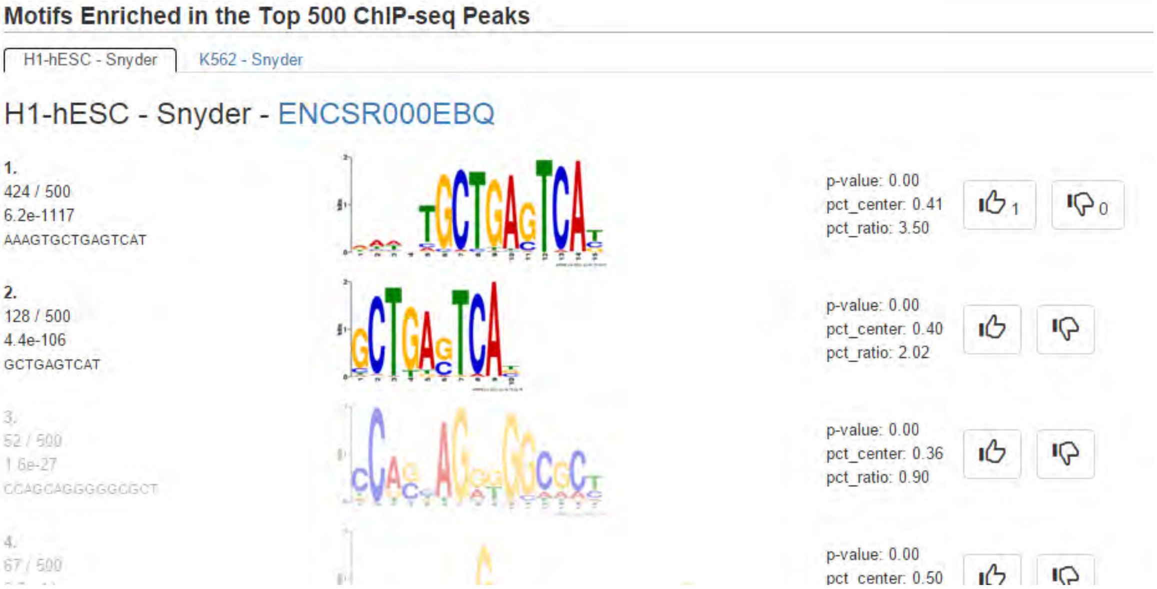
Proximal:

H2AFZ H3K36me3 H3K4me1 H3K4me2 H3K4me3 H3K79me2 H3K9ac
H3K9me :1

Distal:

H2AFZ H3K27ac H3K27me3 H3K36me3 H3K4me1 H3K4me2 H3K4me3 H3K79me2 H3K9ac
H3K9me1 H3K9me3 H4K20me1

Figure III-3 | Factorbook: Histone Profiles
Average histone modification profiles are shown for the [-2 kb, +2 kb] window around the summits of TF ChIP-seq peaks. Profiles are shown separately for peaks that are proximal ([-1 kb, +1 kb]; ending in '-p') to an annotated transcript start site (TSS) and for peaks that are distal (>1 kb; ending in '-d') to all annotated TSS. TSS-proximal profiles are arranged such that the nearest transcript proceeds towards the right. (Wang et al. 2013)



Motifs Enriched in the Top 500 ChIP-seq Peaks

H1-hESC - Snyder - ENCSR000EBQ

1.
424 / 500
6.2e-1117
AAAGTGCTGAGTCAT

p-value: 0.00
pct_center: 0.41
pct_ratio: 3.50

2.
128 / 500
4.4e-106
GCTGAGTCAT

p-value: 0.00
pct_center: 0.40
pct_ratio: 2.02

3.
52 / 500
1.6e-27
CCAGCAGGGGGCGCT

p-value: 0.00
pct_center: 0.36
pct_ratio: 0.90

4.
67 / 500

p-value: 0.00
pct_center: 0.50

Figure III-4 | Factorbook: Motifs
The top 500 TF ChIP-seq peaks were used to identify enriched motifs de novo, using the MEME-ChIP suite of tools. Up to five motifs are reported (1 to 5) if they meet the criteria defined by our filtering pipeline. (Wang et al. 2013)

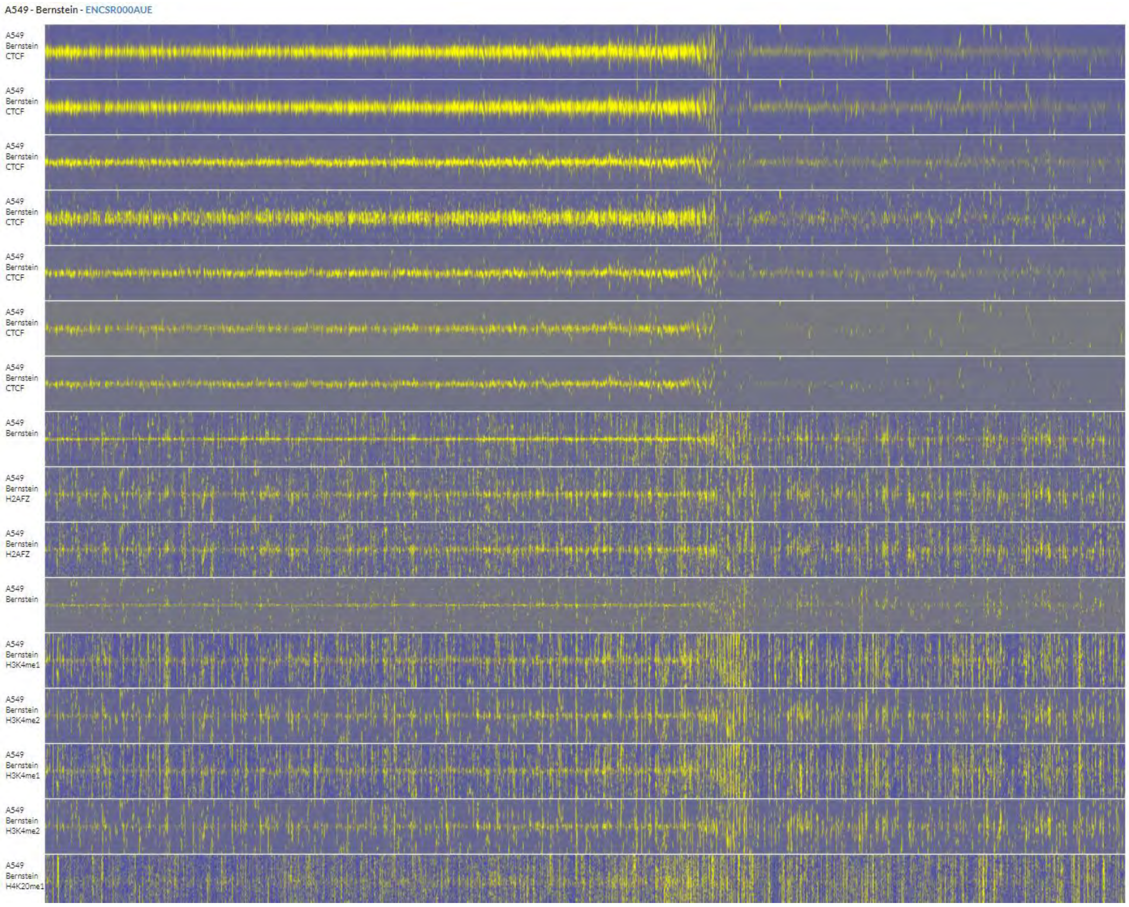


Figure III-5 | Factorbook: Histone Heatmaps
ChIP-seq peaks (left to right in the heatmap) for the current TF are sorted by descending TF ChIP-seq signal. The ChIP-seq signals of histone modifications are plotted for the genomic regions that correspond to the peaks of the current TF in the same order. (Wang et al. 2012) The Pearson correlation coefficient (r) of the histone modification ChIP-seq signal with the TF ChIP-seq signal is also computed.

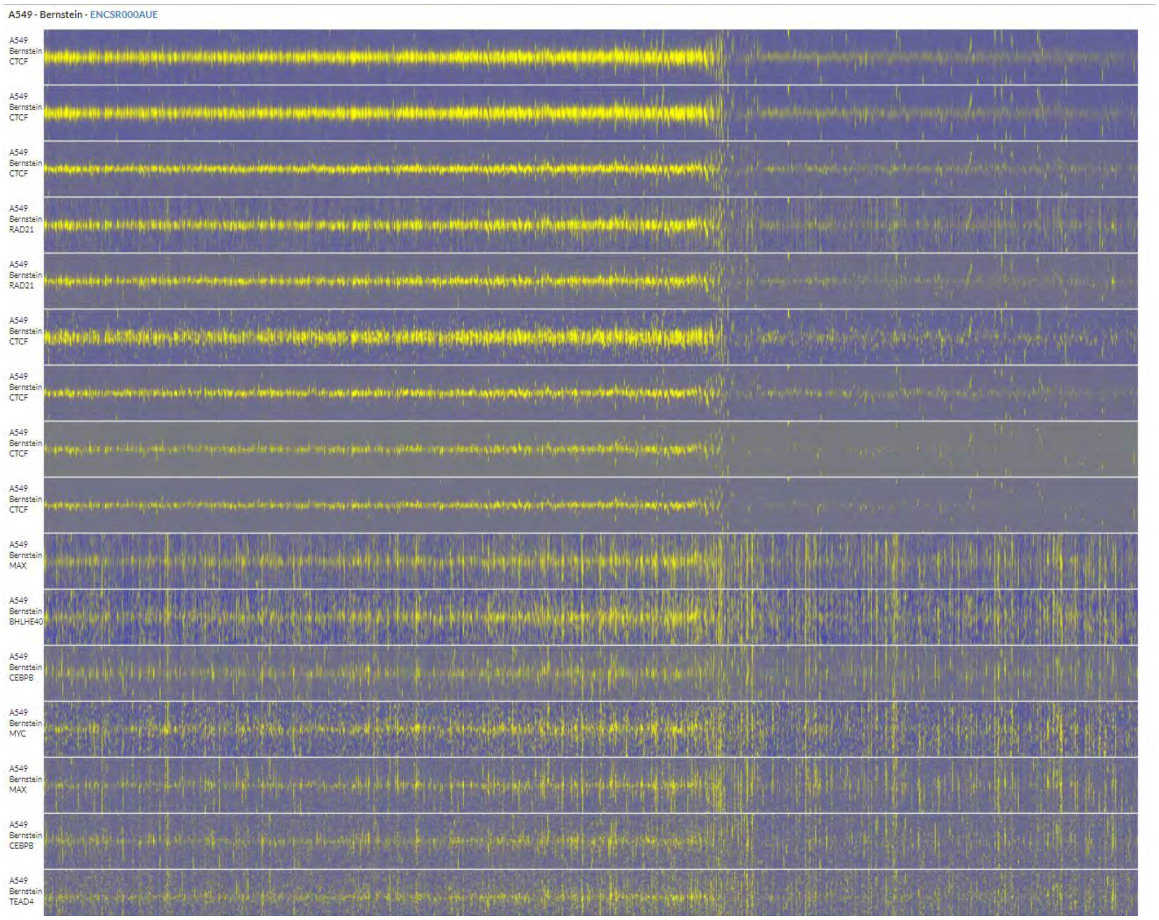


Figure III-6 | Factorbook: Transcription Factor Heatmaps
ChIP-seq peaks (left to right in the heatmap) for the current TF are sorted by descending TF ChIP-seq signal. ChIP-seq TF signals are plotted for the genomic regions that correspond to the peaks of the current TF in the same order. (Wang et al. 2012) The Pearson correlation coefficient (r) of the TF ChIP-seq signals signal is also computed.

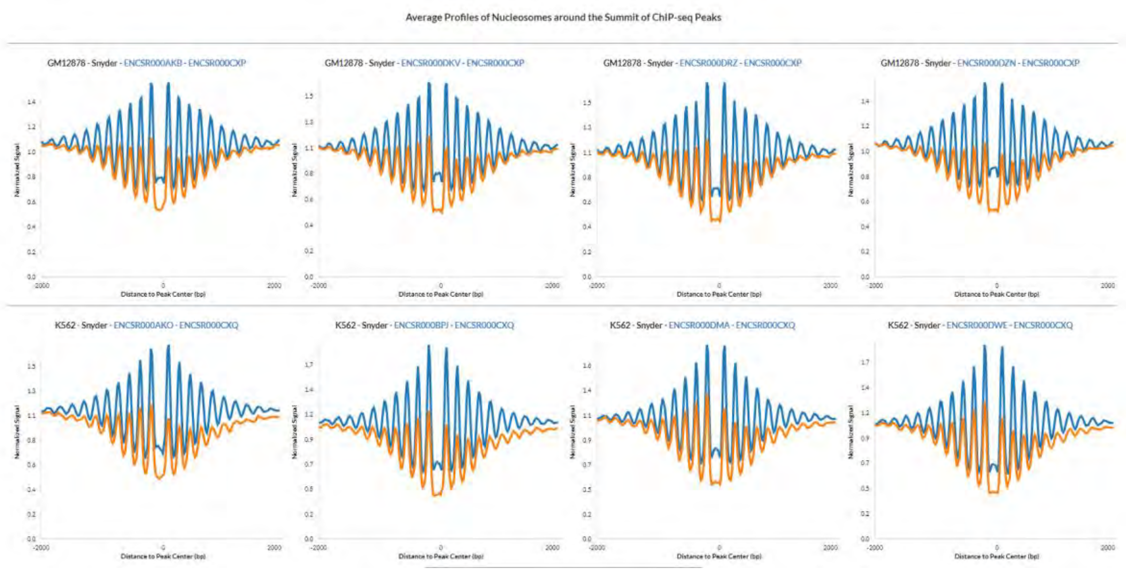


Figure III-7 | Factorbook: Nucleosome Profiles
Average nucleosome occupancy profiles in GM12878 and K562 cells are shown for the [-2 kb, +2 kb] window centered on the summits of the TF ChIP-seq peaks, separately for peaks that are proximal to an annotated transcription start site (red lines) and for peaks that are distal to all annotated transcription start sites (blue lines), as defined in the Histone section. (Wang et al. 2012)

IV. Chapter IV: Machine Learning Epigenomic Data

IV.1 Preface

This research chapter encompasses unpublished work performed primarily in conjunction with Arjan van der Velde and Zhiping Weng, along with Bill Noble, Sowmya Iyer, Jill Moore, Anurag Sethi, and Eugenio Mattei.

Locating sites with non-random, potentially functional sequence is a computational problem more than 20 years old (Bailey and Elkan 1994). The importance of locating a subclass of these non-random sequences—transcription factor binding sites (TFBSs)—is increasing, as the number, cost, and complexity of generating perhaps millions of ChIP-seq TF experiments to cover all possible cell and tissues types, developmental time points, and experiment conditions is obviously becoming intractable (Ebert and Bock 2015). I saw applying supervised machine learning techniques that incorporated the large volume of ChIP-seq data ENCODE already had collected as a potential new avenue to whole-genome prediction of TFBSs. Previous supervised and unsupervised models focused on DNA sequence features, conservation, etc., but did not incorporate expensive, hard-won experiments already collected. Additionally, I was the first to perform base-pair resolution prediction across the whole genome on across hundreds of datasets. Most previous approaches used genomic binning to reduce the computational burden in time and space. I was able to apply techniques I learned from this approach to other large-scale data problems, such as predicting enhancer activity.

IV.2 Imputation of ChIP-seq TF Data

IV.2.1 Introduction

Transcription factors (TFs) are proteins that regulate transcription of genetic information from DNA to RNA. Typically having multiple functional domains, TFs can activate or repress transcription via many mechanisms: they may complex with other TFs (Maston, Evans, and Green 2006), RNA polymerase II, chromatin remodeling complexes, and/or noncoding RNA molecules (Phillips 2008). DNA-binding TFs recognize short (6-15 base pair) sequence motifs in the genome. These motifs are highly conserved evolutionarily. Particular instances of the motif in genomic DNA that a TF binds to are called motif sites or TF binding sites. TF binding regions in living cells can be mapped genome wide using the ChIP-seq technique—chromatin immunoprecipitation using an antibody specific for the TF, followed by deep sequencing of the genomic DNA (Johnson et al. 2007). Public databases such as the Cistrome Project have indexed more than 7,000 ChIP-seq datasets (Liu et al. 2011). With thousands of different DNA-bound TFs in the human genome (Wilson et al. 2008), there are large numbers of TFs for which no ChIP-seq currently exist. This situation is compounded by differential TF binding dependent upon cell type specificity, phase of development, or experimental design. Lack of a detailed overview of TF binding and interactions with other molecules has hampered development of a more comprehensive understanding of regulation and limited development of transcription factor network models (Rieck and Wright 2014). To overcome these limitations, several different models for predicting TF binding sites have been pursued (Pique-Regi et al. 2011; Elemento and Tavazoie 2005). Input data for these models typically include sequence information, DNase-seq data, and/or histone modification data (Cuellar-Partida et al. 2012; Sherwood et al. 2014).

These models, though, don't exploit the hundreds of experimental ChIP-seq datasets already collected to improve model prediction; ChIP-seq data was only used to benchmark the predictive ability of the computational models. Recently, machine learning algorithms have been used for enhancer prediction (Erwin et al. 2014) and imputation of complete signal tracks for ChIP-seq histone marks (Ernst and Kellis 2015b) with great success. Locations of TFBSs inherently varies across cell types and experimental conditions. However, many TFs exhibit binding similarity that can be exploited in predicting TFBS in other cell types. Machine learning models can exploit these similarities in binding or open chromatin to build models predicting TF binding.

Historically, several different models for predicting TF binding sites have been pursued, with limited success (Pique-Regi et al. 2011; Elemento and Tavazoie 2005). One major shortcoming of these predictive models is the restriction of input data to sequence information, DNase-seq data, and/or histone modification data (Cuellar-Partida et al. 2012; Sherwood et al. 2014); when utilized, ChIP-seq data was only used to benchmark the predictive ability of the computational models. These models all assume TFs bind to DNA directly, and at canonical motif sites. Direct use of TFBS motif sequence with DNase-seq data is insufficient to capture the complexity of TF binding (Farnham 2009). TFs may: a) directly bind DNA as well as another TF; b) only bind another TF; c) bind DNA after stabilization by another TF; or d) bind DNA after other molecules create a more favorable chromatin state. No software package as of yet is accurate enough to displace the need for ChIP-seq experiments.

Accurate prediction of TF binding sites could eliminate the need for large numbers of ChIP-seq TF experiments on new cell types, and allow experimentalists to focus on a core set of experiments (composed, at the minimum, of DNase-seq and a core set of ChIP-seq experiments). These data, combined with cell type genomic sequencing, etc. accurately predict TFBS for all other TFs in that cell type. Additionally, these models could also be used to impute data values for missing or noisy data in previously run experiments and improve overall experimental accuracy. Accurate TF motif site data would potentially allow construction of more comprehensive models of TF binding during development or disease states (Maurano et al. 2012; Rieck and Wright 2014). For example, models could be run on cancer biopsy samples to determine what TFs are involved in producing the disease state, thus allowing development and selection of future therapeutic interventions that target the epigenome.

IV.2.2 Methods

IV.2.2.1 Epigenetic datasets

In this study, we used all publicly available ENCODE data and downloaded from the ENCODE consortium website (Sloan et al. 2016); Figure IV-1 on page 179 shows a part of the ENCODE TF matrix. We selected all cell types in which at least one DNase-seq experiment and at least one ChIP-seq TF experiment were available. We downloaded bigwig files of raw signal data. For the datasets with multiple biological replicates, we averaged the bigwig files of raw signal data across the replicates. For ChIP-seq TF datasets, we also downloaded the peak files generated using the ENCODE uniform processing pipeline (Consortium 2012). When multiple ChIP-seq TF experiments were available for a given cell type and TF, we chose one by lab preference and the most recent experiment. We also download the corresponding JSON metadata and parsed them using SnoPlowPy. We used the HG19 reference genomic sequence, downloaded from UCSC Genome Browser.

IV.2.2.2 Preprocessing

We preprocessed all signal and peak files to reduce memory and I/O intensity of later model-building steps. These data files were split by chromosome, processed, and stored in custom binary memory-mapped (citation?) files using custom C++ software. For DNase-seq, bigwig signal files were read using the UCSC genome browser utilities library, smoothed using a MJP base-pair sliding window, normalized, and thresholded based upon background signal levels. A similar process of smoothing and normalizing was applied to ChIP-seq TF data files. Additionally, per-base FIMO scores (Grant, Bailey, and Noble 2011) for all TFs were computed using position weight matrices (PWMs) from Factorbook. Regions of the human genome that were either IDR blacklisted (Li et al. 2011) or ambiguous were assigned the average signal level in a MJP window during preprocessing, and later excluded from analysis. Pearson correlation coefficients of DNase-seq data were computed per chromosome over 10 kb bins for all pairs of cell types; these values form an estimation of cell type similarity, and are used during feature selection.

IV.2.2.3 Features

Two features were formulated for each DNase-seq and ChIP-seq dataset, one feature representing the average signal over a 200-bp window centered on each genomic position, and one feature quantifying the percent of the window that contained signal above the threshold determined during pre-processing. We used one feature for FIMO scores: the maximum FIMO score in a 200 bp window centered on each genomic position. In addition, we computed the average ChIP-seq signal for each TF by averaging all ChIP-seq signal tracks for a particular TF across all available cell types (excluding the testing and training cell types). Potential motif sites from the JASPAR, TRANSFAC, and UniPROBE databases were retrieved, and motif binding prior probabilities computed for each motif using FIMO (Grant, Bailey, and Noble 2011). Mono- and di-nucleotide frequencies were also computed for the hg19 human genomic sequence.

We performed feature selection in several stages to evaluate the contribution of each feature to the overall model performance. Figure IV-2 on page 180 demonstrates the iterative approach to adding a new feature, then reevaluating the model. For each stage, training and testing were performed using the same set of features, but in two different cell types. For example, we may train a model based on the DNase-seq and E2F1 ChIP-seq datasets to predict CTCF binding in GM12878 cells, and then test this model using the DNase-seq and E2F1 ChIP-seq datasets to predict CTCF binding in HepG2 cells.

IV.2.2.4 Model building

In supervised learning, the learning algorithm used for model training is given input training data (as described above), as well as the desired output data the algorithm is to attempt to match. This approach will be performed genome wide on base-pair resolution, with a set of features computed for each base.

Logistic regression is a canonical binary classifier. Logistic regression computes the probability P(output is true | input data) for each sample in the input dataset. For an intuitive explanation of how the model works, consider an input dataset x: logistic regression will take x and assign a weight (β) to each piece of data in x such that the (hypothesis) function $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots = \beta^T x$ can be converted to a probability by the logistic (or sigmoid) function $\frac{1}{1+e^{-a}} = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\cdots)}}$. For example, for our logistic regression model to predict TF binding probability at a base b, the hypothesis function would equal (bias constant) + (average DNase-seq data at base *b* over a given window) + (% of window with DNase-seq data > threshold) + (average TF_1 data at base *b* over a given window) + (% of window with TF_1 > threshold) + [remaining features]. More technically, logistic regression will ultimately find a decision boundary that separates the input data samples into two categories. Solving the weights in the hypothesis function can be computed by transforming the problem into a numerical minimization problem that can be solved using methods such as Newton's method. Logistic regression models can be rapidly trained on large datasets using software such as LIBLINEAR, and provide a good performance baseline.

In addition to logistic regression, we also trained Support Vector Machines (SVMs) to locate TFBSs. Like logistic regression, SVMs work by finding a decision boundary that separates the input data into two classification categories. Unlike logistic regression, though, SVM will find a decision boundary that maximizes the minimum distance from the samples; finding this “large margin” occurs as a consequence of the optimization function used in SVM (Marsland 2009). Another consequence of this optimization function is the ability to transform the input data to improve the margin between categories; these kernel functions allow transformation of input data into higher-dimensional space in a computationally cost-effective (i.e. linear) manner (Marsland 2009). Many different kernel functions exist, giving SVMs a rich parameter space; this may be particularly useful for finding relationships between data experiments when learning TFBSs. The core optimization problem to be numerically solved in SVMs can be solved in polynomial time using quadratic programming, making SVM efficient. SVM software packages capable of handling large datasets include LIBLINEAR and SVMlin (Sindhwani and Keerthi 2006). We ultimately found little performance difference between logistic regression and SVM, and moved on to gradient boosting.

We also used boosting methods to perform supervised learning. In boosting, a collection of simple predictive classifiers are combined by a weighting or voting system; this allows the predictive power of the group of learning methods to exceed the predictive power of any individual predictors. Boosting happens sequentially: early classifiers make simple predictions; these predictions are then analyzed, and training of later classifiers focuses on the errors made earlier. Errors made by early predictors indicate data samples that are difficult to predict. Ultimately, very simple predictive methods get convolved into a complex, weighted classifier. For example, AdaBoost, an early and now canonical boosting algorithm, uses an adaptive approach to boosting. Initially, all data samples are weighed equally. After each stage of prediction, before weight renormalization, samples that were erroneously classified get additional weight, while the weights for samples correctly identified are left unchanged (Marsland 2009). Data samples that are incorrectly predicted receive more weight in an attempt to minimize error. Gradient boosting applies the same technique to regression: a basic regression is first made, the error residual is calculated, and the next regression step attempts to reduce the fitting error. Boosting may also be quite helpful for the experimental datasets being used, given that each dataset may only contribute a small amount of information towards predicting TFBSs. For boosting, we evaluated several packages, starting first with rt-rank, an open-source package with several different boosting algorithms designed for large datasets (Mohan, Chen, and Weinberger 2011), and ultimately used xGBoost (Chen and Guestrin 2016).

The volume of data processed did require extensive time and compute resources. The time complexity for logistic regression on a single compute core is $O(sf^2 + f^3)$, where s represents the number of training samples and f the number of features; SVM has a runtime of $O(s^2f)$ (Chu et al. 2006). There are several possible strategies to reduce runtime. The first and simplest is to train and test the models on a single chromosome instead of working genome wide; this immediately reduces the sample number from $3x10^9$ samples to ~$1.5x10^8$, with corresponding reduction in time and memory usage. Further reduction in time and memory, though, was required, and achieved by sampling on just a subset of the training data; in-house experiments indicated training on 10 million samples for a small (<50) number of features does not adversely affect the model accuracy; additional bootstrapping techniques could also be used. Additionally, training a model on a single chromosome, and then using that model to test on the rest of the genome, also reduced the time cost of training a separate model for every chromosome, without dramatically changing aucPR. Parallel programming techniques were also used to speed up parts of the supervised learning pipeline; data preparation (such as windowing and normalization) benefited from straightforward applications of programming libraries such as OpenMP (Eichenberger et al. 2014).

We built L1-regularized logistic regression models as implemented by the LIBLINEAR software package (Fan, Chang et al. 2008). For each chromosome, we randomly selected 15% of the genomic positions as training data, with the ratio of bound to unbound TF sites preserved. We then applied the trained model predict a binding probability for every genomic position per target TF in a different cell type. We did not use more than 15% genomic positions in the training cell type because the complexity of our models did not require more training data. We also utilized xGBoost to implement gradient boosting models.

IV.2.2.5 Performance evaluation

All models trained output a probability of a transcription factor binding at each base position in the genome. Traditionally, a confusion matrix could be created with the total number of true positive, true negative, false positive and false negative predicted given a probability cutoff threshold, and thus leading to calculation of a sensitivity, specificity, positive predictive value, etc. for each model. Metrics such as receiver operator characteristic (ROC) and precision-recall (PR) curves, however, allow better characterization and visualization of model performance, as the cutoff threshold is moved through the range [0,1]. ROC curves plot false positive rates vs true positive rates (aka sensitivity), demonstrating how the number of correctly predicted examples varies with the number of incorrectly predicted negative examples. The area under the ROC curves (after being normalized) signifies the probability that the classifier will classify a randomly chosen positive sample higher than a randomly chosen negative sample. In PR curves, the true positive rate (sensitivity) is plotted vs the positive predictive value (aka precision).

Mathematically, so long as models are predicting on the same datasets, ROC and PR curves are inter-convertible, and curves that dominate in ROC space will also dominate in PR space (Davis and Goadrich 2006). What distinguishes use of one curve over another (for our purposes) is how each curve visualizes highly asymmetric datasets. Given that <1% of a chromosome will have binding sites for a particular transcription factor, the set of negative examples for each model constitutes the majority of what must be predicted. In asymmetric data cases like these, PR curves offer a better way of representing actual performance of the models (Davis and Goadrich 2006).

While building models on a per-base resolution permits maximum resolution, ultimately the output binding probability vector for a given model is scaled and run through a peak caller such as MACS2 (Zhang et al. 2008). This permits direct visual comparison between the predicted data and actual experimental data, as shown in Figure IV-5 on page 183. Peak overlaps between predicted and actual ChIP-seq data can then be computed. A smoothing process was applied to the prediction vector.

Ultimately, for each TF in each cell type, we custom built a model by selecting all available features based on ENCODE data. Other available TFs in the cell type were used, given the potential interaction of other TFs in the cell type with the desired TF. Also, the same TF from other cell types was used, since there is some correlation between the same TF in different cell types.

We evaluated the performance of the model by comparing the binding probabilities computed by the model with the peaks called from the ChIP-seq data of the TF in the corresponding cell type. We initially plotted precision-recall (PR) curves at nucleotide resolution using ROCR (Sing et al. 2005), but this method was extremely slow. Given that a chromosome may have up to 250 million bases, a naïve algorithm to calculate the PR curve will also output millions of points to plot. We implemented a custom algorithm to generate PR curves that downsamples the output curves to 10,000 to 20,000 points for reasonable plotting times.

We built a peak caller in C++, and called peaks using the predicted binding probabilities; we then compared these predicted peaks with the peaks called from the corresponding ChIP-seq data. An overlap of at least one nucleotide was counted as a correctly predicted peak. We also plotted PR curves at peak resolution. These curves were then monotonized to compute the sensitivity (also known as recall) at q-value cutoffs of 0.5 and 0.25. We compared the performance of our models with the performance of single features DNase-seq, FIMO scores, or average ChIP-seq signal in other cell types.

IV.2.3 Results

IV.2.3.1 Initial Model

To demonstrate the potential of improved TFBS prediction, we implemented a proof-of-concept model using logistic regression as implemented by the LIBLINEAR software package (Fan et al. 2008). We utilized DNase-seq data and 19 transcription factor datasets from the ENCODE and Roadmap Epigenomics projects; the TFs were chosen to avoid any with known functional interactions or binding partnerships. We built a model using LIBLINEAR on the GM12878 cell line using DNase-seq and 18 TF datasets as the input data, with a 100 base-pair smoothing window; the model was trained on 5% of the data on chromosome 7 to predict the 19th TF (ATF3 for this example). We then used the trained model to predict ATF3 using data for the HepG2 cell line. For a baseline of comparison, I ranked DNase-seq data to directly predict TF binding—a predictive strategy that has been noted to be competitive with more complex models (Cuellar-Partida et al. 2012). As demonstrated by Figure IV-3 on page 181, the addition of 18 TFs substantially improved model sensitivity.

IV.2.3.2 Full Models

We can predict where transcription factors bind on the basis of the DNA sequence, DNase-seq and ChIP-seq for this TF from other cell types, and DNase or other TFs in this cell type. In general, our models perform better than DNase-seq strawman, FIMO, or average of other ChIP-seq experiments, and compete with more complex methods such as ChromImpute. Our methods nearly always have increased performance over using an individual ChIP-seq assay from another cell type to predict binding for a given cell type.

Straightforward methods of imputing ChIP-seq datasets have poor performance at both base-pair and peak-centric levels. Simply using DNase-seq (an indicator of open chromatin) gives poor performance (Figure IV-3 on page 181). For example, for CTCF in HepG2, DNase-seq performs poorly (Figure IV-4 on page 182). Utilizing the average of other TFs in the same cell type also performs poorly, as does the average of the TF across all other cell types, and FIMO scores (surrogates for sequence). Machine learning techniques can be utilized to better capture information between assays and other features. For instance, using logistic regression, we build increasingly complex models. We find that many ChIP-seq TF datasets can be captured well using our methods. TFs such as CTCF, while having high correlation across cell types, can be imputed with strong performance.

As another example, a model was built using LIBLINEAR on the GM12878 cell line using DNase-seq and 420 TF datasets as the input data, with a 200 base-pair smoothing window. The model was trained on 20% of the data on chromosome 7, and used to predict a binding probability for every base for a given target TF. We then used the trained model to predict the same target TF in the HepG2 cell line. For a baseline of comparison, we ranked DNase-seq data to directly predict TF binding—a predictive strategy that has been shown to be competitive with more complex models (Cuellar-Partida et al. 2012). As Figure IV-6 on page 184 demonstrates, most ChIP-seq TF models benefitted from features added beyond simply DNase-seq. This trend generally continues when examining predicted TF experiments grouped by TF, as shown in Figure IV-7 on page 185; careful examination of predictions in just HepG2 in Figure IV-8 on page 186 does demonstrate feature engineering is often required, as merely adding 420 other ChIP-seq datasets does not guarantee improved performance. Our performance metrics did improve when switching from base-pair resolution imputation to peak-level (via our internal peak caller), as generally demonstrated in Figure IV-9 on page 187.

ChomImpute, a regression tree-based imputation software package utilized for imputing DNase, DNA methylation, and ChIP-seq histone marks for ROADMAP Epigenetics datasets, was also run. Both our logistic regression models and gradient boosting models failed to consistently outperform ChomImpute (see Figure IV-10 and Figure IV-11) when run on the exact same feature matrices.

IV.3 DREAM *in vivo* Transcription Factor Binding Site Prediction Challenge

IV.3.1 Introduction

The DREAM (Dialogue for Reverse Engineering Assessments and Methods) Challenges support open scientific, computational-centric research on biological data (Stolovitzky, Monroe, and Califano 2007). Anshul Kundaje from ENCODE DAC recently helped lead an effort for in-vivo transcription factor binding prediction. This challenge used a restricted set of data, over a binned set of genomic coordinates (Figure IV-12 on page 190) to help advance machine learning approaches to TF prediction. We applied lessons learned from out ChIP-seq TF imputation (which used base-pair resolution, over a large set of data) to work on the challenge.

We took a multiscale approach to generating features using DNase-seq, RNA-seq and sequence totaling to about 65 features per 200bp window. Since the 200bp windows are overlapping and have step size of 50bp, we effectively generated features for the middle 50bp each window. Based on correlation of DNase-seq datasets, for each target cell type we chose the most appropriate training cell type. The same feature sets were used for all TFs (motif-based features were omitted for TFs without a motif present in jasper), and the same methods were used for leaderboard and final submission. We then applied gradient boosting (using xGBoost) as the machine learning technique of choice. This allowed for efficient training on all ~50M genomic locations and for inherent feature selection and modeling of interactions between features, with each feature being on (potentially) different scale/ranges. An partial example of the boosted tree xGBoost outputs is shown in Figure IV-15 on page 193).

IV.3.2 Methods

IV.3.2.1 Processing DNase-seq data

DNase-seq data was processed into counts (normalized to 1M reads) of just the 5`-ends of the reads. BigWig files were generated for both Watson and Crick strands separately. For samples for which more than one technical replicate exists, the BAM files were merged. An all-to-all correlation matrix was generated for all tracks (using *BigWigCorrelate (Kent et al. 2010)*). The signal files were then further processed in the following ways:

1. Gaussian smoothing using kernel with standard deviations of 0.5, 1, 2, 3, 4, 5, 10, 15, 20, 30, 40, 50, 100, 125, 150, 200, 300, 400, 500, 1,000, 2,000, 3,000 was applied and for each 50bp center of the training/testing regions the maximum of the smoothed signal was taken.

2. The sum of the signal in the 50bp center of the training/testing windows was taken.
3. The sum of the signal in the 200bp center of the training/testing windows was taken.
4. The maximum of the signal in each 200bp window was taken.
5. The average footprint score (described below) for the 50bp center of each window was taken (covered bases only).
6. The maximum footprint score for the 50bp center of each window was taken.

In 1, 5 and 6, for every 50bp window, the cleavage signal was normalized to the mean signal in the local region of 10,000bp (counting only covered bases).

IV.3.2.2 Footprint score

The entire genome was scanned using a kernel constructed to pick up the characteristic footprint signal (flanking region on one strand, followed by a cleavage-depleted region and the same signal mirrored on the opposite strand, as noted in (Piper et al. 2013). For the Watson strand signal, we used a kernel based on the first derivative of a Gaussian PDF and for the Crick strand we used the same kernel (see Figure IV-14 on page 192) but mirrored. Figure IV-14 also shows the kernel for both Watson and Crick strand in black and yellow. The convolution was run at several scales and summed.

IV.3.2.3 Features for RNA-seq

The provided RNA-seq data were transformed into BED files containing regions for each expressed gene and their associated TPM value. These regions we then used to produce multiscale features similar to what was done for DNase-seq. That is, for each expressed gene (TPM > 1.0) a region around the transcription start site (TSS) was marked with a flag. The sizes of the marked regions we used were 1kb, 2kb, 3kb, 5kb, 10kb, 20kb, 50kb, 100kb and 500kb. Also, several different types of marks were used: a simple binary flag, a Gaussian smoothed binary flag, the actual Gaussian smoothed TPM and a smooth neglog(p-value) for the f-score of each gene (i.e. indicating the variance of the respective gene's expression across all training cell types). Colloquially known as the "eugenio-score," for its inventor.

IV.3.2.4 Features based on sequence

For the all provided 200bp regions, the GC content was calculated and provided as a feature. Also, for all TFs, if a PWM was found in the JASPAR 2016 core motif database (as part of the MEME suite (Bailey et al. 2009)), the maximum -log(p-value) was calculated in each 200bp window, using FIMO.

IV.3.3 Results

We generated a series of 4 whole genome models, using different combinations of features and model parameters:

1. DNase + FIMO features (tree depth 10)
2. DNase + FIMO + gene expression (tree depth 10), +/- 200 bp offsets
3. DNase + FIMO + gene expression (tree depth 10)
4. DNase + FIMO + gene expression (tree depth 6)

Comparison of these models for the Leaderboard training TFs appears in Figure IV-17 on page 195, with model 4 having best performance across the board. While this result may initially appear surprising—model 4 utilizes trees not as deep as model #3—this is a common scenario in machine learning (and frequently with XGBoost) where the model either is overfitting the data, or other model parameters were not tuned in concordance with the tree depth (Hawkins 2004). We examined the models for feature importance; as shown in Figure IV-16 on page 194, TF motif and DNA sequence (via FIMO and GC content) were the most highly important features. We then utilized model #4 for the TFs being publically tested on the DREAM leaderboard. Our performance in general was highly competitive; as shown in Figure IV-18 and Figure IV-19, a sub-selection of the TF Leaderboard, we scored second for AIRD3A and REST and fourth for ATF7, by team. We ultimately finished Round 1in 9th place in the competition. Figure IV-20 on page 198 shows a comparison of our performance against the best performing model for the Leaderboard TFs. On average, our model trailed the best-performing model by an average aucPR of 0.17 with a standard deviation of 0.13 aucPR. We were the best or near-best model for several TFs, including REST and STAT3, as well as TCF7L2, a TF with established involvement in glucose metabolism and Type 2 diabetes (Savic et al. 2011).

IV.4 Enhancer Prediction

IV.4.1 Introduction

Enhancers are sequences of DNA involved in increasing gene expression. This *cis*-acting regulatory elements are typically short (50 to 1500 bp) and activate their target genes under specific cellular conditions or developmental time points (Blackwood and Kadonaga 1998). The structure of the enhancer complex and formation of the enhanceosome were first studied structurally in the human IFNβ gene; activation of this gene in response to viral infection required recruitment of multiple transcription factors and HMG-1Y (Thanos and Maniatis 1995). Studying enhancers has become increasingly complex. While hundreds of enhancers were found through discovery of ultraconservative regions of the mouse, rat, and human genomes (Bejerano et al. 2004), we now know the number of putative enhancer regions is in the hundreds of thousands (Shen et al. 2012): far higher than the approximately 20,000 protein-coding genes in the human genome (Ezkurdia et al. 2014).

Enhancers are typically within a ~1 MB window upstream or downstream of a gene (Gillies et al. 1983), but are scattered through the 98% of the genome that doesn't code for protein—billions of bases (Pennacchio et al. 2013). This distal regulation by enhancers is possible through loops in the 3D structure of the genome (Sanyal et al. 2012), allowing multiple distal elements to interact and regulate multiple target genes (Mohrs et al. 2001), furthering complicating determination of the target gene for a particular enhancer. Additionally, since enhancers may only be activated under very particular cellular states or developmental time points, or only in specific cell types or disease conditions, experimental validation of enhancer activity is also fraught with difficulty (Pennacchio et al. 2013).

Experimentally, mouse transgenic assays have had success in decoding *in vivo* enhancer activity for a limited number of enhancer regions. In the transgenic assay, candidate regions are selected, amplified by PCR from human genomic DNA, and then cloned into a reporter vector containing a minimal promoter (Hsp68) and a LacZ reporter gene (Kothary et al. 1988). Primer designs must be done semi-automatically, since primers must be long enough to be effective but are sensitive to exons and highly repetitive sequences in the flanking regions around candidate regions[3]. These vectors are packaged into plasmids that are inserted into mouse eggs, and allowed to grow. The mouse embryos are ultimately viable are harvested on embryonic days 10.5 through 16.5; regions with blue LacZ expression are visually inspected, and, if the same region is found active in at least 3 embryos, the region is classified as active by that enhancer (Pennacchio et al. 2006). Initial studies using this technique found half of ultraconserved regions had enhancer-like activity at a specific developmental time point in a very specific CNS subregion (Visel et al. 2008). These studies also found that, while many enhancer regions have conservation, an even greater number of enhancers are not conserved in vertebrate evolution (Blow et al. 2010; Schmidt et al. 2010).

Patterns of transcription factor binding and histone modification have also been found in enhancer regions, with, unfortunately, no single TF or mark distinguishing all enhancer sites (Visel, Rubin, and Pennacchio 2009). The EP300 transcription factor, a known acetyltransferase and transcriptional coactivator, binds to enhancer regions; only a few thousand EP300 binding sites, though, are found in the genome in ChIP-seq experiments (Visel et al. 2009). Likewise, DNA must be accessible for transcription factors to bind, and DNase-seq experiments have establish many putative enhancer locations (with a corresponding large number of false positives) (Dorschner et al. 2004; Thurman et al. 2012). H3K4me1 was the first histone modification to be found enriched at enhancers (Heintzman et al. 2007), but the mark is broad, covering large genomic

[3] http://wiki.encodedcc.org/index.php/ENCODE_Teleconference_Information

regions well beyond the enhancer, and is also found at the 5’ end of actively transcribed genes (Calo and Wysocka 2013). The H3K27ac histone modification has been experimentally demonstrated to distinguish active from inactive enhancers (Creyghton et al. 2010), and we have utilized this mark extensively when building the ENCODE Encyclopedia (see Chapter II: Building and Visualizing an Encyclopedia of ENCODE candidate Regulatory Elements on page 9). These TFs and histone modifications have been utilized in computational models to predict enhancer locations in supervised and unsupervised models (Ernst and Kellis 2012; Rajagopal et al. 2013; Erwin et al. 2014). These models all suffer from lack of experimentally validated true positive enhancer regions, though, and either find a fraction of putative enhancers (Erwin et al. 2014), or have a large number of false positives (Zacher et al. 2017).

NHGRI and the ENCODE Project have recently made functional characterization of putative enhancers a priority for the consortium[4]. ENCODE now helps not only select regions for transgenic mouse assays for enhancer activity, but also sponsors “bakeoffs” to compare sensitivity and specificity of computational models predicting enhancer activity in a cell-type specific fashion. I have been developing and refining supervised machine learning models for the last several rounds of ENCODE enhancer predictions, building upon the work of Sowmya Iyer (Iyer 2015).

IV.4.2 Methods

As the basis for our models, we utilized experimentally validated enhancers as found in the VISTA Enhancer Browser (Visel et al. 2007), which contains a list of all mouse

[4] https://grants.nih.gov/grants/guide/rfa-files/RFA-HG-16-003.html

transgenic enhancer assays. We have worked on combining hundreds of new ENCODE mouse datasets into multiple machine learning models to best predict whether a given genomic region acts as an enhancer *in vivo*. We have also now trained models across multiple mouse embryonic developmental time points. The goal is to ultimately produce a genome-wide set of predictions predicting enhancer activity for different biosamples and developmental time points.

We used up to 723 ENCODE mouse ChIP-seq histone experiments, as well as 2 mouse bisulfide experiments, from embryonic days 10.5 through 16.5. We also used RNA-seq data, GC content, and mouse conservation datasets. Features were based upon the interval mean of q-values (when available) or signal from narrowPeak or broadPeak files. Features were variance normalized and rescaled to be [0,1]. To select regions for training and testing, we binned the whole genome into 1500 bp bins with 500 bp overlap. Training was done in two stages (like EnhancerFinder): first, a model was trained on positive enhancers in any tissue. Then, the genomic intervals with p-value >0.5 were selected, and used as intervals to train separate random forest models on biosample specific VISTA enhancers. Models were trained on 80% of the samples in the feature matrix, with 30x cross-validation.

Training labels were collected from the VISTA enhancer set of experimentally validated true positives and true negatives; Table IV-1 on page 178 shows the number of tested enhancer regions per biosample. Negatives were re-shuffled using “bedtools shuffle” to avoid areas of high evolutionary conservation. While 8 classifiers were initially used, random forest and gradient boosting consistently performed best. When then tested the models on the remaining 20% of the data. We also the R package RankAggreg to combine output from random forest and gradient boosted models methods using weighted rank aggregation (Pihur, Datta, and Datta 2009).

IV.4.3 Results

We trained models on the entire VISTA Enhancer set, and tested on genomic coordinates provided by the assay lab. During testing, we found no one optimal machine learning methods or feature set. For example, for predicting forebrain enhancer activity in forebrain tissue, building a random forest model on the entire ChIP-seq dataset performed best (Figure IV-21 on page 199). On the other hand, to predict forebrain enhancers that also may share enhancer activity in other tissues, a gradient boosting model with a more limited set of developmental time point-specific enhancers had an aucPR improvement of 0.059 over the above random forest model on a large set of features (Figure IV-22 on page 200). Likewise, for heat-specific enhancers, a gradient boosting model on developmental time point-specific data combined with whole genome bisulfide data outperformed all other models (Figure IV-23 on page 201), while a random forest model on developmental time point-specific data bested models for heart enhancers active in other tissues (Figure IV-24 on page 202).

For predicting enhancers in midbrain on embryonic day 11.5, my gradient boosting model with 823 ENCODE ChIP-seq datasets surpassed all other models (including ensemble approaches combining all models together) for predicting midbrain-specific enhancers (Figure IV-25 on page 203). For predicting midbrain e11.5 enhancers active in other tissues, random forest on the same feature matrix was the best performer, bested only by ensemble learners based on all models (Figure IV-26 on page 204).

IV.5 Overall Machine Learning Discussion

Machine learning of epigenetic data is on the precipice of heralding in a new world of hybrid experimental and computational techniques; while experiment will always be gold standard for data, predictive computational techniques can help bridge the gap when experiments are too expensive, technically difficult, or numerous. Imputation of experimental data is also one possible source of quality control for when the actual experiment is performed (Ebert and Bock 2015). Imputation of transcription factor binding sites may shed light on more intricate co-binding and tethered binding patterns: for instance, we purposefully eliminated CTCF as a feature when predicting members of the complex SMC3 or RAD21 (and vice-versa), since inclusion of these known co-binding partners made the models artificially accurate (Holwerda and de Laat 2013).

Several TFs demonstrate high difficulty in imputing. Reasons for poor performance can be broken down into several categories: too few datasets to impute on, low quality data, and large differences in TF binding as indicated by low Pearson correlation between cell types for a given TF. Some experiments (in particular DNase-seq) have datasets from labs that use incompatible protocols, making combination of replicates across labs infeasible; just arbitrarily selecting one lab's experiment, though, may skew the results in unintentional ways. Many of the cell lines present in these projects are cancerous; this could skew the model training and evaluation of model performance, as we notice in Figure IV-7 on page 185. Two recent studies on ChIP-seq data have noted an increase in false positive reads from genomic areas with high rates of transcription (Teytelman et al. 2013; Park et al. 2013); this problem is intrinsic to the ChIP-seq experiment, and reduces the accuracy of models.

We feel we have found a viable core combination of features and machine learning algorithms. Due to our own time constraints, we did not fully determine the optimal combination of XGBoost parameters. We also didn't have sufficient time to add in offsetted features in additional to the multi-scale features. After examining the feature importance using XGBoost python scripts, motif is by far the most important feature. A winning group from the first 2 rounds of DREAM also noted sequence and motif-based features were the two most important features (Keilwagen, Posch, and Grau 2017). In the future, we will further investigate this feature, as well as making sure all TFs have motif information. We will also better explore the parameter space. For TFs that don't bind with sequence-specificity, motif will not help, and other approaches may be required.

Machine of learning of enhancers in different tissues and developmental time points poses a different set of challenges. Some tissues undergo significantly less developmental dynamics at the developmental time points we were predicting in. This developmental stability lead to better predictive models, since the state of enhancers was not changing as drastically between time points. The relatively small number of samples from VISTA, though, has made machine learning enhancers difficult, as does the intrinsic bias when selecting regions to test (since regions have been selected by high conservation or high DNase and H3K27ac signal, thereby skewing results towards models using those features).

IV.6 Tables

# enhancers	Tissue
75	Facial mesenchyme
588	Brain
73	Nose
0	Kidney
298	Midbrain
262	Hindbrain
50	Cranial nerve
356	Forebrain
192	Heart
41	Trigeminal V (ganglion, cranial)
196	Neural tube
80	Eye
136	Branchial arch
227	Limb
65	Dorsal root ganglion

Table IV-1 | VISTA Datasets
~2600 enhancers (as of August 2015)

IV.7 Figures

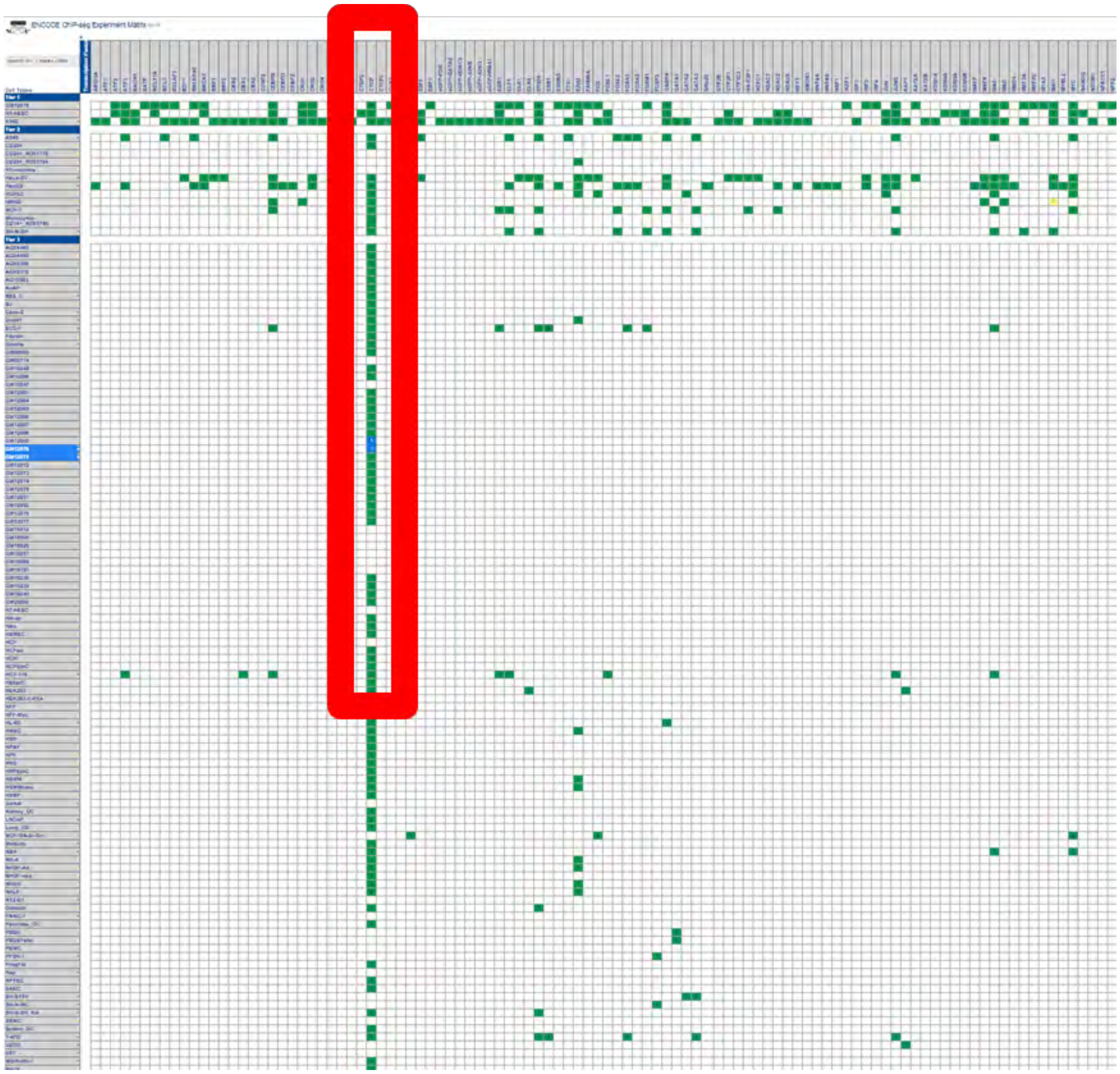


Figure IV-1 | Validate predictions for known ChIP-seq experiments
Matrix of ENCODE TFs vs biosamples; note great sparsity of matrix.

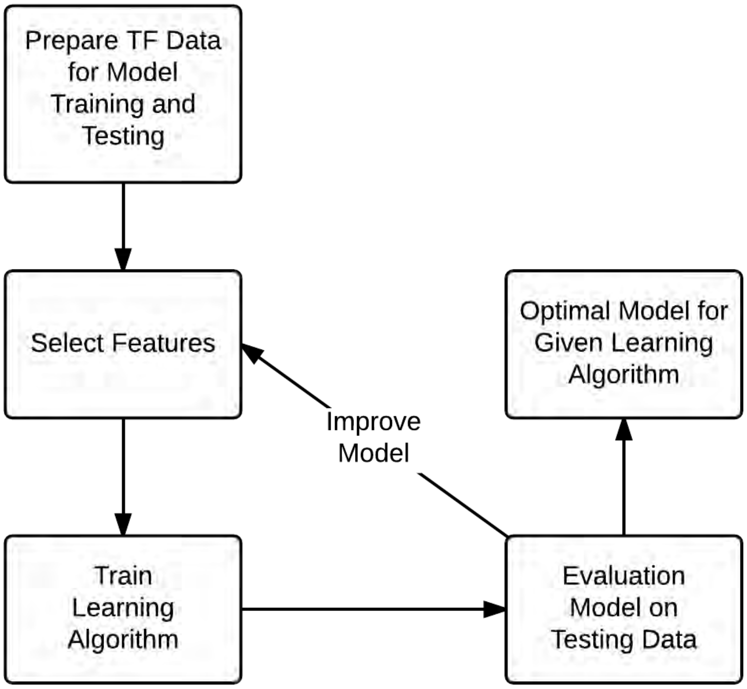

Prepare TF Data for Model Training and Testing
Select Features
Optimal Model for Given Learning Algorithm
Improve Model
Train Learning Algorithm
Evaluation Model on Testing Data

Figure IV-2 | Imputation Flowchart
Typical supervised machine leaning approach; feature engineering is typically the most difficult—and important—step.

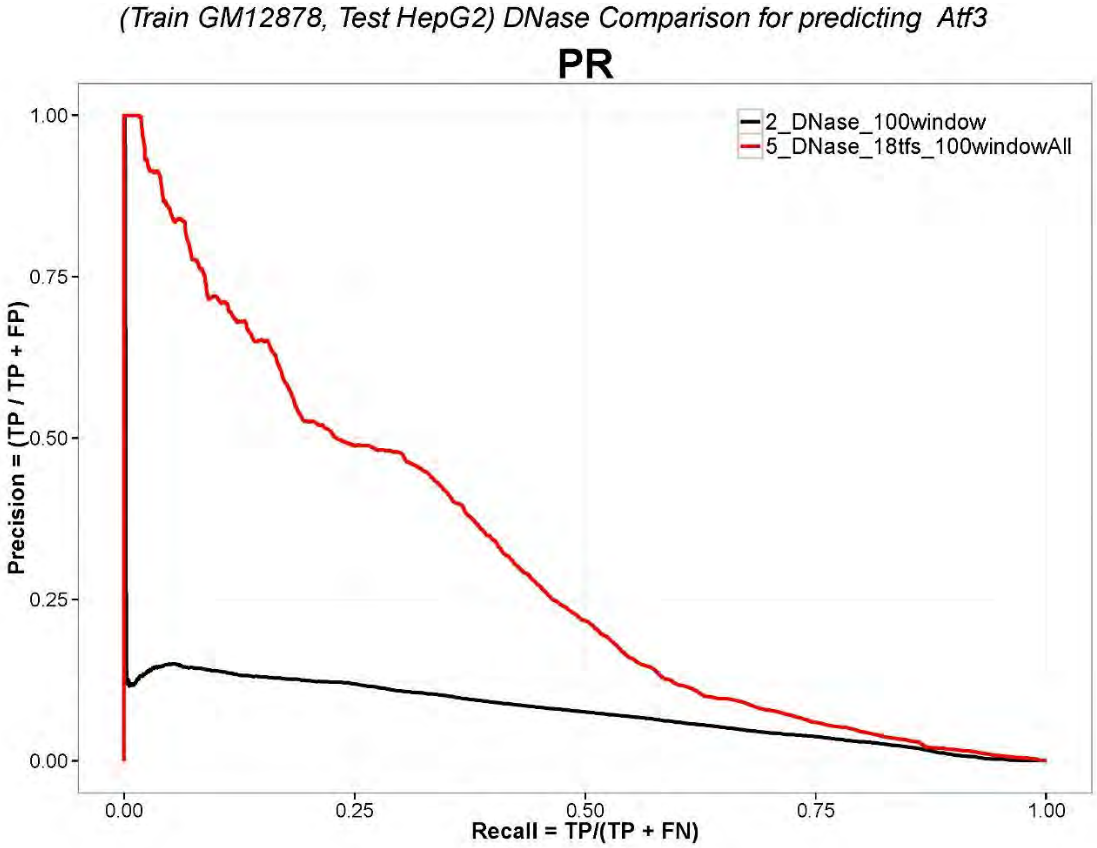


Figure IV-3 | Predict ATF3 binding sites
Predict ATF3 binding sites using just DNase (black line) and DNase with 18 other transcription factors (red line). Smoothing window size of 100 was used for both models.

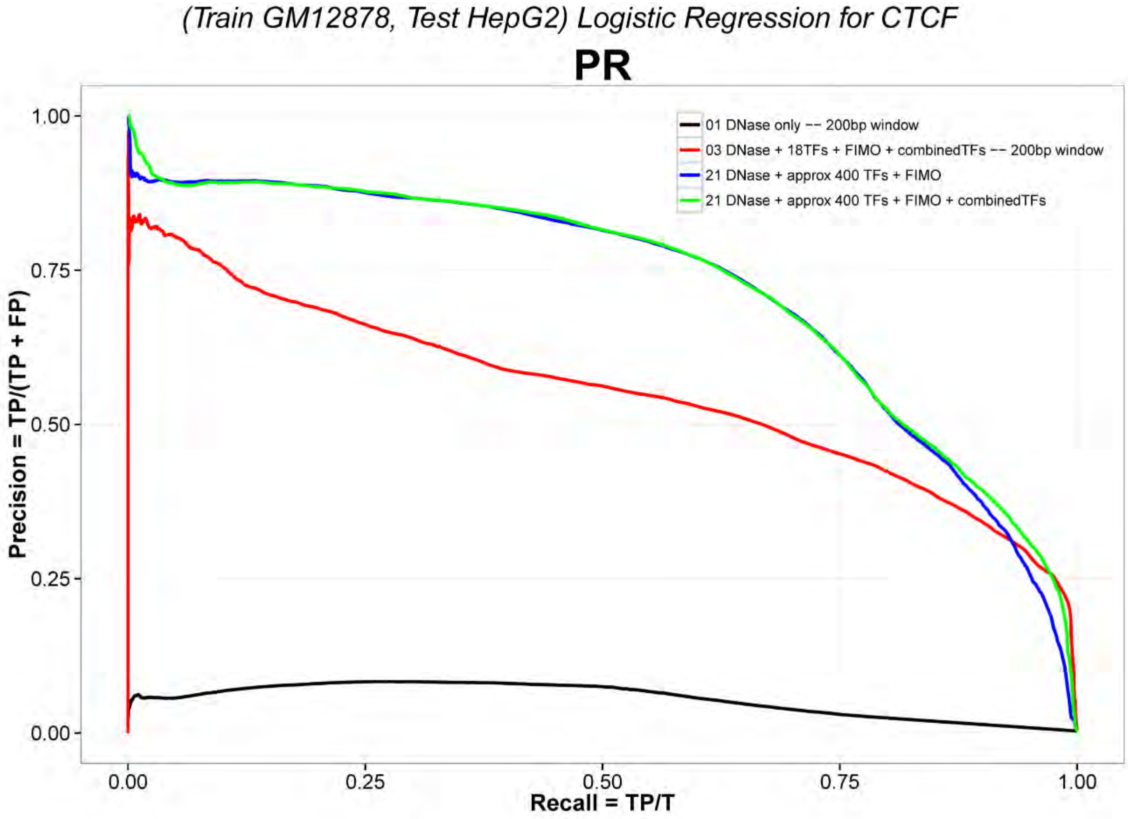


Figure IV-4 | CTCF PR Curve
PR curves for predicting CTCF TFBSs in HepG2 after training in GM12878

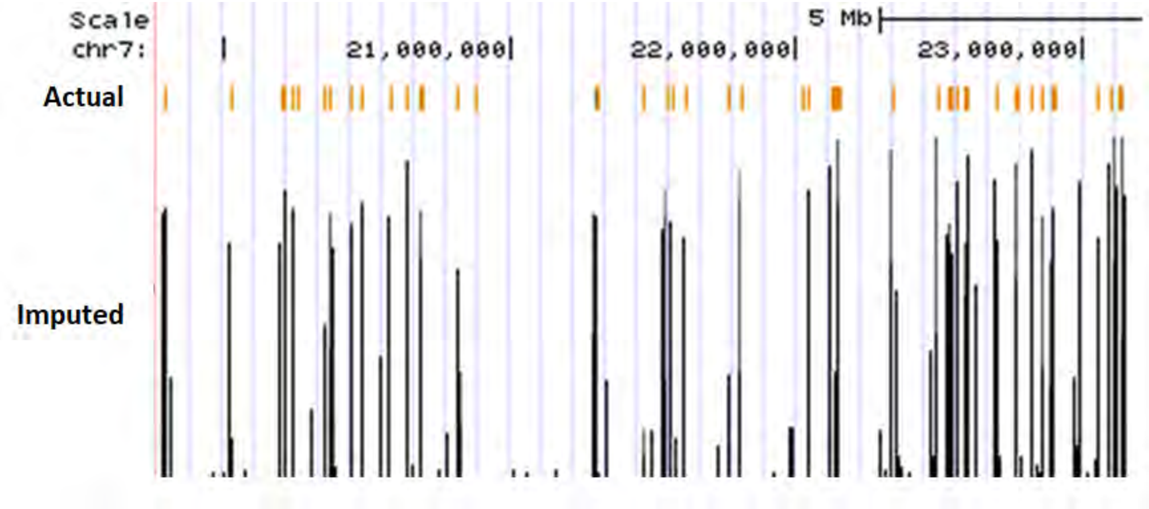


Figure IV-5 | Actual vs Imputed TFBS in UCSC Genome Browser
MACS2 output of logistic regression imputed TFBSs.

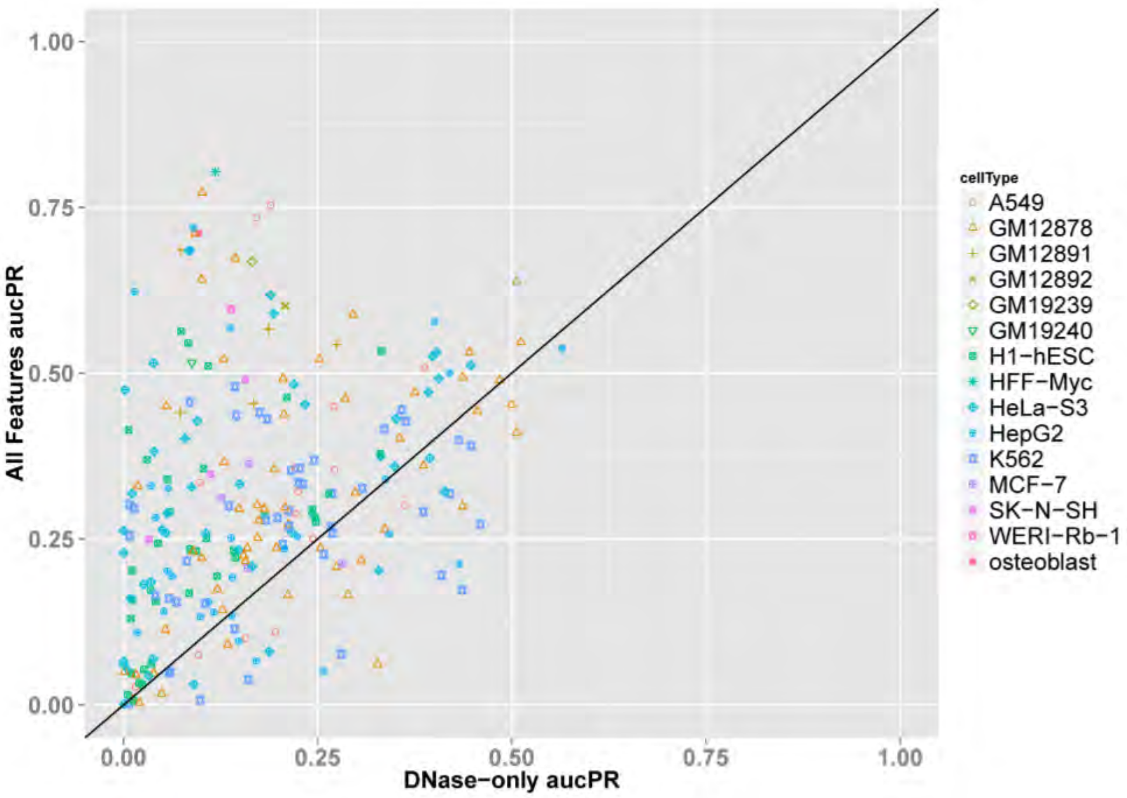


Figure IV-6 Predicting TFBSs in 246 ChIP-seq Datasets
Scatter plot of aucPRs at base-pair resolution in hg19 chromosome 7

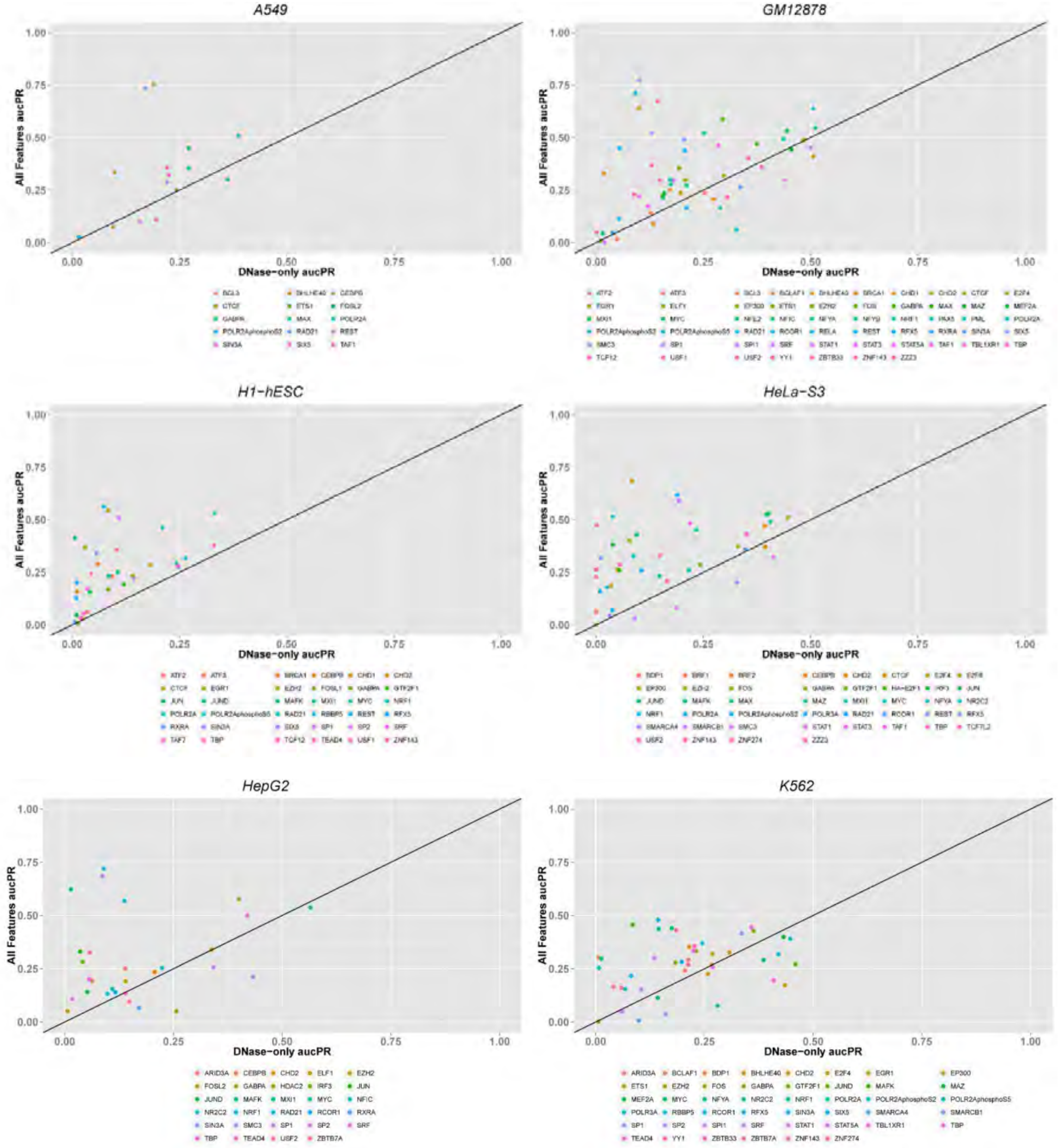


Figure IV-7 | Predicting TFBSs in 246 ChIP-seq Datasets
Scatter plots of aucPR, grouped by biosample, at base-pair resolution in hg19 chromosome 7

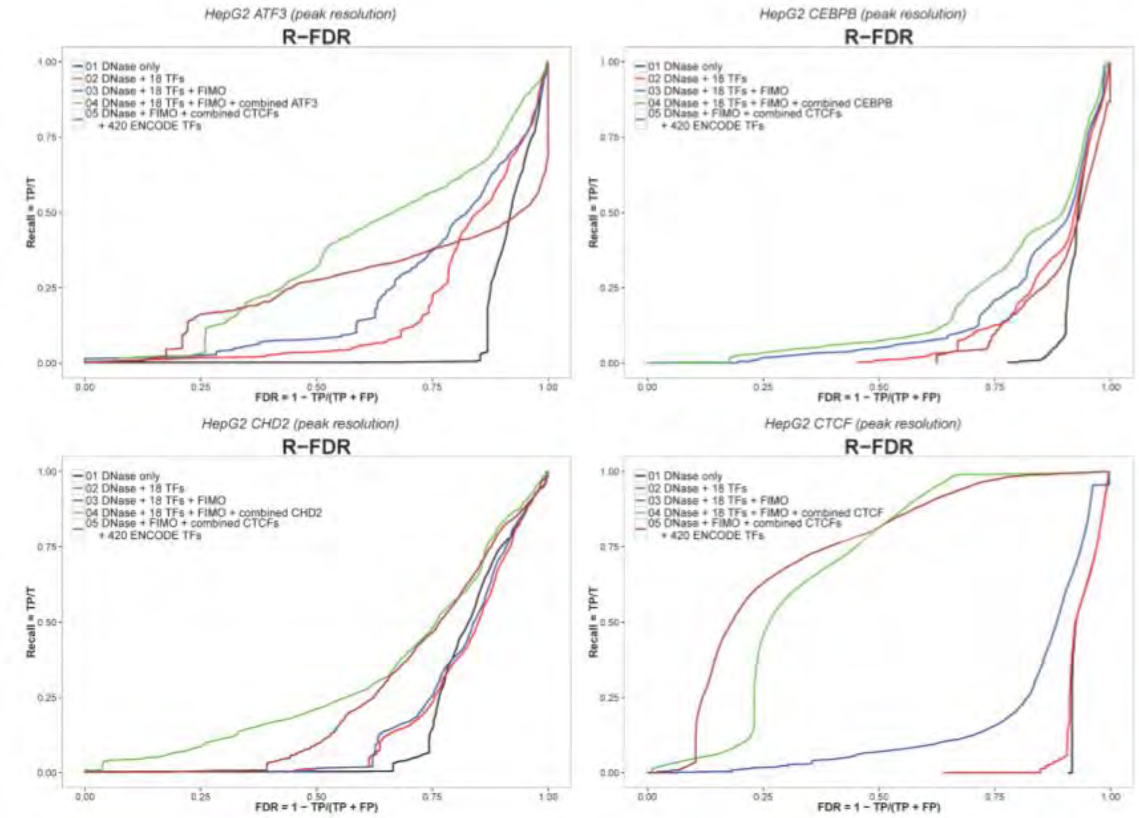


Figure IV-8 | Predicting TFBSs in HepG2
Recall/FDR curves of predictions in HepG2 at base-pair resolution on hg19 chromosome 7. Features include DNase-seq, FIMO, 420 ChIP-seq, and average of target TF. Note: for these plots, highest performance occurs when curves extend furthest into upper-left hand corner of plots.

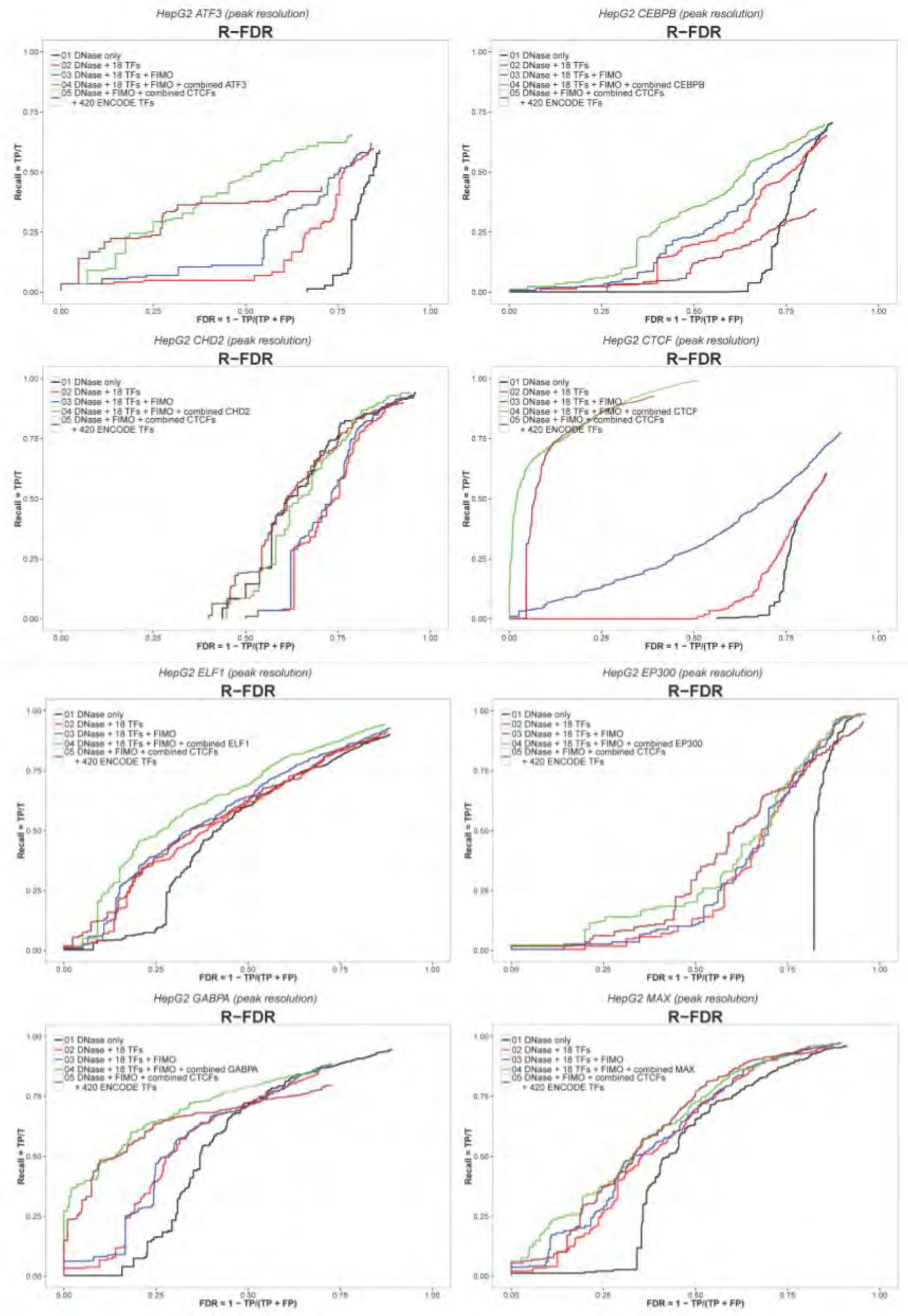


Figure IV-9 | Predicting TFBSs in HepG2

Recall/FDR curves of predictions in HepG2 at peak resolution on hg19 chromosome 7. Features include DNase-seq, FIMO, 420 ChIP-seq, and average of target TF. Note: for these plots, highest performance occurs when curves extend furthest into upper-left hand corner of plots.

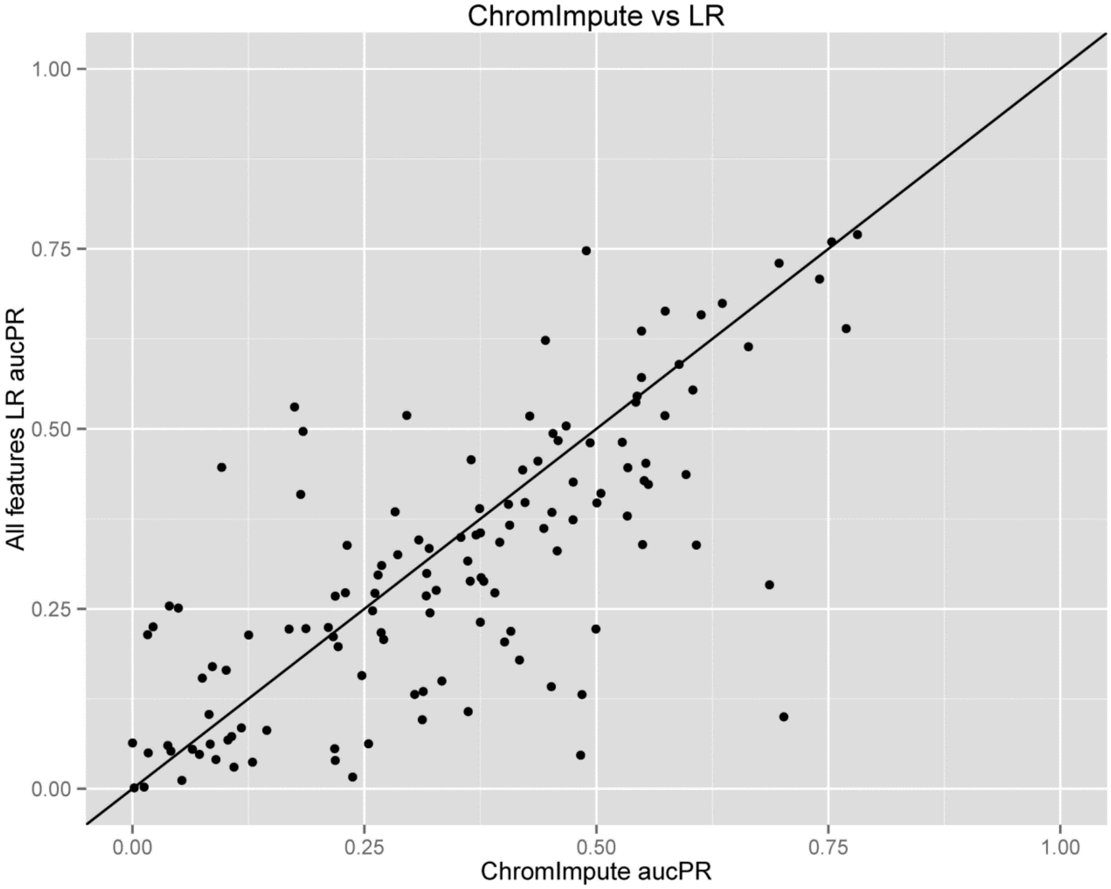


Figure IV-10 | aucPR Curves for ChomImpute vs LR
prAUCs are computed for a test set of ChIP-seq TF experiments for both ChomImpute and logistic regression, and plotted pair-wise.

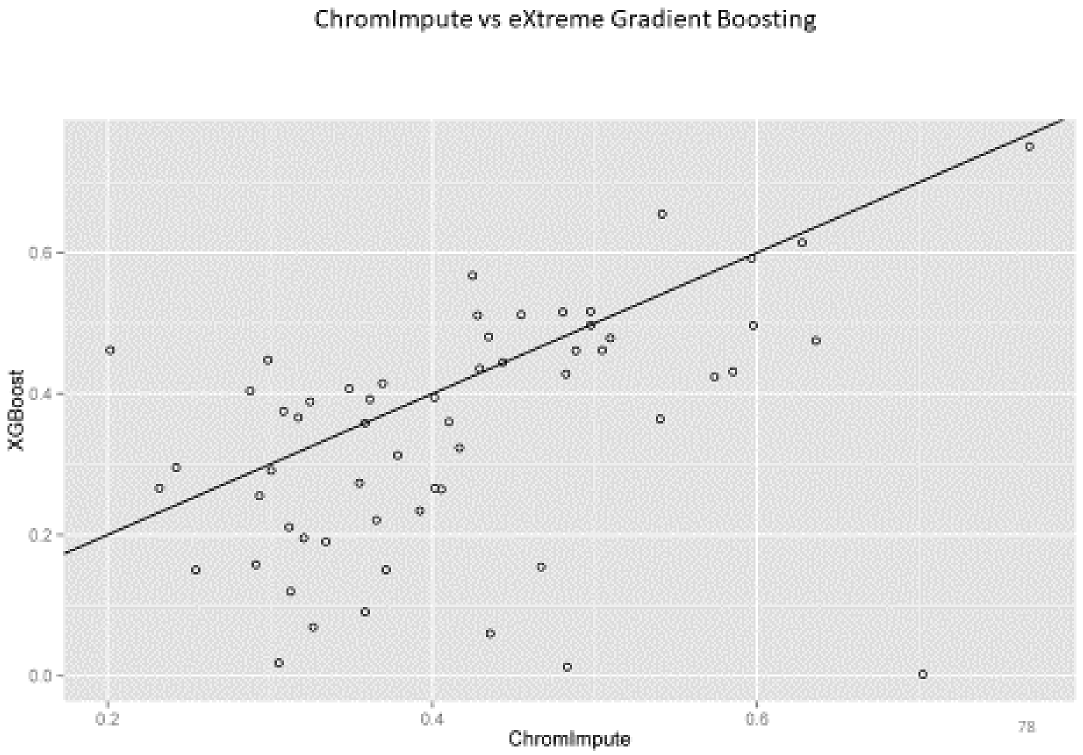
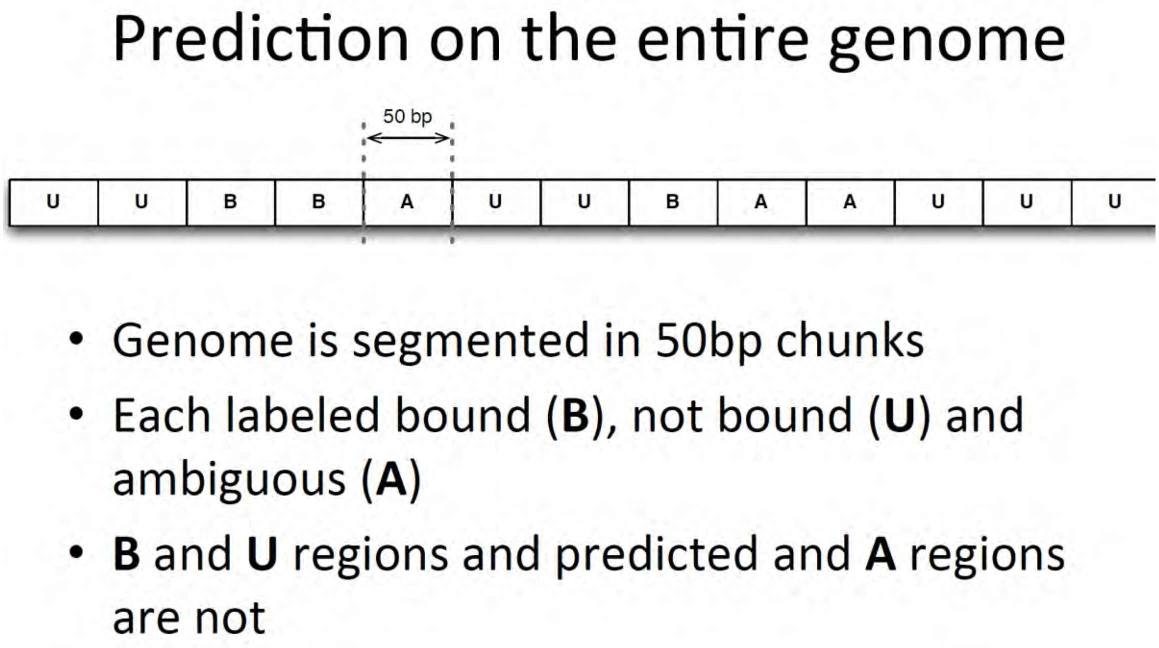


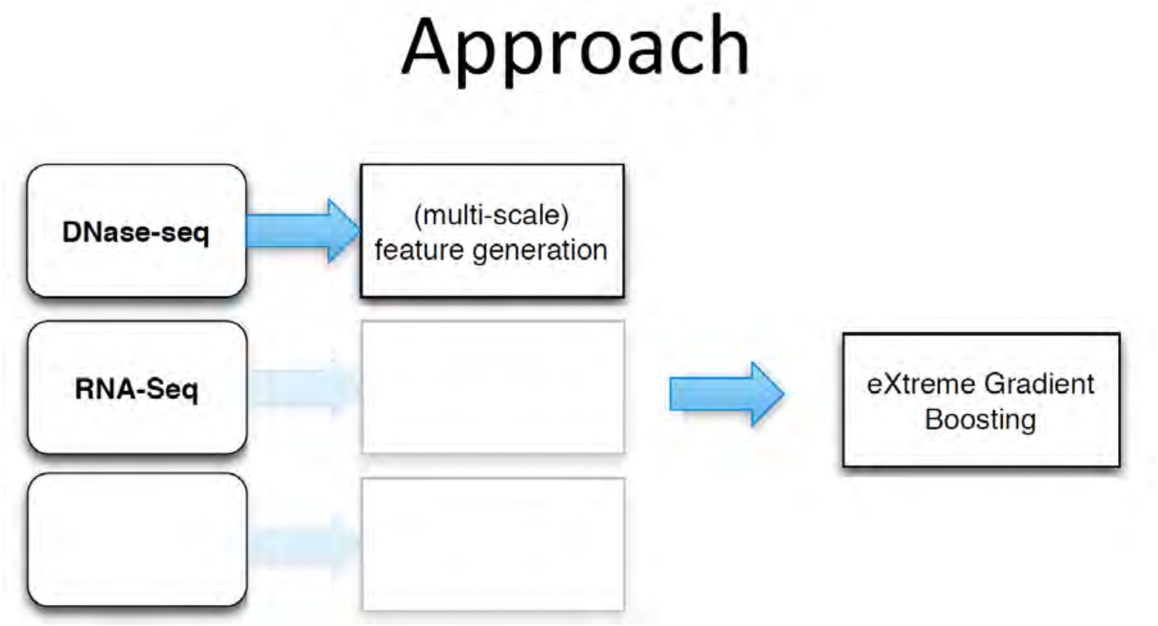
Figure IV-11 | aucPR curves ChromImpute vs xGBoost
prAUCs are computed for a test set of ChIP-seq TF experiments for both ChomImpute and xGBoost, and plotted pair-wise.



Prediction on the entire genome

- Genome is segmented in 50bp chunks
- Each labeled bound (**B**), not bound (**U**) and ambiguous (**A**)
- **B** and **U** regions and predicted and **A** regions are not

Figure IV-12 | DREAM Genomic Binning
Binning of hg19 into 50bp bins as dictated by the DREAM planners. This binning greatly simplifies computational time and space requirements.



Approach

DNase-seq

RNA-Seq

(multi-scale)
feature generation

eXtreme Gradient
Boosting

Figure IV-13 | Machine Learning Approach
We mostly utilized DNase-seq and RNA-seq data (as dictated by the DREAM planners), and used xgBoost for the supervised machine learning.

$$g'(x) = -\frac{1}{2\sigma^2} 2xe^{-\frac{x^2}{2\sigma^2}} = -\frac{x}{\sigma^2} e^{-\frac{x^2}{2\sigma^2}}$$

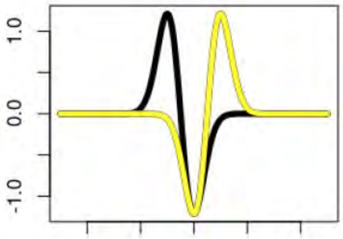


Figure IV-14 | Kernel
Kernel based on the first derivative of a Gaussian PDF. For the Crick strand, we used the same kernel but mirrored. The plot shows the kernel for both Watson and Crick strand in black and yellow.

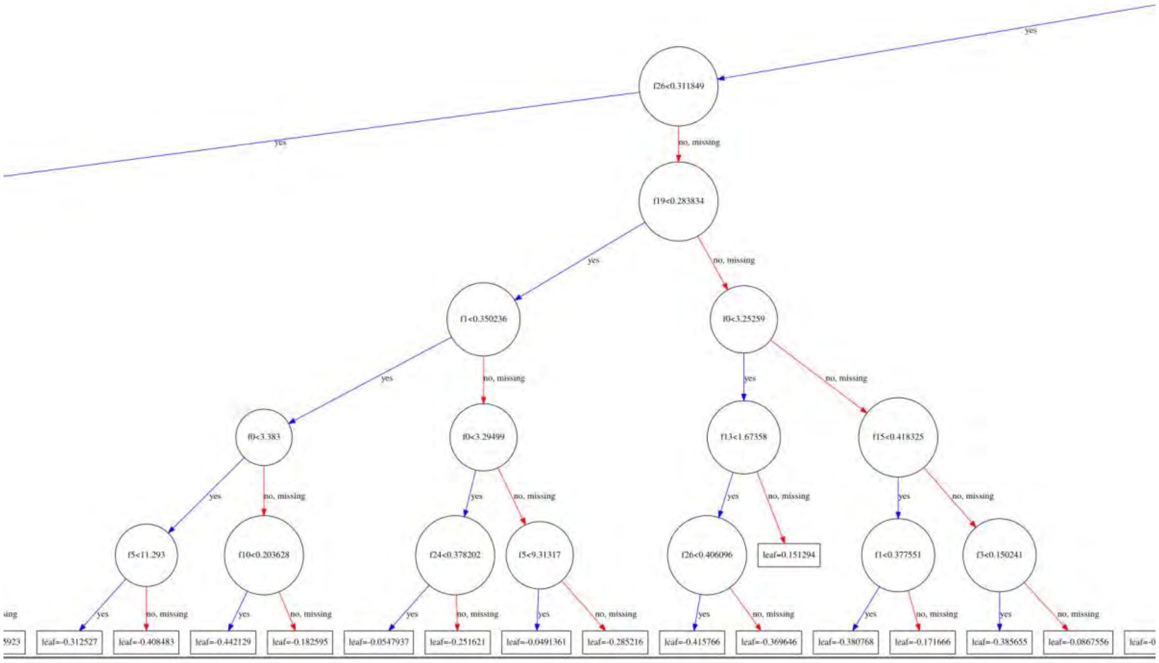


Figure IV-15 | Gradient Boosting Tree Example
Tree of depth 10 for FOXA1 on MCF-7

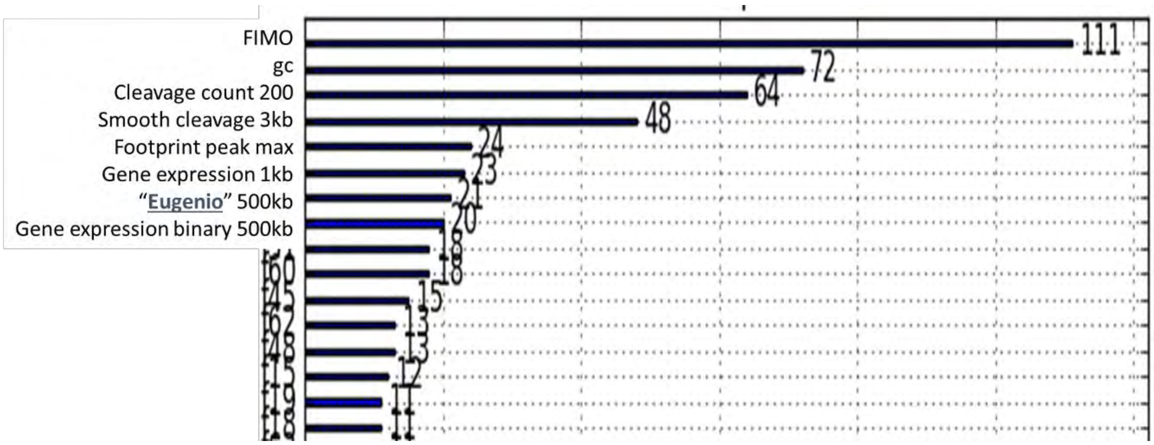


Figure IV-16 | Feature Importance
Example feature importance plot for testing of REST (64 features, 6 tree depth)

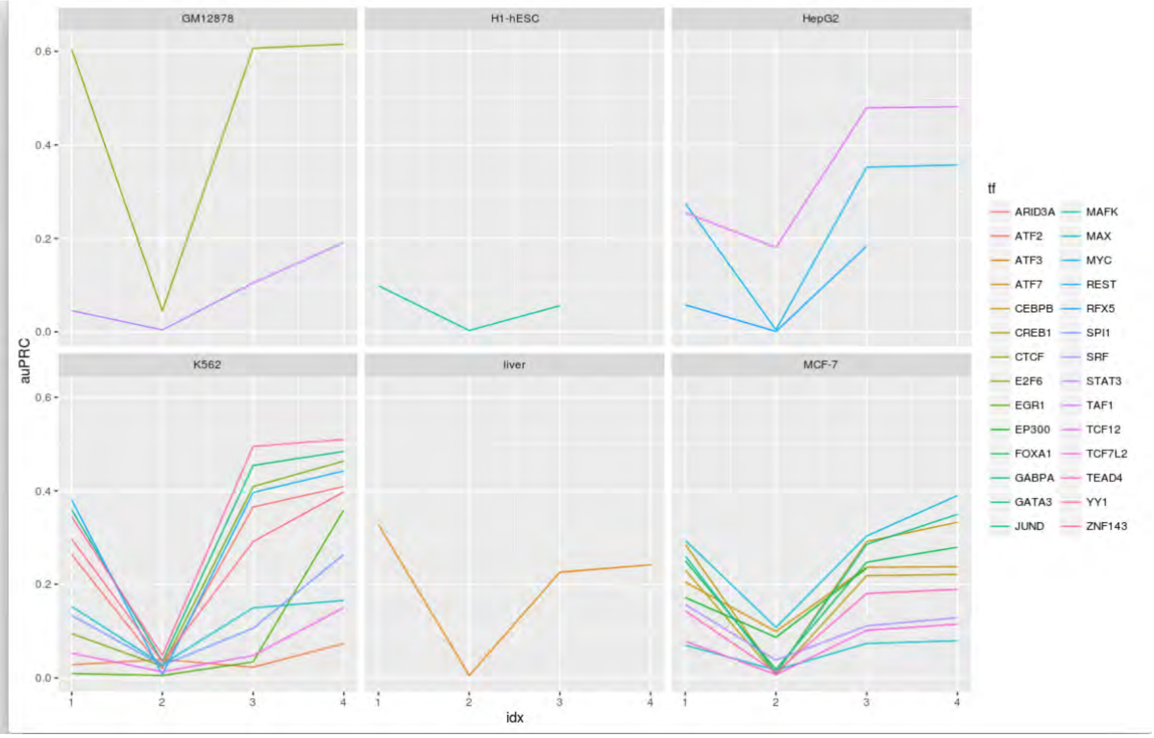
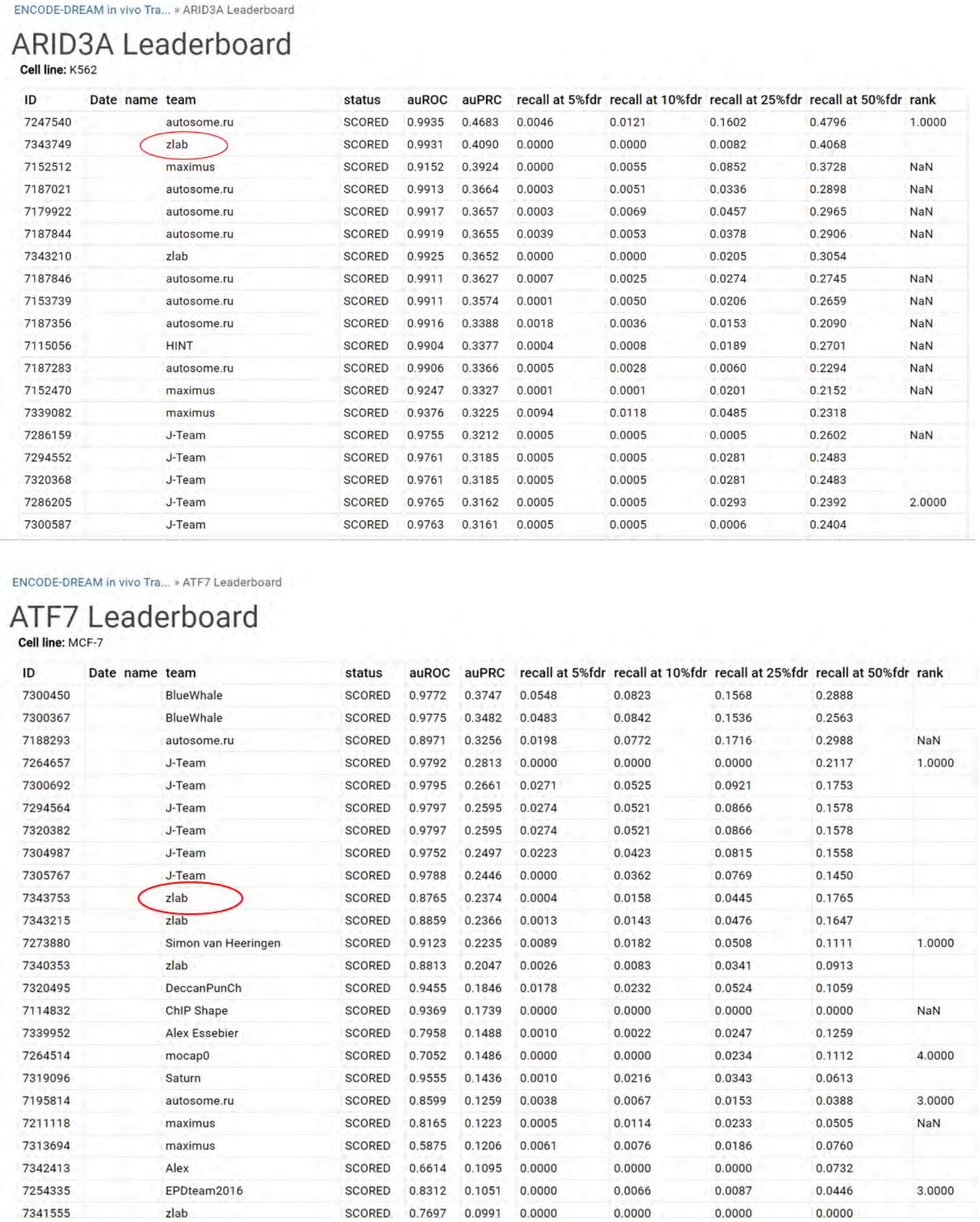


Figure IV-17 | Performance Comparison

Comparison of aucPR curves across different biosamples and TFs for each of our models (numbered 1-4) during Leaderboard Training Round 1



ENCODE-DREAM in vivo Tra... » ARID3A Leaderboard

ARID3A Leaderboard

Cell line: K562

ID	Date	name	team	status	auROC	auPRC	recall at 5%fdr	recall at 10%fdr	recall at 25%fdr	recall at 50%fdr	rank
7247540			autosome.ru	SCORED	0.9935	0.4683	0.0046	0.0121	0.1602	0.4796	1.0000
7343749			zlab	SCORED	0.9931	0.4090	0.0000	0.0000	0.0082	0.4068	
7152512			maximus	SCORED	0.9152	0.3924	0.0000	0.0055	0.0852	0.3728	NaN
7187021			autosome.ru	SCORED	0.9913	0.3664	0.0003	0.0051	0.0336	0.2898	NaN
7179922			autosome.ru	SCORED	0.9917	0.3657	0.0003	0.0069	0.0457	0.2965	NaN
7187844			autosome.ru	SCORED	0.9919	0.3655	0.0039	0.0053	0.0378	0.2906	NaN
7343210			zlab	SCORED	0.9925	0.3652	0.0000	0.0000	0.0205	0.3054	
7187846			autosome.ru	SCORED	0.9911	0.3627	0.0007	0.0025	0.0274	0.2745	NaN
7153739			autosome.ru	SCORED	0.9911	0.3574	0.0001	0.0050	0.0206	0.2659	NaN
7187356			autosome.ru	SCORED	0.9916	0.3388	0.0018	0.0036	0.0153	0.2090	NaN
7115056			HINT	SCORED	0.9904	0.3377	0.0004	0.0008	0.0189	0.2701	NaN
7187283			autosome.ru	SCORED	0.9906	0.3366	0.0005	0.0028	0.0060	0.2294	NaN
7152470			maximus	SCORED	0.9247	0.3327	0.0001	0.0001	0.0201	0.2152	NaN
7339082			maximus	SCORED	0.9376	0.3225	0.0094	0.0118	0.0485	0.2318	
7286159			J-Team	SCORED	0.9755	0.3212	0.0005	0.0005	0.0005	0.2602	NaN
7294552			J-Team	SCORED	0.9761	0.3185	0.0005	0.0005	0.0281	0.2483	
7320368			J-Team	SCORED	0.9761	0.3185	0.0005	0.0005	0.0281	0.2483	
7286205			J-Team	SCORED	0.9765	0.3162	0.0005	0.0005	0.0293	0.2392	2.0000
7300587			J-Team	SCORED	0.9763	0.3161	0.0005	0.0005	0.0006	0.2404	

ENCODE-DREAM in vivo Tra... » ATF7 Leaderboard

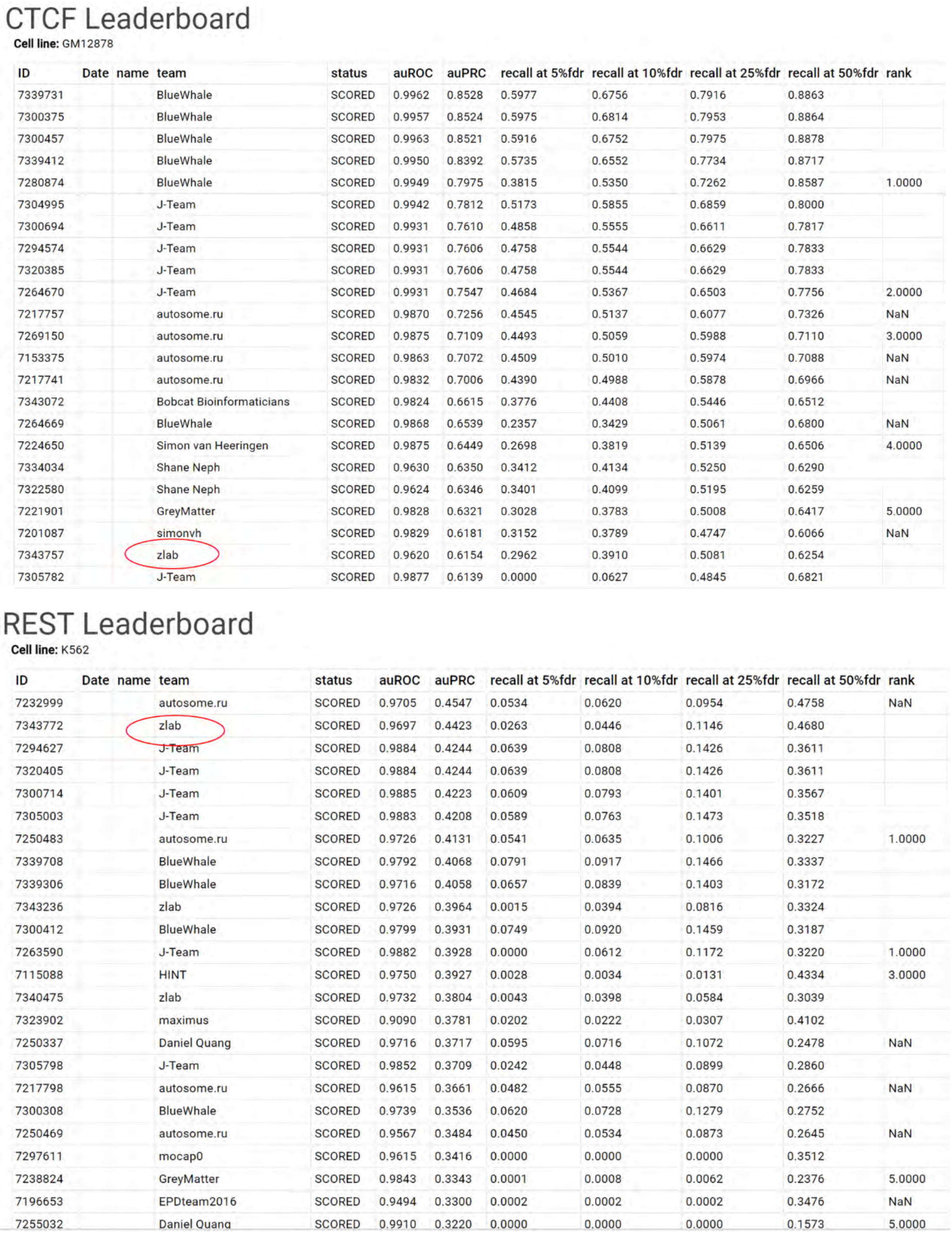
ATF7 Leaderboard

Cell line: MCF-7

ID	Date	name	team	status	auROC	auPRC	recall at 5%fdr	recall at 10%fdr	recall at 25%fdr	recall at 50%fdr	rank
7300450			BlueWhale	SCORED	0.9772	0.3747	0.0548	0.0823	0.1568	0.2888	
7300367			BlueWhale	SCORED	0.9775	0.3482	0.0483	0.0842	0.1536	0.2563	
7188293			autosome.ru	SCORED	0.8971	0.3256	0.0198	0.0772	0.1716	0.2988	NaN
7264657			J-Team	SCORED	0.9792	0.2813	0.0000	0.0000	0.0000	0.2117	1.0000
7300692			J-Team	SCORED	0.9795	0.2661	0.0271	0.0525	0.0921	0.1753	
7294564			J-Team	SCORED	0.9797	0.2595	0.0274	0.0521	0.0866	0.1578	
7320382			J-Team	SCORED	0.9797	0.2595	0.0274	0.0521	0.0866	0.1578	
7304987			J-Team	SCORED	0.9752	0.2497	0.0223	0.0423	0.0815	0.1558	
7305767			J-Team	SCORED	0.9788	0.2446	0.0000	0.0362	0.0769	0.1450	
7343753			zlab	SCORED	0.8765	0.2374	0.0004	0.0158	0.0445	0.1765	
7343215			zlab	SCORED	0.8859	0.2366	0.0013	0.0143	0.0476	0.1647	
7273880			Simon van Heeringen	SCORED	0.9123	0.2235	0.0089	0.0182	0.0508	0.1111	1.0000
7340353			zlab	SCORED	0.8813	0.2047	0.0026	0.0083	0.0341	0.0913	
7320495			DeccanPunCh	SCORED	0.9455	0.1846	0.0178	0.0232	0.0524	0.1059	
7114832			ChIP Shape	SCORED	0.9369	0.1739	0.0000	0.0000	0.0000	0.0000	NaN
7339952			Alex Essebier	SCORED	0.7958	0.1488	0.0010	0.0022	0.0247	0.1259	
7264514			mocap0	SCORED	0.7052	0.1486	0.0000	0.0000	0.0234	0.1112	4.0000
7319096			Saturn	SCORED	0.9555	0.1436	0.0010	0.0216	0.0343	0.0613	
7195814			autosome.ru	SCORED	0.8599	0.1259	0.0038	0.0067	0.0153	0.0388	3.0000
7211118			maximus	SCORED	0.8165	0.1223	0.0005	0.0114	0.0233	0.0505	NaN
7313694			maximus	SCORED	0.5875	0.1206	0.0061	0.0076	0.0186	0.0760	
7342413			Alex	SCORED	0.6614	0.1095	0.0000	0.0000	0.0000	0.0732	
7254335			EPDteam2016	SCORED	0.8312	0.1051	0.0000	0.0066	0.0087	0.0446	3.0000
7341555			zlab	SCORED	0.7697	0.0991	0.0000	0.0000	0.0000	0.0000	

Figure IV-18 | Leaderboard Training Performance

aucPR values across competitors for Leaderboard Training Round 1 for ARID3A and ATF7



CTCF Leaderboard

Cell line: GM12878

ID	Date	name	team	status	auROC	auPRC	recall at 5%fdr	recall at 10%fdr	recall at 25%fdr	recall at 50%fdr	rank
7339731			BlueWhale	SCORED	0.9962	0.8528	0.5977	0.6756	0.7916	0.8863	
7300375			BlueWhale	SCORED	0.9957	0.8524	0.5975	0.6814	0.7953	0.8864	
7300457			BlueWhale	SCORED	0.9963	0.8521	0.5916	0.6752	0.7975	0.8878	
7339412			BlueWhale	SCORED	0.9950	0.8392	0.5735	0.6552	0.7734	0.8717	
7280874			BlueWhale	SCORED	0.9949	0.7975	0.3815	0.5350	0.7262	0.8587	1.0000
7304995			J-Team	SCORED	0.9942	0.7812	0.5173	0.5855	0.6859	0.8000	
7300694			J-Team	SCORED	0.9931	0.7610	0.4858	0.5555	0.6611	0.7817	
7294574			J-Team	SCORED	0.9931	0.7606	0.4758	0.5544	0.6629	0.7833	
7320385			J-Team	SCORED	0.9931	0.7606	0.4758	0.5544	0.6629	0.7833	
7264670			J-Team	SCORED	0.9931	0.7547	0.4684	0.5367	0.6503	0.7756	2.0000
7217757			autosome.ru	SCORED	0.9870	0.7256	0.4545	0.5137	0.6077	0.7326	NaN
7269150			autosome.ru	SCORED	0.9875	0.7109	0.4493	0.5059	0.5988	0.7110	3.0000
7153375			autosome.ru	SCORED	0.9863	0.7072	0.4509	0.5010	0.5974	0.7088	NaN
7217741			autosome.ru	SCORED	0.9832	0.7006	0.4390	0.4988	0.5878	0.6966	NaN
7343072			Bobcat Bioinformaticians	SCORED	0.9824	0.6615	0.3776	0.4408	0.5446	0.6512	
7264669			BlueWhale	SCORED	0.9868	0.6539	0.2357	0.3429	0.5061	0.6800	NaN
7224650			Simon van Heeringen	SCORED	0.9875	0.6449	0.2698	0.3819	0.5139	0.6506	4.0000
7334034			Shane Neph	SCORED	0.9630	0.6350	0.3412	0.4134	0.5250	0.6290	
7322580			Shane Neph	SCORED	0.9624	0.6346	0.3401	0.4099	0.5195	0.6259	
7221901			GreyMatter	SCORED	0.9828	0.6321	0.3028	0.3783	0.5008	0.6417	5.0000
7201087			simonvh	SCORED	0.9829	0.6181	0.3152	0.3789	0.4747	0.6066	NaN
7343757			zlab	SCORED	0.9620	0.6154	0.2962	0.3910	0.5081	0.6254	
7305782			J-Team	SCORED	0.9877	0.6139	0.0000	0.0627	0.4845	0.6821	

REST Leaderboard

Cell line: K562

ID	Date	name	team	status	auROC	auPRC	recall at 5%fdr	recall at 10%fdr	recall at 25%fdr	recall at 50%fdr	rank
7232999			autosome.ru	SCORED	0.9705	0.4547	0.0534	0.0620	0.0954	0.4758	NaN
7343772			zlab	SCORED	0.9697	0.4423	0.0263	0.0446	0.1146	0.4680	
7294627			J-Team	SCORED	0.9884	0.4244	0.0639	0.0808	0.1426	0.3611	
7320405			J-Team	SCORED	0.9884	0.4244	0.0639	0.0808	0.1426	0.3611	
7300714			J-Team	SCORED	0.9885	0.4223	0.0609	0.0793	0.1401	0.3567	
7305003			J-Team	SCORED	0.9883	0.4208	0.0589	0.0763	0.1473	0.3518	
7250483			autosome.ru	SCORED	0.9726	0.4131	0.0541	0.0635	0.1006	0.3227	1.0000
7339708			BlueWhale	SCORED	0.9792	0.4068	0.0791	0.0917	0.1466	0.3337	
7339306			BlueWhale	SCORED	0.9716	0.4058	0.0657	0.0839	0.1403	0.3172	
7343236			zlab	SCORED	0.9726	0.3964	0.0015	0.0394	0.0816	0.3324	
7300412			BlueWhale	SCORED	0.9799	0.3931	0.0749	0.0920	0.1459	0.3187	
7263590			J-Team	SCORED	0.9882	0.3928	0.0000	0.0612	0.1172	0.3220	1.0000
7115088			HINT	SCORED	0.9750	0.3927	0.0028	0.0034	0.0131	0.4334	3.0000
7340475			zlab	SCORED	0.9732	0.3804	0.0043	0.0398	0.0584	0.3039	
7323902			maximus	SCORED	0.9090	0.3781	0.0202	0.0222	0.0307	0.4102	
7250337			Daniel Quang	SCORED	0.9716	0.3717	0.0595	0.0716	0.1072	0.2478	NaN
7305798			J-Team	SCORED	0.9852	0.3709	0.0242	0.0448	0.0899	0.2860	
7217798			autosome.ru	SCORED	0.9615	0.3661	0.0482	0.0555	0.0870	0.2666	NaN
7300308			BlueWhale	SCORED	0.9739	0.3536	0.0620	0.0728	0.1279	0.2752	
7250469			autosome.ru	SCORED	0.9567	0.3484	0.0450	0.0534	0.0873	0.2645	NaN
7297611			mocap0	SCORED	0.9615	0.3416	0.0000	0.0000	0.0000	0.3512	
7238824			GreyMatter	SCORED	0.9843	0.3343	0.0001	0.0008	0.0062	0.2376	5.0000
7196653			EPDteam2016	SCORED	0.9494	0.3300	0.0002	0.0002	0.0002	0.3476	NaN
7255032			Daniel Quang	SCORED	0.9910	0.3220	0.0000	0.0000	0.0000	0.1573	5.0000

Figure IV-19 | Leaderboard Training Performance
aucPR values across competitors for Leaderboard Training Round 1 for CTCF and REST

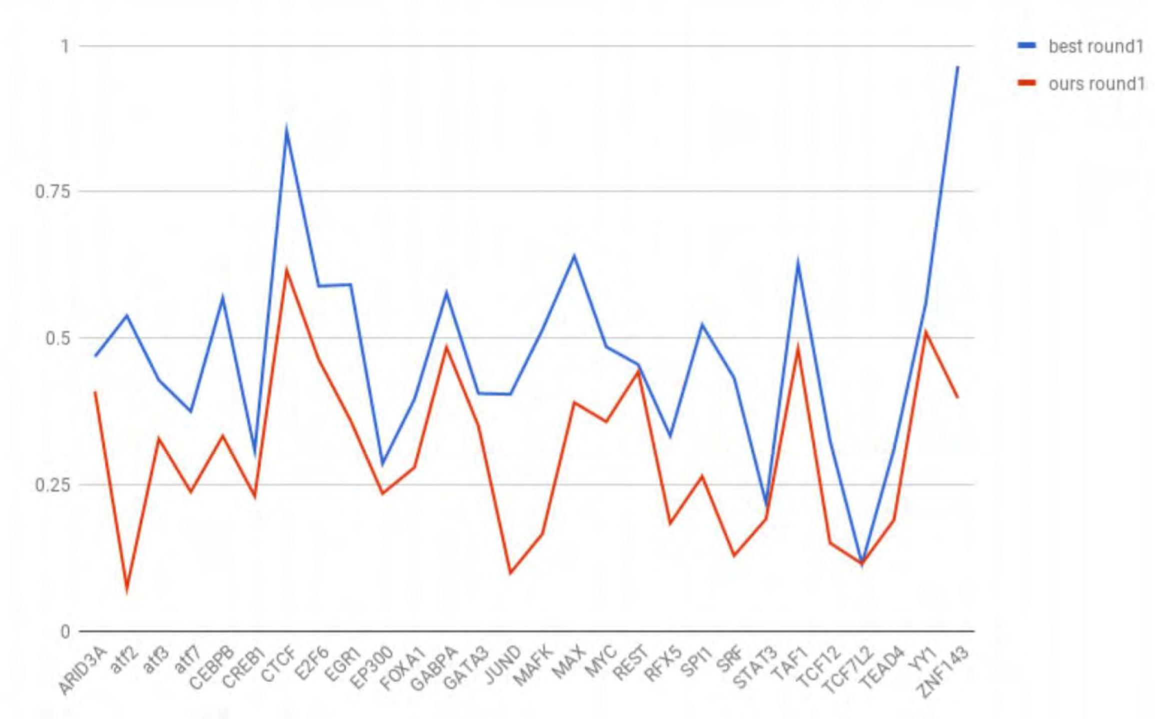
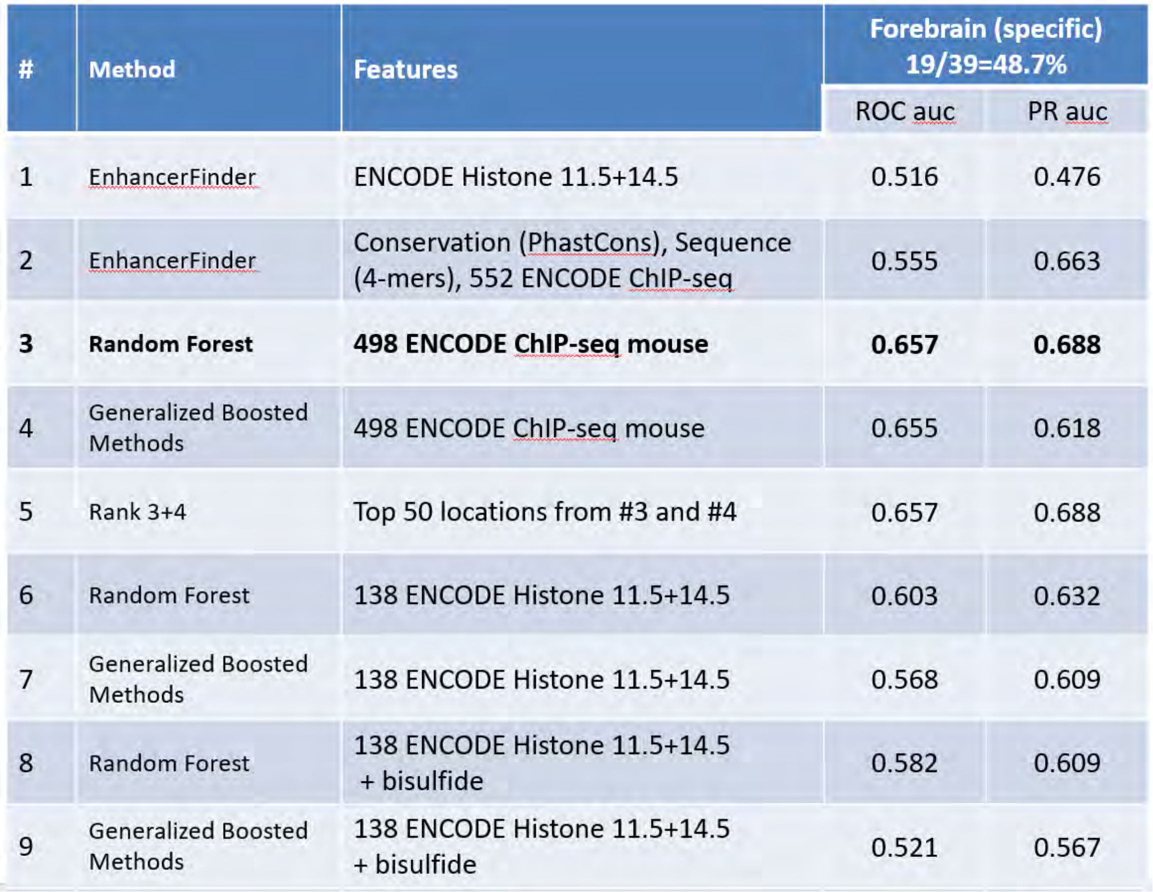
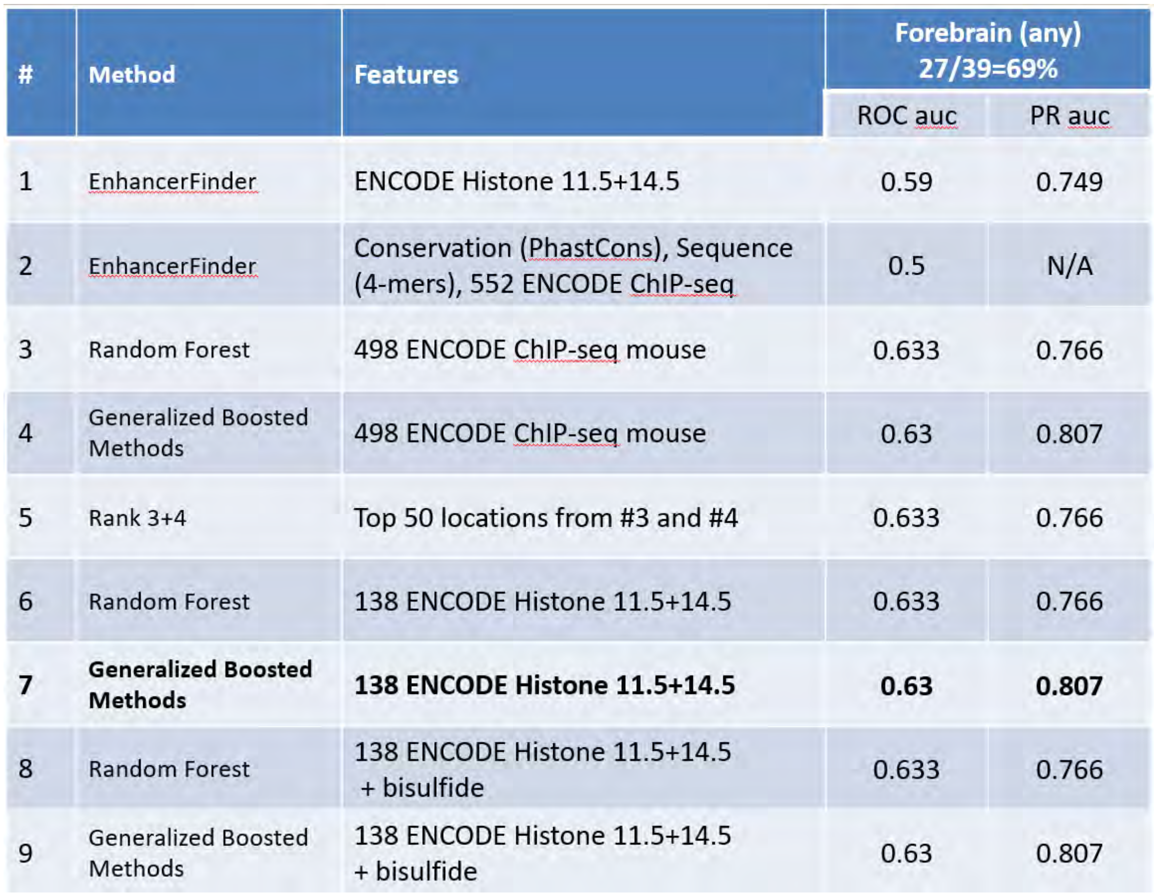


Figure IV-20 | DREAM Leaderboard Testing Performance
Line plots of aucPR across all Leaderboard TFs in Round 1.



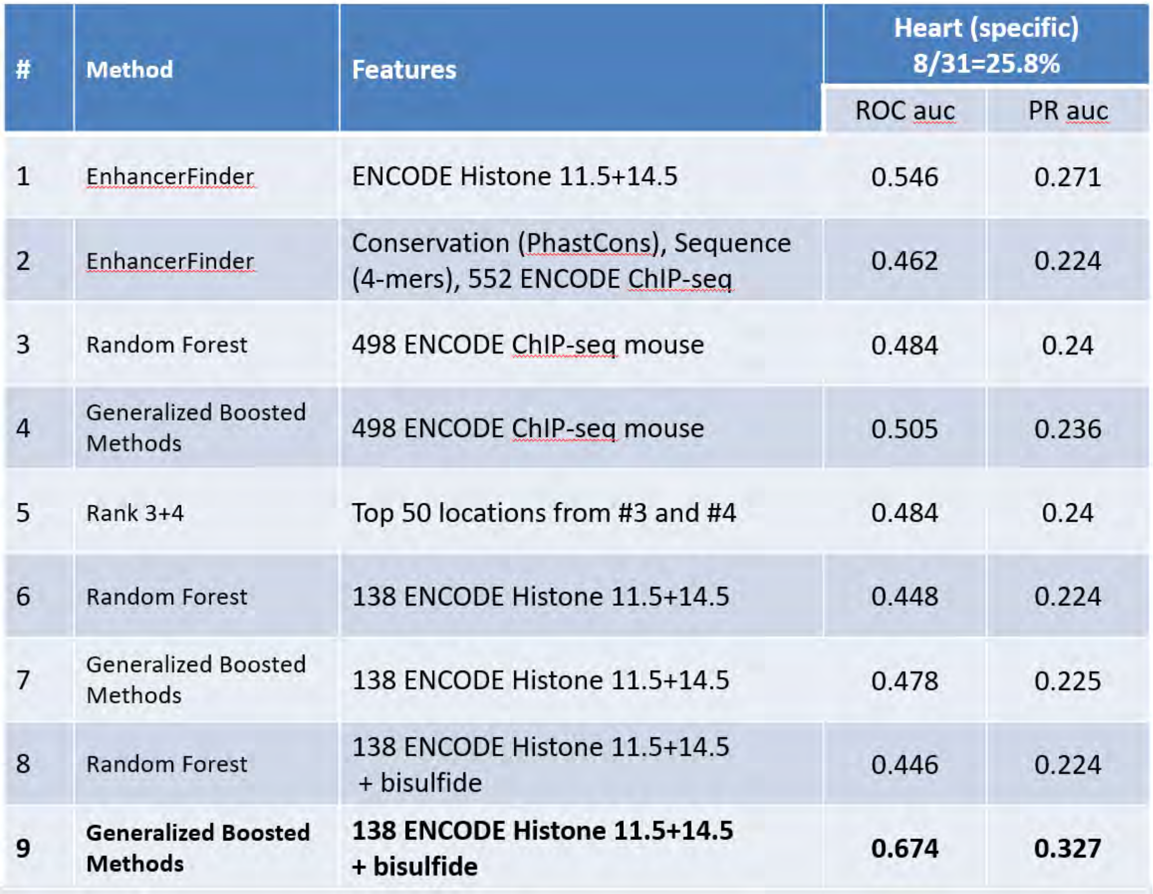
#	Method	Features	Forebrain (specific) 19/39=48.7%	
			ROC auc	PR auc
1	EnhancerFinder	ENCODE Histone 11.5+14.5	0.516	0.476
2	EnhancerFinder	Conservation (PhastCons), Sequence (4-mers), 552 ENCODE ChIP-seq	0.555	0.663
3	**Random Forest**	**498 ENCODE ChIP-seq mouse**	**0.657**	**0.688**
4	Generalized Boosted Methods	498 ENCODE ChIP-seq mouse	0.655	0.618
5	Rank 3+4	Top 50 locations from #3 and #4	0.657	0.688
6	Random Forest	138 ENCODE Histone 11.5+14.5	0.603	0.632
7	Generalized Boosted Methods	138 ENCODE Histone 11.5+14.5	0.568	0.609
8	Random Forest	138 ENCODE Histone 11.5+14.5 + bisulfide	0.582	0.609
9	Generalized Boosted Methods	138 ENCODE Histone 11.5+14.5 + bisulfide	0.521	0.567

Figure IV-21 | Method Comparison for Predicting Forebrain Enhancers
Comparison of our 9 methods for predicting forebrain enhancers in forebrain tissue on training data. Random forest on 498 ENCODE ChIP-seq mouse datasets performs best.



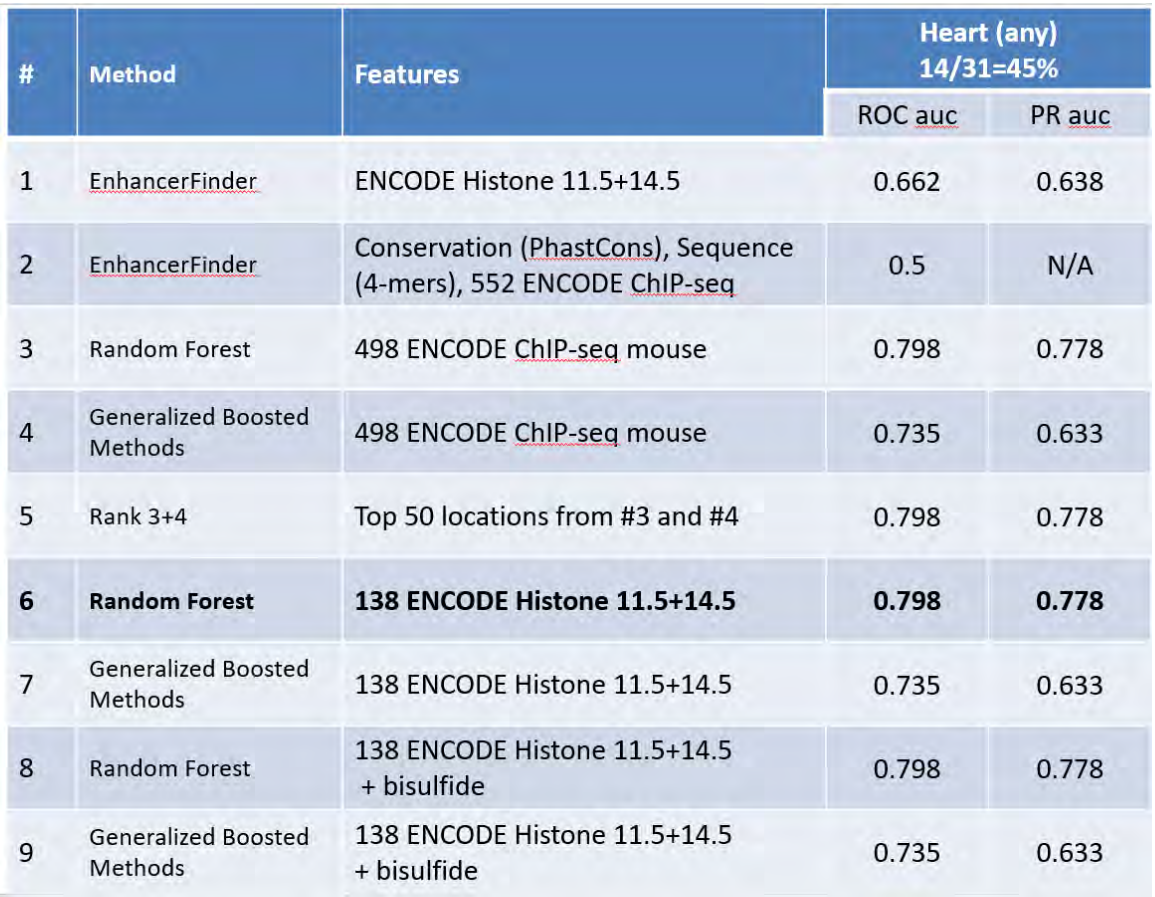
#	Method	Features	Forebrain (any) 27/39=69%	
			ROC auc	PR auc
1	EnhancerFinder	ENCODE Histone 11.5+14.5	0.59	0.749
2	EnhancerFinder	Conservation (PhastCons), Sequence (4-mers), 552 ENCODE ChIP-seq	0.5	N/A
3	Random Forest	498 ENCODE ChIP-seq mouse	0.633	0.766
4	Generalized Boosted Methods	498 ENCODE ChIP-seq mouse	0.63	0.807
5	Rank 3+4	Top 50 locations from #3 and #4	0.633	0.766
6	Random Forest	138 ENCODE Histone 11.5+14.5	0.633	0.766
7	**Generalized Boosted Methods**	**138 ENCODE Histone 11.5+14.5**	**0.63**	**0.807**
8	Random Forest	138 ENCODE Histone 11.5+14.5 + bisulfide	0.633	0.766
9	Generalized Boosted Methods	138 ENCODE Histone 11.5+14.5 + bisulfide	0.63	0.807

Figure IV-22 | Method Comparison for Predicting Forebrain Enhancers
Comparison of our 9 methods for predicting forebrain enhancers in any tissue on training data. Generalized boosting on 138 ENCODE ChIP-seq mouse datasets (just on embryonic days 11.5 and 14.5) performs best.



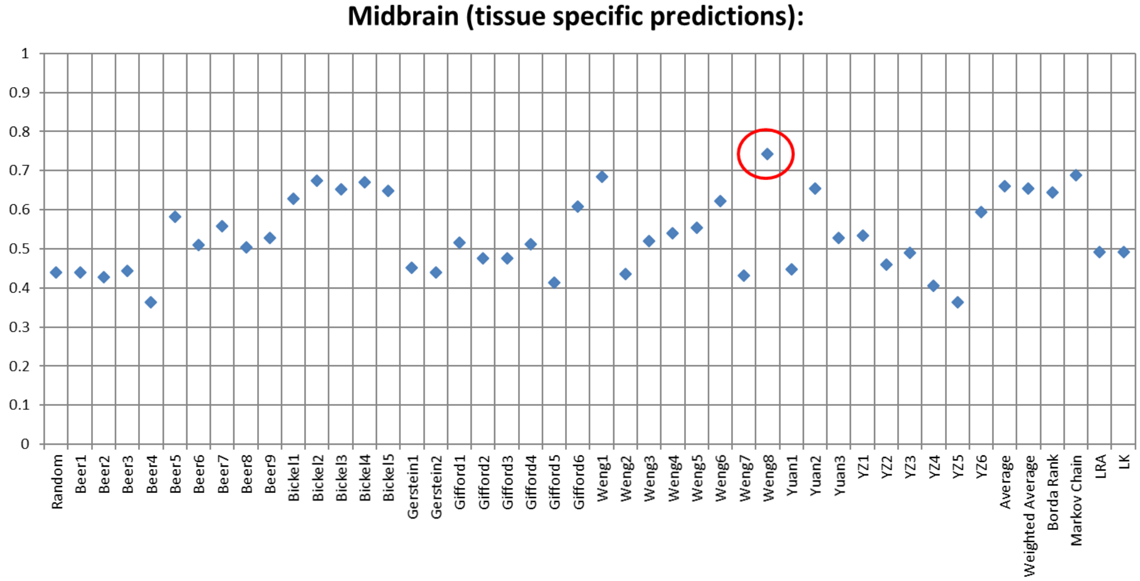
#	Method	Features	Heart (specific) 8/31=25.8%	
			ROC auc	PR auc
1	EnhancerFinder	ENCODE Histone 11.5+14.5	0.546	0.271
2	EnhancerFinder	Conservation (PhastCons), Sequence (4-mers), 552 ENCODE ChIP-seq	0.462	0.224
3	Random Forest	498 ENCODE ChIP-seq mouse	0.484	0.24
4	Generalized Boosted Methods	498 ENCODE ChIP-seq mouse	0.505	0.236
5	Rank 3+4	Top 50 locations from #3 and #4	0.484	0.24
6	Random Forest	138 ENCODE Histone 11.5+14.5	0.448	0.224
7	Generalized Boosted Methods	138 ENCODE Histone 11.5+14.5	0.478	0.225
8	Random Forest	138 ENCODE Histone 11.5+14.5 + bisulfide	0.446	0.224
9	**Generalized Boosted Methods**	**138 ENCODE Histone 11.5+14.5 + bisulfide**	**0.674**	**0.327**

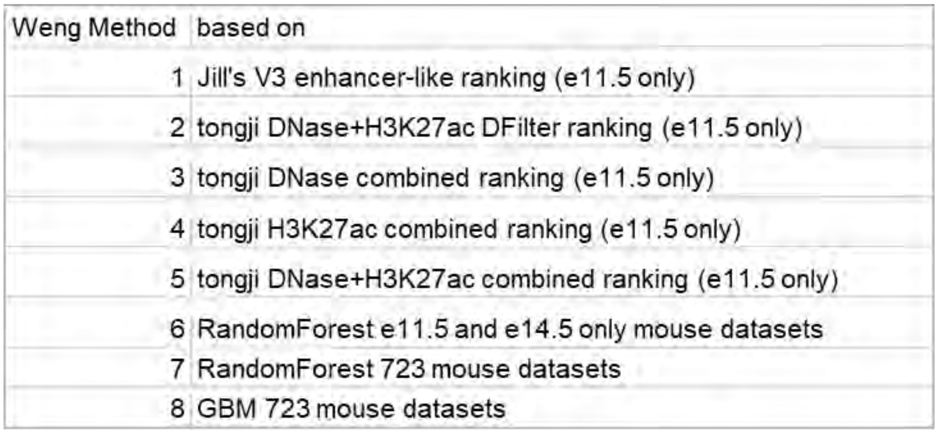
Figure IV-23 | Method Comparison for Predicting Heart Enhancers
Comparison of our 9 methods for predicting heart enhancers in heart tissue on training data. Generalized boosting on 138 ENCODE ChIP-seq mouse datasets (just on embryonic days 11.5 and 14.5) with 2 mouse whole genome bisulfide experiments performs best.



#	Method	Features	Heart (any) 14/31=45%	
			ROC auc	PR auc
1	EnhancerFinder	ENCODE Histone 11.5+14.5	0.662	0.638
2	EnhancerFinder	Conservation (PhastCons), Sequence (4-mers), 552 ENCODE ChIP-seq	0.5	N/A
3	Random Forest	498 ENCODE ChIP-seq mouse	0.798	0.778
4	Generalized Boosted Methods	498 ENCODE ChIP-seq mouse	0.735	0.633
5	Rank 3+4	Top 50 locations from #3 and #4	0.798	0.778
6	**Random Forest**	**138 ENCODE Histone 11.5+14.5**	**0.798**	**0.778**
7	Generalized Boosted Methods	138 ENCODE Histone 11.5+14.5	0.735	0.633
8	Random Forest	138 ENCODE Histone 11.5+14.5 + bisulfide	0.798	0.778
9	Generalized Boosted Methods	138 ENCODE Histone 11.5+14.5 + bisulfide	0.735	0.633

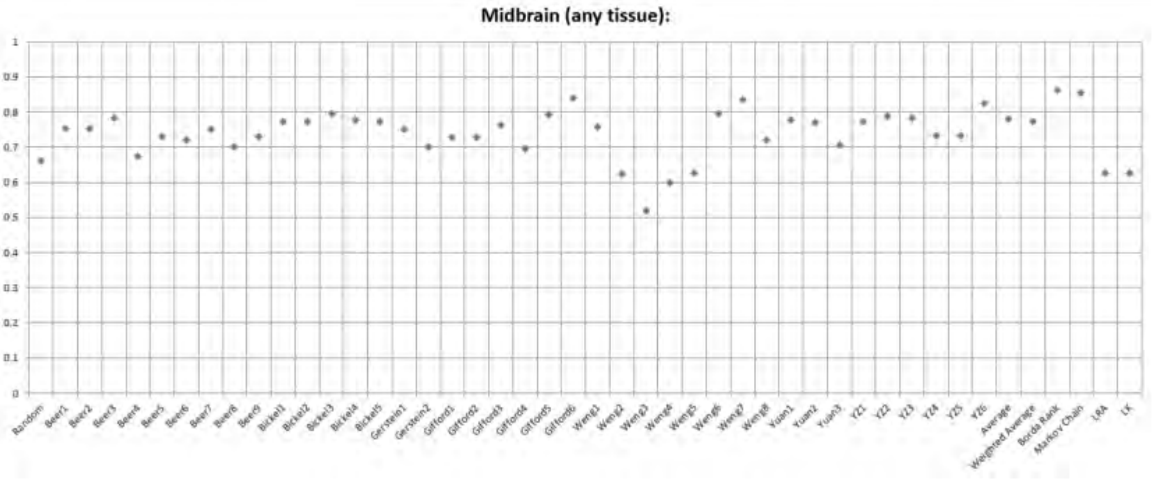
Figure IV-24 | Method Comparison for Predicting Heart Enhancers
Comparison of our 9 methods for predicting heart enhancers in any tissue on training data. Random forest on 138 ENCODE ChIP-seq mouse datasets (just on embryonic days 11.5 and 14.5) performs best.

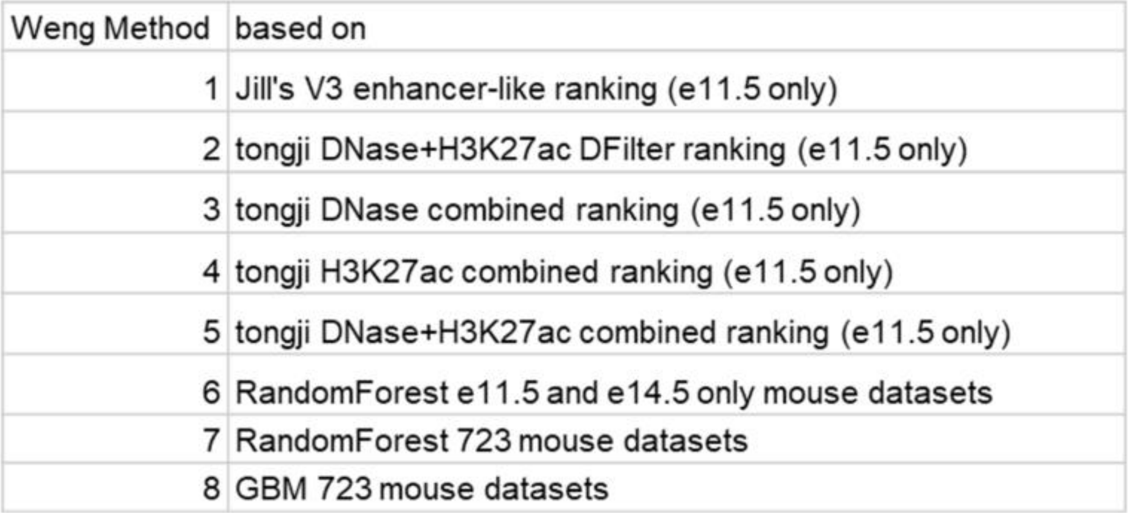




Weng Method	based on
1	Jill's V3 enhancer-like ranking (e11.5 only)
2	tongji DNase+H3K27ac DFilter ranking (e11.5 only)
3	tongji DNase combined ranking (e11.5 only)
4	tongji H3K27ac combined ranking (e11.5 only)
5	tongji DNase+H3K27ac combined ranking (e11.5 only)
6	RandomForest e11.5 and e14.5 only mouse datasets
7	RandomForest 723 mouse datasets
8	GBM 723 mouse datasets

Figure IV-25 | Round 2 Midbrain Prediction Results
Round 2 results for predicting midbrain enhancers in midbrain tissue on embryonic day 11.5 for 150 regions. Gradient boosting with an enlarged set of ENCODE mouse data performed best across all competitors and ensemble models.





Weng Method	based on
1	Jill's V3 enhancer-like ranking (e11.5 only)
2	tongji DNase+H3K27ac DFilter ranking (e11.5 only)
3	tongji DNase combined ranking (e11.5 only)
4	tongji H3K27ac combined ranking (e11.5 only)
5	tongji DNase+H3K27ac combined ranking (e11.5 only)
6	RandomForest e11.5 and e14.5 only mouse datasets
7	RandomForest 723 mouse datasets
8	GBM 723 mouse datasets

Figure IV-26 | Round 2 Midbrain Prediction Results
Round 2 results for predicting midbrain enhancers in any tissue on embryonic day 11.5 for 150 regions. Random forest with an enlarged set of ENCODE mouse data performed best across all competitors and almost equaled the ensemble models.

V. Chapter V: SnoPlowPy: Advanced ENCODE Data Manipulation Tools

V.1 Preface

This research chapter encompasses work performed in conjunction with Henry Pratt, Arjan van der Velde, and Xiao-Ou Zhang, and will be drafted into a bioinformatics applications note.

As I've discussed at the start of this thesis, ENCODE data and metadata are difficult to work with for many reasons. I particularly found the ENCODE metadata to be difficult to work with in bulk: trying to parse metadata for thousands of experiments for Factorbook, ChIP-seq imputation, and SCREEN took hours for operations that should complete in minutes. I started work on my own tool to store and manipulate this metadata; it has grown into quasi-metadata system in its own right, and, if not elegant, it is straightforward to adapt and expand. The tools described here have severed the needs for several users in the lab for processing datasets hundreds of thousands of times.

V.2 Introduction

The ENCODE datasets we utilize as input data to our pipelines and analysis tools are all publically available, versioned, stored in the cloud, and have rich and actively maintained metadata (Davis et al. 2017). Objects are all uniquely identifiable, and cannot be deleted. Ontological information is built into the metadata, as ENCODE has helped develop some of the ontologies. The datasets themselves are of high quality and individually curated, with strict quality-control metrics. File formats are regulated, and tools developed to make sure submitted files adhere to the format their metadata reports. Our analysis products follow similar standards: we submit results back to ENCODE with appropriate metadata, which is reviewed before the products are publically released. Analysis metadata even contains the list of files the product was derived from, as well as the software tools and parameters used, so that the product can be reproduced by others.

While the experiment metadata available from the ENCODE portal is very rich, and provides most of the pertinent metadata information needed per experiment, the JSON data is large, bulky, and difficult to process on a large-scale in a timely fashion. The JSON metadata itself is >3GB of data. Our solution to manipulating these data resulted in our construction of an efficient, performant metadata tool called SnoPlowPy. SnoPlowPy forms the core of our large analysis pipelines, and allows us to walk across core metadata for thousands of ENCODE experiments in a few minutes. A microcosm of tools to support the analysis pipeline and metadata manipulation have also been integrated into SnoPlowPy.

V.3 Tools

V.3.1 Experiment and experiment file metadata objects

The core of SnoPlowPy is an Experiment class (named *Exp*) that wraps and abstracts whatever experiment we are trying to manage (ENCODE, ROADMAP, psychENCODE, etc.). The overview of this class is available in Figure V-1 on page 210. This class can be loaded in multiple ways: directly from on-disk JSON files (updated daily or on-demand); directly from a web request from ENCODE DCC (slower, but with the most recent metadata); or from our metadata web service API (updated weekly, but fastest). The *Exp* class can contains information on what type of assay the experiment is, lab, release date, etc. It also contains Experiment File classes (*ExpFile*) that wrap metadata for particular files for that experiment. These *ExpFile* classes contain both the metadata for that file (MD5 checksum, size in bytes, etc.), but also methods to automatically download and store the file on the local file system in an organized fashion, and has been refactored into its own set of classes (Figure V-2 on page 211 and Figure V-3 on page 212). In addition, I have written *QueryDCC* to directly query the ENCODE DCC metadata system, and merge results into our on-disk metadata store (Figure V-4 on page 213). Overall, these classes are lightweight, and easily manageable; they form the core objects for which the entire Factorbook pipeline is designed. These classes are also flexible enough to be adapted to other datasets; essentially, any experiment that can be represented as an Experiment object with files can be housed in the system.

V.3.2 ENCODE submission system

The ENCODE DAC has become a submitter of original data and analysis products to ENCODE DCC. We have now submitted several TBs of data to DCC, ranging from our own analysis products (such as the cREs from Chapter II: Building and Visualizing an Encyclopedia of ENCODE candidate Regulatory Elements), missing NarrowPeak files for ChIP-seq TF analysis, all the way to entire HiC experiments and a majority of the ROADMAP data. We have developed our own system for doing the data upload, both to simplify the process, but also the handle the wide range of data and metadata we have to submit. We automatically login and authorize to the ENCODE portal with our submitter credentials, then can upload most of the common ENCODE metadata and data objects (including biosamples, documents, files, epigenomes, and cREs). The system also uploads files to the ENCODE DCC AWS bucket, and can automatically resume file transfers if it detects an interruption.

V.3.3 api.wenglab.org

The increasingly large amount of data and metadata (both in bytes and in number of objects) that must be retrieved for projects such as Factorbook, SCREEN, etc. requires more formalized data abstraction layers. These layers can provide unified access interfaces to the data (called APIs), and allow easy programmatic manipulation of the data, far exceeding what can be done using conventional file systems. Centralized location of these webservices also allows the public a way to access the data in a coherent manner. We have migrated our webservices for Factorbook, SCREEN, and metadata to api.wenglab.org, and will be providing specifications in GraphQL language for the public to access the data.

V.3.4 JobRunner and JobMonitor

To better automate creating jobs, we implemented *JobRunner*, which can automatically create the execution files necessary to run 10,000s of Factorbook jobs (see Figure V-5 on page 214). We also implemented *JobMonitor*, to monitor the succeeded and failed jobs. *JobMonitor* can also synchronize the output analysis products from the cluster back to our central store, a necessary task, given the 16 TBs of data produced by the pipeline. In addition, in order to speed-up rerunning of partially-completed jobs, a *Checkpoint* system was developed that permits just to be resume from the last-known good stage of the pipeline. This was especially important when large Factorbook jobs exceeded their allowed run-time on the cluster, a frequent problem encountered while developing jobs pipeline.

V.3.5 Helpers

Finally, I have collected all our common-used Python code into several helper classes. In particular, the *Utils* class (Figure V-7 on page 216) contains code commonly used throughout many of the projects in the Weng lab. Utils has been unit tests for most, if not all, of its methods. In addition, much of the logic needed to parse and manipulate NarrowPeak/Broad Peak/Gapped Peak files has also been refactored and condensed into a helper class *Peaks* (Figure V-6 on page 215).

V.4 Discussion

Managing and manipulating >70TB of locally-stored ENODE data and metadata is a challenge. We've developed several tools to help addresses these problems, while allowing them to be light-weight enough to be adapted to other datasets, and easy modified and extended. Our approach has enabled running hundreds of thousands of jobs on the cluster, even as the metadata at ENCODE changes and evolves. I hope these tools will continue to improve, simplifying use of the vast wealth of ENCODE data.

V.5 Figures

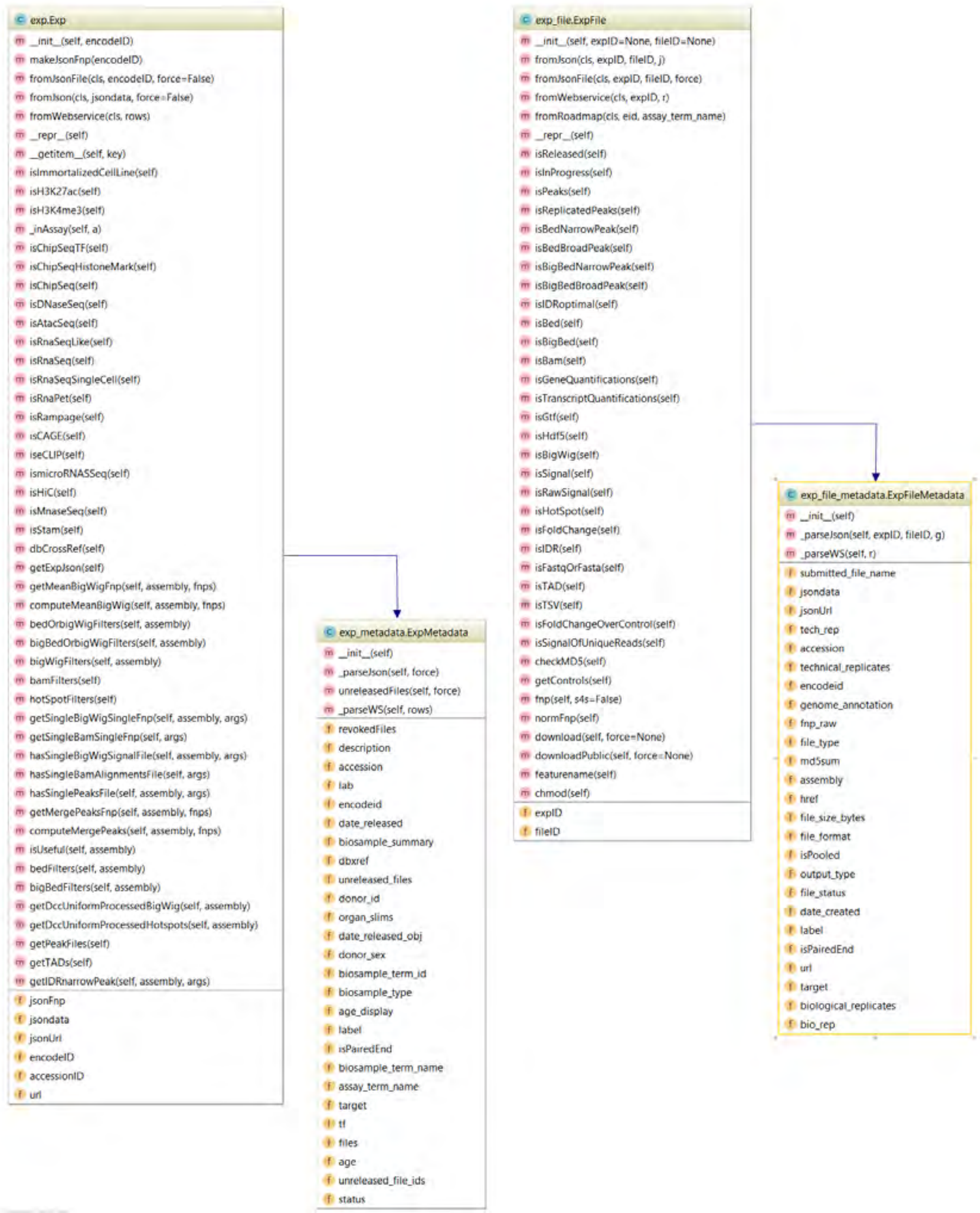


Figure V-1 | Exp and ExpFile Class Diagram
Overview of classes, and the methods and attributes in each class.

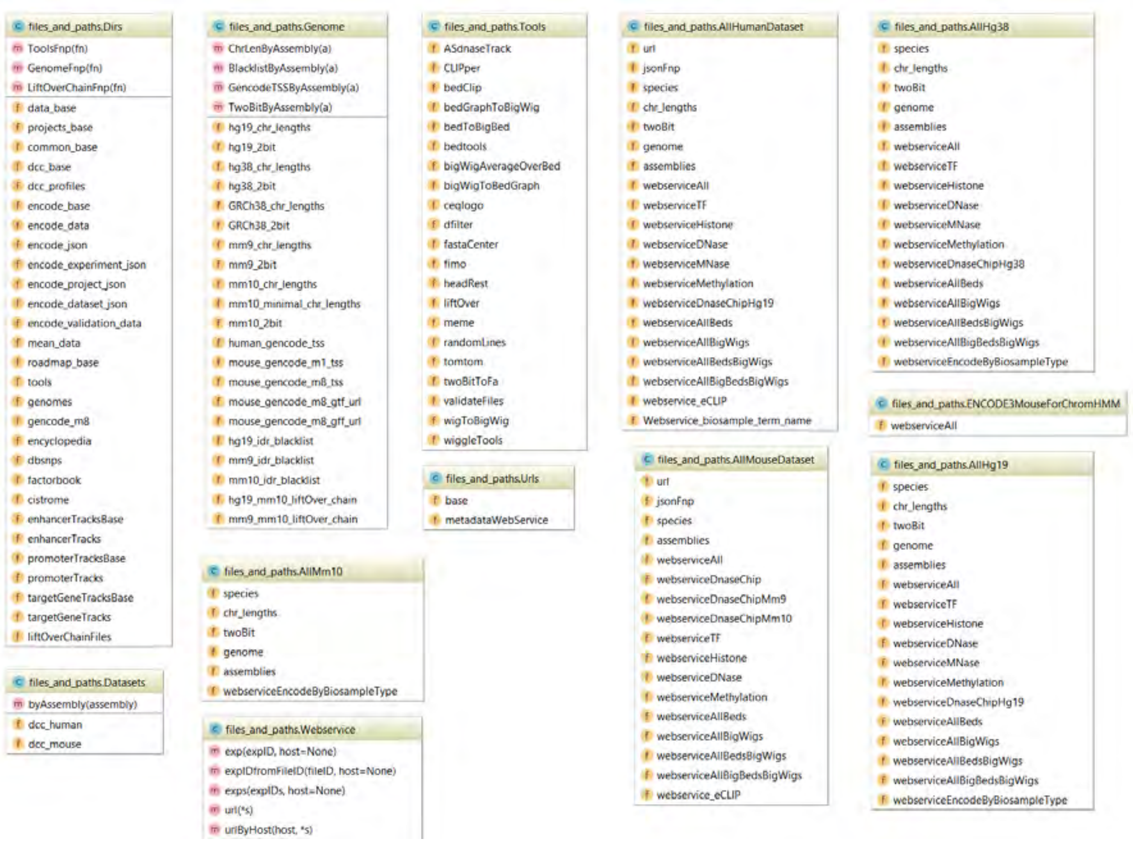
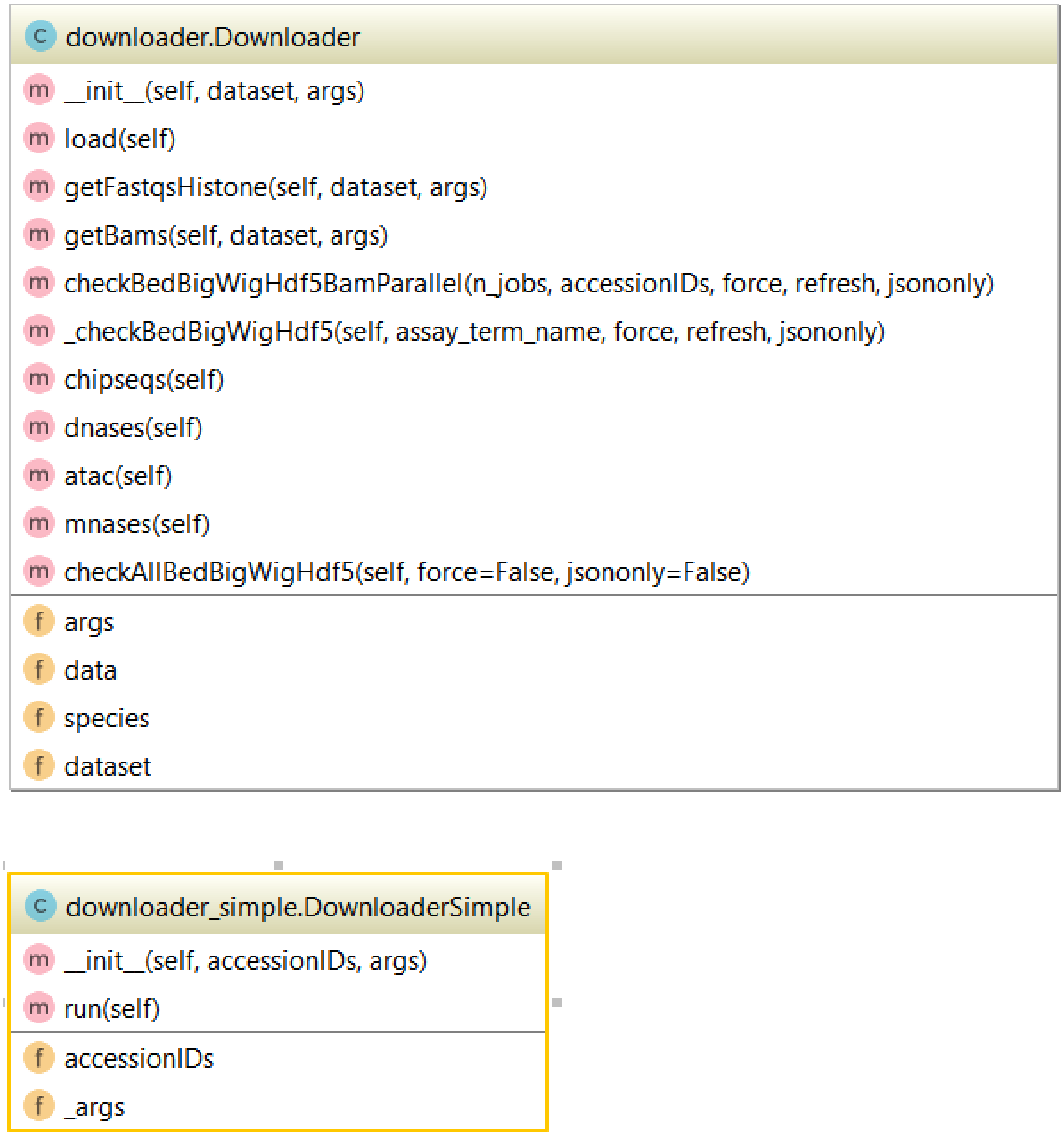


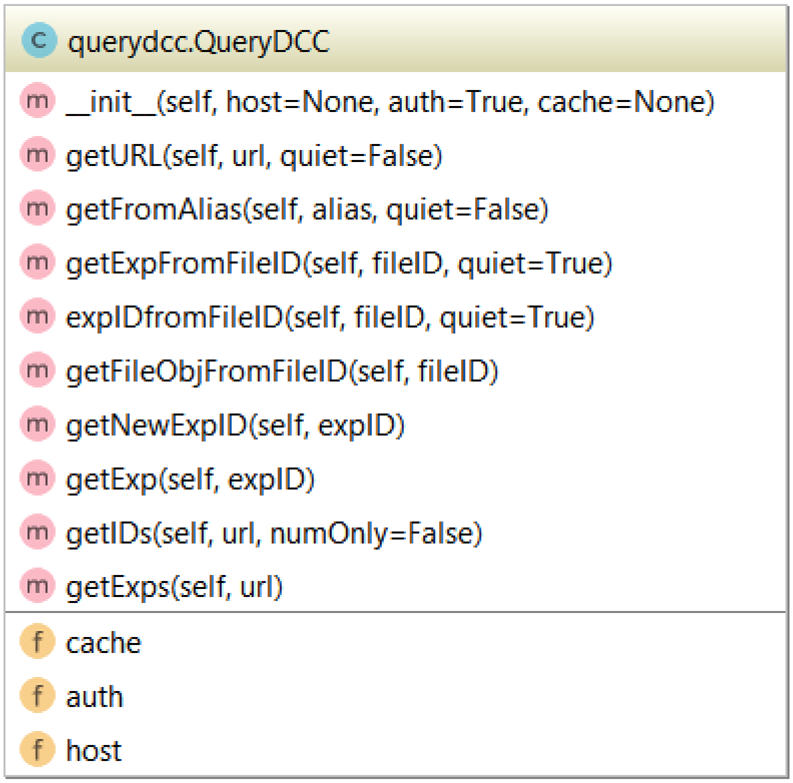
Figure V-2 | File and Paths Class Diagram
Overview of classes, and the methods and attributes in each class.



```
downloader.Downloader
__init__(self, dataset, args)
load(self)
getFastqsHistone(self, dataset, args)
getBams(self, dataset, args)
checkBedBigWigHdf5BamParallel(n_jobs, accessionIDs, force, refresh, jsononly)
_checkBedBigWigHdf5(self, assay_term_name, force, refresh, jsononly)
chipseqs(self)
dnases(self)
atac(self)
mnases(self)
checkAllBedBigWigHdf5(self, force=False, jsononly=False)
args
data
species
dataset
```

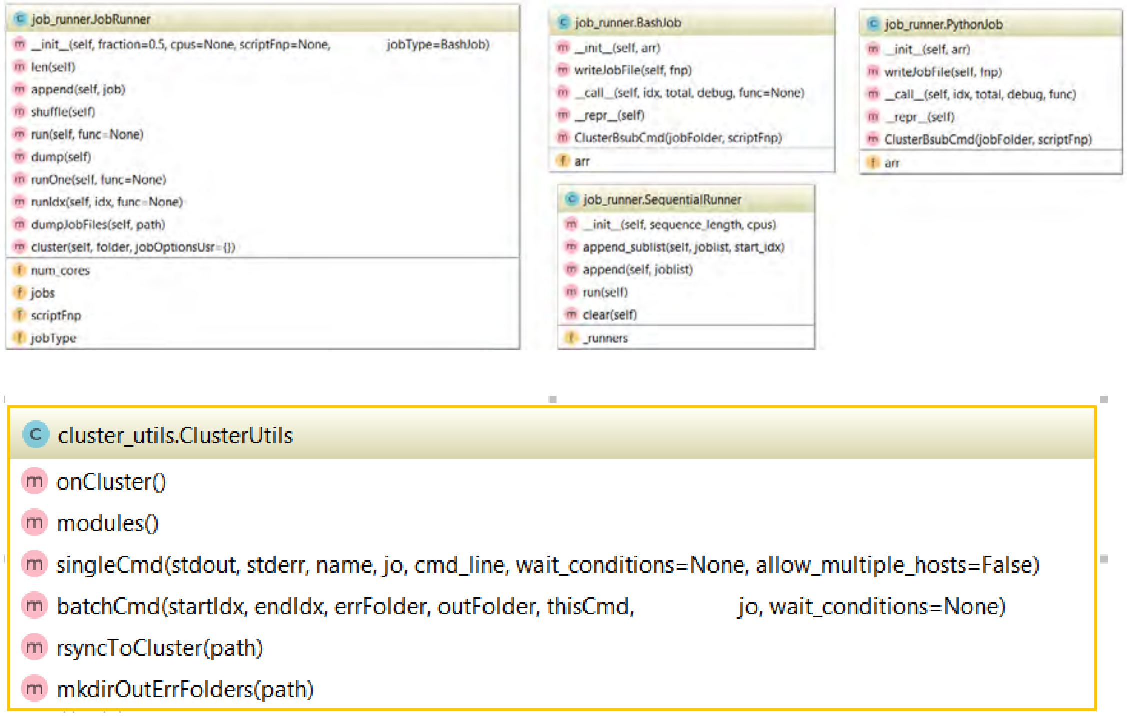
```
downloader_simple.DownloaderSimple
__init__(self, accessionIDs, args)
run(self)
accessionIDs
_args
```

Figure V-3 | Downloader Class Diagram
Overview of classes, and the methods and attributes in each class.



```
c querydcc.QueryDCC
m __init__(self, host=None, auth=True, cache=None)
m getURL(self, url, quiet=False)
m getFromAlias(self, alias, quiet=False)
m getExpFromFileID(self, fileID, quiet=True)
m expIDfromFileID(self, fileID, quiet=True)
m getFileObjFromFileID(self, fileID)
m getNewExpID(self, expID)
m getExp(self, expID)
m getIDs(self, url, numOnly=False)
m getExps(self, url)
f cache
f auth
f host
```

Figure V-4 | QueryDCC Class Diagram
Overview of classes, and the methods and attributes in each class.



```
job_runner.JobRunner
__init__(self, fraction=0.5, cpus=None, scriptFnp=None,          jobType=BashJob)
len(self)
append(self, job)
shuffle(self)
run(self, func=None)
dump(self)
runOne(self, func=None)
runIdx(self, idx, func=None)
dumpJobFiles(self, path)
cluster(self, folder, jobOptionsUsr={})
num_cores
jobs
scriptFnp
jobType
```

```
job_runner.BashJob
__init__(self, arr)
writeJobFile(self, fnp)
__call__(self, idx, total, debug, func=None)
__repr__(self)
ClusterBsubCmd(jobFolder, scriptFnp)
arr
```

```
job_runner.PythonJob
__init__(self, arr)
writeJobFile(self, fnp)
__call__(self, idx, total, debug, func)
__repr__(self)
ClusterBsubCmd(jobFolder, scriptFnp)
arr
```

```
job_runner.SequentialRunner
__init__(self, sequence_length, cpus)
append_sublist(self, joblist, start_idx)
append(self, joblist)
run(self)
clear(self)
_runners
```

```
cluster_utils.ClusterUtils
onCluster()
modules()
singleCmd(stdout, stderr, name, jo, cmd_line, wait_conditions=None, allow_multiple_hosts=False)
batchCmd(startIdx, endIdx, errFolder, outFolder, thisCmd,          jo, wait_conditions=None)
rsyncToCluster(path)
mkdirOutErrFolders(path)
```

Figure V-5 | Job Runner Class Diagrams
Overview of classes, and the methods and attributes in each class.

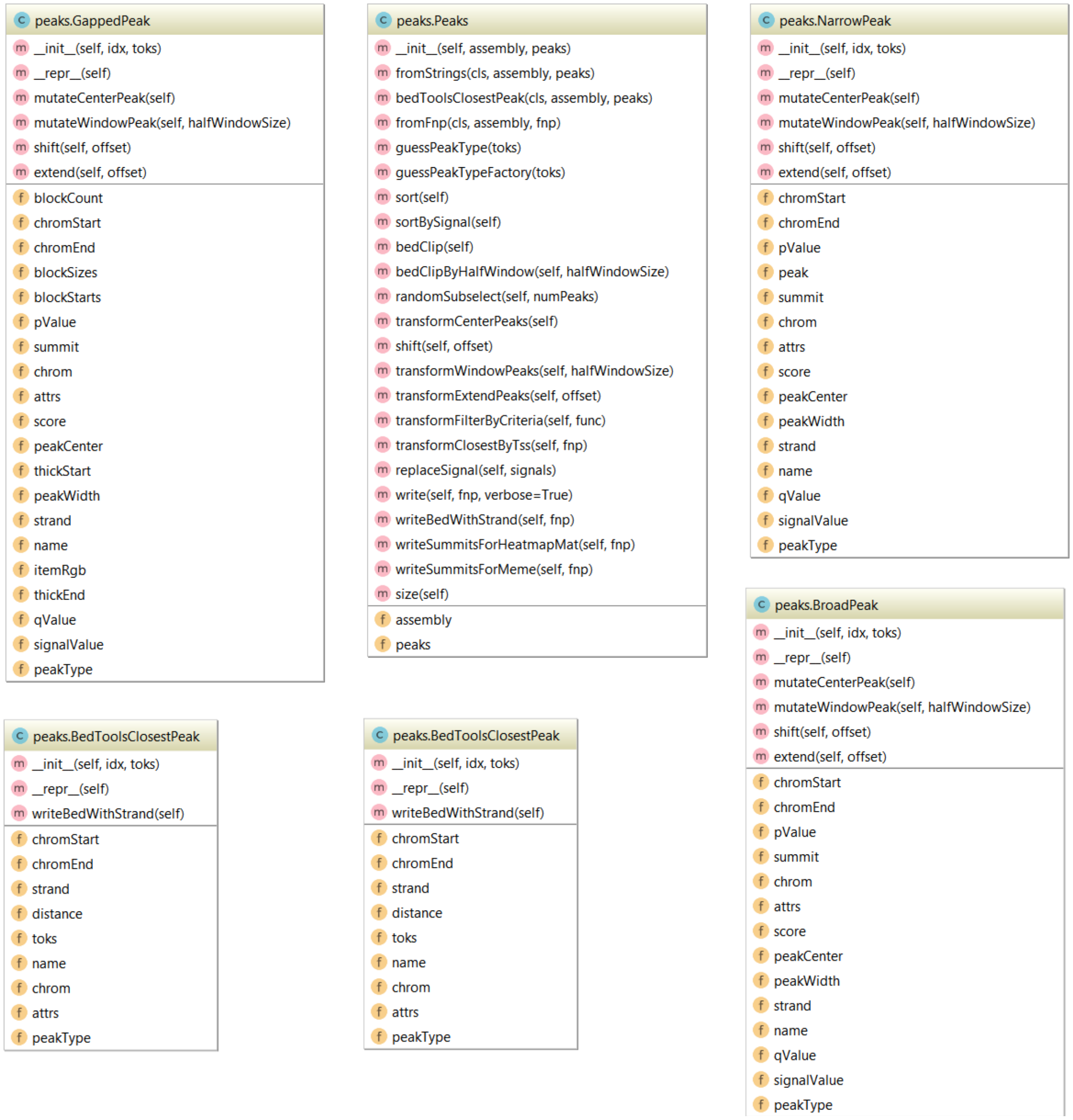


Figure V-6 | Peaks Class Diagram

Overview of classes, and the methods and attributes in each class.

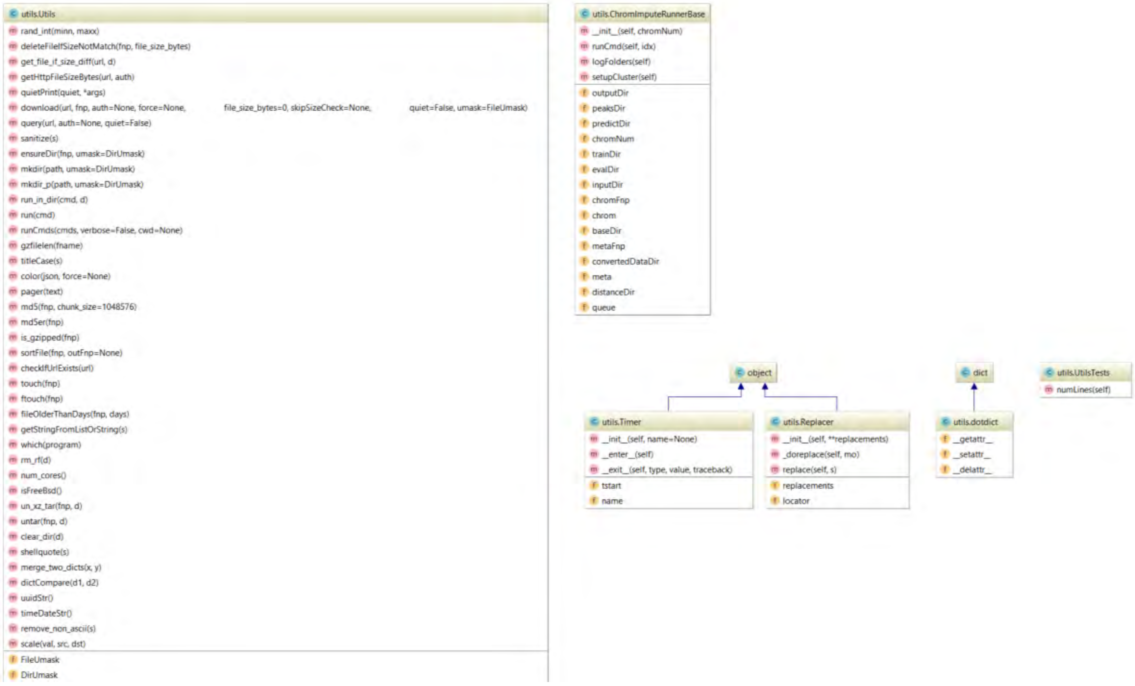


Figure V-7 | Utils Class Diagram
Overview of classes, and the methods and attributes in each class.

VI. Chapter VI: Discussion

VI.1 Preface

This discussion is based on my discussions in chapters II, III, IV, and V.

VI.2 Introduction

The work in this thesis describes efforts to better understand the epigenome through creation of an Encyclopedia of candidate Regulatory Elements and a visualizer (SCREEN) for these elements. We have also described Factorbook, another visualization tool to help understand aggregated histone mark and transcription factor occupancy patterns around TF peaks. We also described our work on computationally predicting transcription factor binding using supervised machine learning methods. Lastly, we described a set of programming tools we have implemented to aid in the large scale processing needed for the above work.

VI.3 ENCODE Encyclopedia and SCREEN

With ~90% of disease-associated Single Nucleotide Polymorphisms (SNPs) occurring to be in intronic or intergenic regions (Hindorff et al. 2009), the systematic location and study of genetic variants outside of protein-coding regions is critical to better understanding and potentially treating disease pathology. There are millions of these potentially functional regions in the genome, working as enhancer, promoters, repressors, or insulators; having a catalog of these elements, and a way to easily interrogate these regions, will be essential to keep researchers afloat in an ocean of data. Researchers require tools to aid in both biological questions and practical bioinformatics problems;

users require an encyclopedia synthesizing the low-level data into a more manageable product that can be analyzed and effectively investigated.

We have identified the first real catalog of putative regulatory regions, locating nearly 2 millions cREs across human and mouse genomes in regions with open chromatin and enhancer-like or promoter-like signatures (based on histone modification marks and other genomic distance information). These elements are numbered and versioned, permitting direct reference in future papers. cREs are anchored on DNase-seq representative DHSs (rDHSs), condensed first from >30 million DHS sites across more than 400 individual samples into a set of non-overlapping regions number ~1.3 million in human and ~400 thousand in mice.

The tools we have developed have enabled an integrated approach to discovering new biological insights. For example, as we discussed in Chapter 2, the *Ogn* mouse gene that encodes osteoglycin was known to increase expression on mouse embryonic day 12, in concordance with increased bone formation (Taher et al. 2011). Not only do we demonstrate this differential change on expression using our SCREEN Differential Gene Expression tool in Figure II-27a on page 113, as well as on the UCSC Genome Browser in Figure II-28 on page 114, but we also find putative regulatory elements whose own expression levels change in step with the *Ogn* gene expression increases. The human homolog of this gene—OGN—has been correlated with left ventricular hypertrophy (Petretto et al. 2008), and may account for hypertrophic responses to hypertension or aortic stenosis. Future studies could utilize our candidate Regulatory Elements in mouse models to pursue several experimental approaches (including mouse transgenic assay or

CRISPR/Cas9 targeted gene editing) to definitely define the elements regulating this gene, and start on the path to therapeutic control.

We have also explored developing new hypotheses for epigenetic malfunctions in autoimmune and immune-mediated diseases like inflammatory bowel disease. We cross-referenced our curated set of GWAS SNP studies with our cRE database to determine which biosamples are most enriched in active cREs overlapping SNPs. We hypostasized that we could locate a putative promoter involved in autoimmune dysregulation in inflammatory bowel disease. For the top 10 most active biosamples for a IDB study, we found that 9 were in leukocyte biosamples, and the tenth biosample was from rectal mucosa. We pursued the top-ranked leukocyte biosample, and then located a putative promoter region within 2kb of a TSS for *LSP1*, an immune-related gene. We demonstrated that *LSP1* was immune-related, as well as active in immune-related tissues, based upon ENCODE RNA-seq gene expression and RAMPAGE TSS data. We then cross-reference our putative promoter with Ensembl, and found the region overlapped a SNP whose A allele was already correlated with *LSP1* isoform expression changes. To enable possible future mouse model studies, we also located an orthologous region in mouse, and verified putatively regulated the promoter in the first exon of *Lsp1*, also an immune-related gene.

The registry of Elements is a powerful new research paradigm, but there are a number of limitations to our approach. Of the hundreds of biosamples used in the Encyclopedia, less than two dozen actually had all 4 of our core assays (DNase-seq, H3K4me3, H3K27ac, and CTCF). This sparsity of core data complicates cross-biosample comparison of elements, and, obviously, limits the possibility of looking for, say enhancer-like signature cREs in biosamples without any histone experimental data. Our z-score based ranking system is practical but overly simplistic; it deals with the many problems of data normalization across experiments, but leaves a great deal of room for more mathematically powerful future models. There are many subtle complications posed by the processed ENCODE data: can ChIP-seq experiments from different labs, using different approaches to controls, be integrated together, or even comparable? Many older, core experiments performed during early years of ENCODE have (by today's standards) very low read counts; at what point do they get replaced by newer, more deeply sequenced experiments? The ENCODE pipelines are composed of a large number of different software packages and "glue" code; how do we verify that the final analysis products are valid and "correct," especially if only one analysis pipeline exists?

During the primordial creation of the Encyclopedia, the epigenome did not fail to remind us of its innate complexity. As we have shown, the classical picture of enhancers and promoters having particular histone modification signatures is only a rough first approximation: it is well established. for instance, that regions with H3K4me3 marks can indicate active enhancers (Pekowska et al. 2011). The current classification scheme of enhancer, promoter, silencer, etc., is overly simplified, and, in many cases, doesn't reflect the more nuanced, complex reality of the genome. Future versions of the Encyclopedia have great potential to bring a more nuanced classification system to life, and more finely elucidate exactly how, and why, each regulatory element works.

The rapid pace of advancements in sequencing technology has exceeded the technology speed increases Moore's Law predicts (Hayden 2014), with the cost of sequencing a human genome falling from >$100 million (International Human Genome Sequencing 2001) to less than $10,000. Sequencing a human genome for $1,000 has gone from the realm of science fiction to near inevitability (Hayden 2014). At this rate, whole exome, if not whole genome sequencing, of patients for medical purposes is inevitable. The amount of genomic information to be collected will be a vast ocean of data. In light of this, the Encyclopedia's utility in selecting a stable set of cREs to—at least—start filtering and distilling this ocean of data is especially important. While future datasets will no doubt greatly expand the Encyclopedia beyond its current scope, our selection of the top 5% of regions to begin with (as demonstrated by saturation analysis) should prove to be an important subset of functional regions for a long time. It is inevitable that the Encyclopedia will need to incorporate weaker putative regulatory regions in some manner—if even to provide a way of annotating regions for future research.

To interrogate the cREs we selected in this version of the Encyclopedia, we needed a tool with several requirements, including dynamic and interactive interfaces that allow users to search and filter cREs in real time, as well as display interactive plots of cREs. The tool needed to integrate and view all the low-level data in some sort of genome browser. The interactive plots also needed to be reproducible, and near-publication quality. Finally, we also needed a mechanism to facilitate users' generation of these figures with custom data. The tool needed to be extensible, providing for the inclusion of analyses involving external annotations and user-submitted annotations.

To fulfill these requirements, we developed SCREEN, the visualizer for cREs. This tool is maturing into an integrated approach to examining cREs in the context of the genome and epigenome. We already a multi-part visualization platform, with mechanisms to search and investigate cREs, show gene expression information, and explore GWAS studies for SNP overlap with cREs. SCREEN is the start of central repository for accessing information on functional regions of human and mouse genomes, integrating cRE searching, sorting, and visualization.

The genomic visualization field, though, is just in its infancy, with many hurdles to overcome in data scale, complexity of information, and visualization techniques. SCREEN is most definitely still in its infancy, with many improvements to be made; hopefully, though, SCREEN will become an increasingly powerful platform of epigenome discovery, especially with the experiments being performed in the ENCODE4.

VI.4 Factorbook

As demonstrated by the chapter, Factorbook is a web-based analysis tool that integrates all public ENCODE ChIP-seq TF data in a peak-centric manner. Factorbook enjoys use by researchers within the ENCODE community and beyond. Factorbook has become a canonical resource for TF motif information; it has been cited by more than 127 publications[5], its motif tracks have been integrated into the UCSC Genome Browser (Rosenbloom et al. 2015), and it is even being mined for machine learning competitions (Keilwagen, Posch, and Grau 2017). Factorbook's centralization of ENCODE TF motif

[5] https://scholar.google.com/scholar?cites=16586749045503397316&as_sdt=40000005&sciodt=0,22&hl=en

information has led to many biological insights: for example, Factorbook motifs assisted the search for SNPs that could disrupt TF binding in chronic lymphocytic leukemia (Law et al. 2017). Likewise, gene regulatory networks have been built incorporating RNA-seq data with Factorbook motifs to better elucidate breast cancer-related TF networks (Janky et al. 2014). Gene expression data visualization was first incorporated into Factorbook because of the known association between genes and increased TF occupancy in the local genomic neighborhood (Wang et al. 2012).

We will continue expanding Factorbook as more ChIP-seq experiments become available at ENCODE during its 4th phase. This expansion includes not only incorporating new raw data to improve the quality and utility of annotations, but also continuing to develop novel analyses and visualizations. Inspired by the representational DHS sites used to initially select cREs, we will be creating representation Transcription Factor Binding Sites (rTDBS) from ChIP-seq TF data to help simplify and anchor the set of motif sites in the genome for each TF. In many ways, the future of Factorbook and SCREEN are intermixed: users will wish to know, for instance, what TFs overlap her enhancer-like signature cRE of interest, or what TFBSs overlap a particular SNP in a given cRE. Ultimately, we will combine SCREEN and Factorbook ("SCREENbook") for this deep integration; time will tell. Another important future development of Factorbook is the integration of a genome browser with the Factorbook motifs, aggregation plots, signal tracks, and cREs; this unified view might greatly aid users, and expand Factorbook's usefulness and utility. Additional crowdsourcing tools must be further developed to help centralize discussion on TF ChIP-seq motifs.

VI.5 Machine Learning Epigenomic Data

Machine learning of epigenetic data is on the precipice of heralding in a new world of hybrid experimental and computational techniques; while experiment will always be gold standard for data, predictive computational techniques can help bridge the gap when experiments are too expensive, technically difficult, or numerous. Imputation of experimental data is also one possible source of quality control for when the actual experiment is performed (Ebert and Bock 2015). Imputation of transcription factor binding sites may shed light on more intricate co-binding and tethered binding patterns: for instance, we purposefully eliminated CTCF as a feature when predicting members of the complex SMC3 or RAD21 (and vice-versa), since inclusion of these known co-binding partners made the models artificially accurate (Holwerda and de Laat 2013).

We demonstrated viability in predicting transcription factor binding sites using supervised machine learning methods. For some TFs (like CTCF), similarity of TF binding across cell types greatly improved performance of TF binding prediction. For TFs with highly variable, biosample specific binding, though, getting good performance from the learned models was difficult. While good performance can be obtained from standard machine learning techniques like logistic regression, gradient boosting, and random forest, it seems inevitable that deep learning will supersede these approaches. The inherent complexity of the epigenome intuitively dictates this—only computational models that can deal with this complexity seem likely to ultimately succeed. The intersection of experimental mapping with imputation appears inevitable. The combined power of both approaches may be maximally leveraged by performing mapping for important, experimentally viable biosamples, while using imputation for biosamples experimentally more difficult to assay (such as primary cancer samples) (Ebert and Bock 2015).

Machine of learning of enhancers in different tissues and developmental time points poses a different set of challenges. Some tissues undergo significantly less developmental dynamics at the developmental time points we were predicting in. This developmental stability lead to better predictive models, since the state of enhancers was not changing as drastically between time points. The relatively small number of samples from VISTA, though, has made machine learning enhancers difficult, as does the intrinsic bias when selecting regions to test (since regions have been selected by high conservation or high DNase and H3K27ac signal, thereby skewing results towards models using those features).

VI.6 SnoPlowPy

Managing and manipulating >70TB of locally-stored ENODE data and metadata is a challenge. We've developed several tools to help addresses these problems, while allowing them to be light-weight enough to be adapted to other datasets, and easy modified and extended. The core purpose of SnoPlowPy—to support our research pipelines and ease use of ENCODE metadata—has been proven in the hundreds of thousands of cluster jobs it has helped organize and run. We are planning expansion of the system to handle Cistrome and psychENCODE metadata, as well as support other labs and projects at UMassMed. The API portion of SnoPlowPy will also house a future central warehouse for all "peak" files (Narrow/Broad/Gapped peaks, DHSs, rDHSs, cREs, etc.), and will allow coordinate-based searching. This simplifies some of our analyses which require large number of peak file intersection. As mentioned above, the API site will also handle the effective merge of "SCREENbook", providing one common API for data for web/tablet/desktop versions of SCREEN/Factorbook. A major task we have not yet performed is to make SnoPlowPy available to the general public, with better documentation and an automated way of installing locally for the user. We also pursuing development of R and C++ wrappers, so users have a unified way of accessing the API regardless of programming language.

VI.7 Conclusion

We have utilized the tools developed in this thesis to begin to build new hypotheses for which elements may be involved in functional regulation of genes; in particular, we have shown ways of finding putative regulators for diseases such as left ventricular hypertrophy and inflammatory bowel disease that, once experimentally validated and targeted through pharmacology or gene editing, could translate into new therapeutic treatments at the bedside.

ENCODE3 has greatly increased the number and variety of experiments available from the ENCODE portal. These new data are allowing a much richer, expansive analysis of the epigenome. There is finally sufficient data to undertake create of an Encyclopedia of candidate Regulatory Elements. While still in its early age, this Registry of cREs is already proving itself to be a useful contribution to the scientific community. For the first time in ENCODE, we are accessioning putative functional regions of the genome, with the hope that papers in the future can directly reference these cREs. We have found a wealth of putative regulatory regions that correlate with experimental datasets within and outside of ENCODE. These regions greatly simplify the search for putative regulatory regions genes or SNPs of interest.

SCREEN, the visualizer for cREs, is maturing into a tool providing an integrated approach to examining cREs in the context of the genome and epigenome. SCREEN is already a multi-part visualization platform, with one app providing cRE search and investigation, another app providing gene expression display, and a third allow GWAS studies to be interrogated for SNP overlap with cREs. Factorbook, a peak-centric visualizer of transcription factor data, has helped centralize TF motif information, TF/TF interactions, and TF/histone interactions, providing another vehicle for developing biological insights. Our machine learning approaches are starting to help shed light on patterns in TF binding, as well as the set of epigenetic markers that truly distinguish enhancer regions.

This thesis has sought to illuminate several new insights into how the epigenome works, and how we can apply computational techniques to better elucidate and visualize functional regions of the epigenome. We've also developed a number of new tools to aid in pipeline development for these projects. In the context of today, we hope we've enabled researchers with new ways of interrogating and investigating the epigenome for both basic research purposes, and for clinical applications for the good of humanity. Decades from now, I hope this work is recognized as a useful—if primitive—stepping stone, rendered obsolete and primordial compared to the scientific discoveries ahead of us.

VII. Appendix A: Encyclopedia V2

VII.1 Preface

This appendix is based off work I performed with Zhiping Weng for version 2 of the Encyclopedia; the methods and tracks are available through ENCODE[6]. This was the first version of the Encyclopedia to be publically disseminated and presented at conferences by myself and Zhiping.

VII.2 Introduction

Annotations made for version 2 of the Encyclopedia expanded Sowmya Iyer's work (Iyer 2015) on version 1. This version directly annotated hg19 (mm10 was drafted[7] but not released) with candidate promoters and enhancers through integration of DNase-seq, histone mark ChIP-seq, and transcription factor (TF) ChIP-seq datasets. In total 177 ENCODE2 and ROADMAP cell types are annotated in this release; among them 94 cell types have both DNase-seq data and ChIP-seq data for one or more of the histone marks H3K27ac, H3K4me1, H3K4me3, H3K9ac (see Table VII-1 on page 232). For each of these 94 cell types, we annotated the DNase peaks with the percentile of each histone mark signal in the matching cell type.

VII.3 Methods

The Stamatoyannopoulos (Stam) Lab (University of Washington) merged all DNase peak data from the Stam and Crawford (Duke University) labs. This merging process formed one combined DNase-seq dataset with non-overlapping DNase hypersensitive regions.

[6] https://www.encodeproject.org/data/annotations/v2/

[7] https://zlab.umassmed.edu/~purcarom/bib5/mouse_pedia/beta1/ucsc_trackhub.txt

The Stam lab then identified the “master” peak in each region, defined as the peak in the region with highest peak height. I then separated the master DNase peaks into TSS proximal and TSS distal groups based on whether or not they intersected a 2000bp window centered on any GENCODE TSS.

I downloaded signal files from Roadmap and ENCODE (using a primordial version of SnowPlowPy). For each DNase master peak, the average histone signals in the matching cell type was calculated in a 1000bp window around the center of the peak. This signal was converted to a percentile using the background distribution of histone signal in the matching cell type in randomly chosen 1000bp genomic regions that were outside all DNase peaks and ENCODE blacklisted regions. DNase master peaks that have at least one cell type with histone signal > 95th percentile of background were retained. Likewise, ChIP-seq TF files were also downloaded from ENCODE project. For each of the distal and proximal DNase master peaks, overlapping TF ChIP-seq peaks across all cell types available were identified. The TF peaks with maximum score in each master DNase peak were retained. All resulting peaks were transformed into bigBed tracks, and a trackhub generated for the UCSC Genome Browser (Figure VII-7 on page 239) and WashU Genome Browser (Figure VII-8 on page 240).

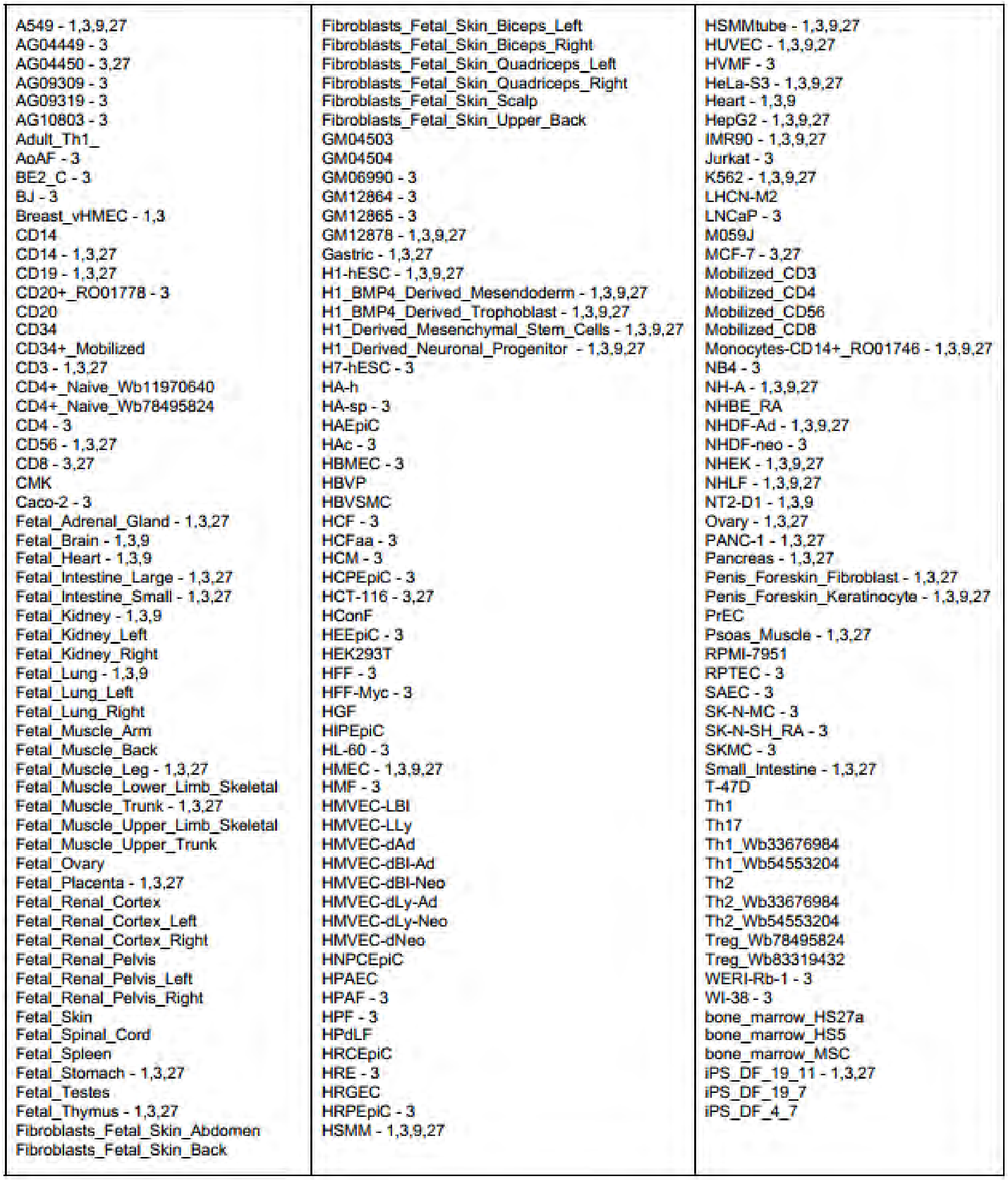
VII.4 Results

My pipeline located 3,166,489 candidate enhancer and promoter regions in hg19, and 1,200,491 candidate enhancer and promoter regions in mm10. To gauge how well the DNase master peaks intersected regions with enhancer-like signatures (i.e. high H3K27ac signal, or called peaks from the ENCODE pipeline), a multitude of ‘bedtools intersect’ operations were performed. In general, as shown in Figure VII-2 on page 234 for human and in Figure VII-5 on page 237 for mouse, H3K27ac peaks intersected DNase master peaks more than 75% of the time. As a sanity check, DNase master peaks were also intersected with H3K27ac peaks; since DHSs indicate genomic areas of open chromatin, where many other post-transcription modifications can occur besides acetylation of H3K27, less than half of the DNase master peaks overlapped with our enhancer histone modification mark, as shown in Figure VII-2 on page 234 and Figure VII-4 on page 236. I also examined master peak distance distributions across biosamples in mm10. As shown in Figure VII-3 on page 235, distances to nearest master peak for different histone marks did vary. Ideally, master peaks should have relatively similar peak distributions for the same histone marks across different biosamples. Very short distances, though, may indicate large number of weak peaks, which was a problem. These DNase master peak "clouds" (Figure VII-9 on page 241) indicated too many master peaks with weak signal were being selected; these weaker peaks increase the number false positive enhancer-like elements found. Lastly, I examined how well master peaks made from just individual biosamples overlapped H3K27ac; as show in Figure VII-6 on page 238, biosample-specific master peaks generally intersected the H3K27ac peaks for the particular biosample well, with some exceptions.

Version 2 of the Encyclopedia was the first to incorporate Roadmap data, and we started exploring the techniques and problems associated with building what would be known as the registry of Candidate Elements. This was the Encyclopedia first presented at several conferences; feedback (in particular, the master peak "cloud" problem, and difficulties with finding elements in the UCSC Genome Browser) encouraged our development of Version 3 of the Encyclopedia.

VII.5 Tables

numbers after cell type: 1 = has H3K4me1 datasets 3 = has H3K4me3 datasets
9 = has H3K9ac datasets 27 = has H3K27ac datasets



A549 - 1,3,9,27
AG04449 - 3
AG04450 - 3,27
AG09309 - 3
AG09319 - 3
AG10803 - 3
Adult_Th1_
AoAF - 3
BE2_C - 3
BJ - 3
Breast_vHMEC - 1,3
CD14
CD14 - 1,3,27
CD19 - 1,3,27
CD20+_RO01778 - 3
CD20
CD34
CD34+_Mobilized
CD3 - 1,3,27
CD4+_Naive_Wb11970640
CD4+_Naive_Wb78495824
CD4 - 3
CD56 - 1,3,27
CD8 - 3,27
CMK
Caco-2 - 3
Fetal_Adrenal_Gland - 1,3,27
Fetal_Brain - 1,3,9
Fetal_Heart - 1,3,9
Fetal_Intestine_Large - 1,3,27
Fetal_Intestine_Small - 1,3,27
Fetal_Kidney - 1,3,9
Fetal_Kidney_Left
Fetal_Kidney_Right
Fetal_Lung - 1,3,9
Fetal_Lung_Left
Fetal_Lung_Right
Fetal_Muscle_Arm
Fetal_Muscle_Back
Fetal_Muscle_Leg - 1,3,27
Fetal_Muscle_Lower_Limb_Skeletal
Fetal_Muscle_Trunk - 1,3,27
Fetal_Muscle_Upper_Limb_Skeletal
Fetal_Muscle_Upper_Trunk
Fetal_Ovary
Fetal_Placenta - 1,3,27
Fetal_Renal_Cortex
Fetal_Renal_Cortex_Left
Fetal_Renal_Cortex_Right
Fetal_Renal_Pelvis
Fetal_Renal_Pelvis_Left
Fetal_Renal_Pelvis_Right
Fetal_Skin
Fetal_Spinal_Cord
Fetal_Spleen
Fetal_Stomach - 1,3,27
Fetal_Testes
Fetal_Thymus - 1,3,27
Fibroblasts_Fetal_Skin_Abdomen
Fibroblasts_Fetal_Skin_Back
Fibroblasts_Fetal_Skin_Biceps_Left
Fibroblasts_Fetal_Skin_Biceps_Right
Fibroblasts_Fetal_Skin_Quadriceps_Left
Fibroblasts_Fetal_Skin_Quadriceps_Right
Fibroblasts_Fetal_Skin_Scalp
Fibroblasts_Fetal_Skin_Upper_Back
GM04503
GM04504
GM06990 - 3
GM12864 - 3
GM12865 - 3
GM12878 - 1,3,9,27
Gastric - 1,3,27
H1-hESC - 1,3,9,27
H1_BMP4_Derived_Mesendoderm - 1,3,9,27
H1_BMP4_Derived_Trophoblast - 1,3,9,27
H1_Derived_Mesenchymal_Stem_Cells - 1,3,9,27
H1_Derived_Neuronal_Progenitor - 1,3,9,27
H7-hESC - 3
HA-h
HA-sp - 3
HAEpiC
HAc - 3
HBMEC - 3
HBVP
HBVSMC
HCF - 3
HCFaa - 3
HCM - 3
HCPEpiC - 3
HCT-116 - 3,27
HConF
HEEpiC - 3
HEK293T
HFF - 3
HFF-Myc - 3
HGF
HIPEpiC
HL-60 - 3
HMEC - 1,3,9,27
HMF - 3
HMVEC-LBl
HMVEC-LLy
HMVEC-dAd
HMVEC-dBl-Ad
HMVEC-dBl-Neo
HMVEC-dLy-Ad
HMVEC-dLy-Neo
HMVEC-dNeo
HNPCEpiC
HPAEC
HPAF - 3
HPF - 3
HPdLF
HRCEpiC
HRE - 3
HRGEC
HRPEpiC - 3
HSMM - 1,3,9,27
HSMMtube - 1,3,9,27
HUVEC - 1,3,9,27
HVMF - 3
HeLa-S3 - 1,3,9,27
Heart - 1,3,9
HepG2 - 1,3,9,27
IMR90 - 1,3,9,27
Jurkat - 3
K562 - 1,3,9,27
LHCN-M2
LNCaP - 3
M059J
MCF-7 - 3,27
Mobilized_CD3
Mobilized_CD4
Mobilized_CD56
Mobilized_CD8
Monocytes-CD14+_RO01746 - 1,3,9,27
NB4 - 3
NH-A - 1,3,9,27
NHBE_RA
NHDF-Ad - 1,3,9,27
NHDF-neo - 3
NHEK - 1,3,9,27
NHLF - 1,3,9,27
NT2-D1 - 1,3,9
Ovary - 1,3,27
PANC-1 - 1,3,27
Pancreas - 1,3,27
Penis_Foreskin_Fibroblast - 1,3,27
Penis_Foreskin_Keratinocyte - 1,3,9,27
PrEC
Psoas_Muscle - 1,3,27
RPMI-7951
RPTEC - 3
SAEC - 3
SK-N-MC - 3
SK-N-SH_RA - 3
SKMC - 3
Small_Intestine - 1,3,27
T-47D
Th1
Th17
Th1_Wb33676984
Th1_Wb54553204
Th2
Th2_Wb33676984
Th2_Wb54553204
Treg_Wb78495824
Treg_Wb83319432
WERI-Rb-1 - 3
WI-38 - 3
bone_marrow_HS27a
bone_marrow_HS5
bone_marrow_MSC
iPS_DF_19_11 - 1,3,27
iPS_DF_19_7
iPS_DF_4_7

Table VII-1 | Datasets Used
List of biosamples and assays used for V2 of hg19 Encyclopedia

VII.6 Figures

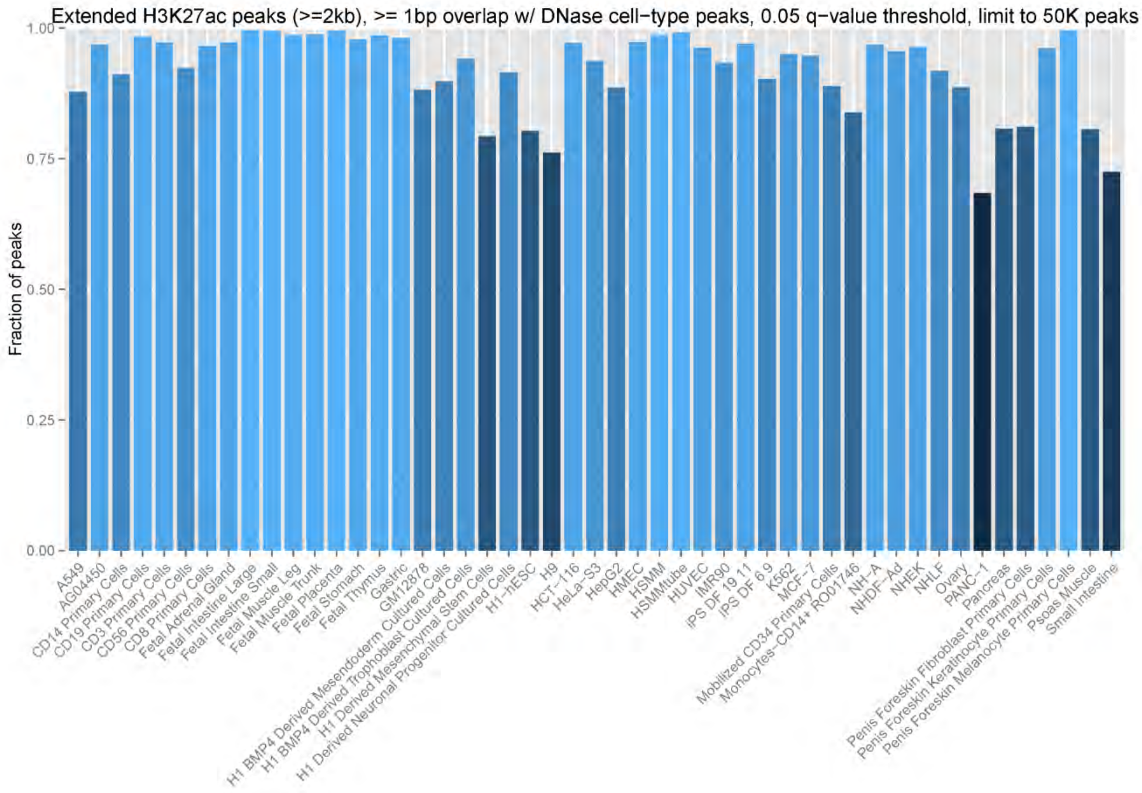


Figure VII-1 | Fraction of H3K27ac peaks that overlap DNase peaks (hg19)
DNase-seq "master peaks" overlap most (if not all) H3K27ac peaks, indicating we're identifying the majority of sites with enhancer-like epigenetic signals based on H3K27ac.

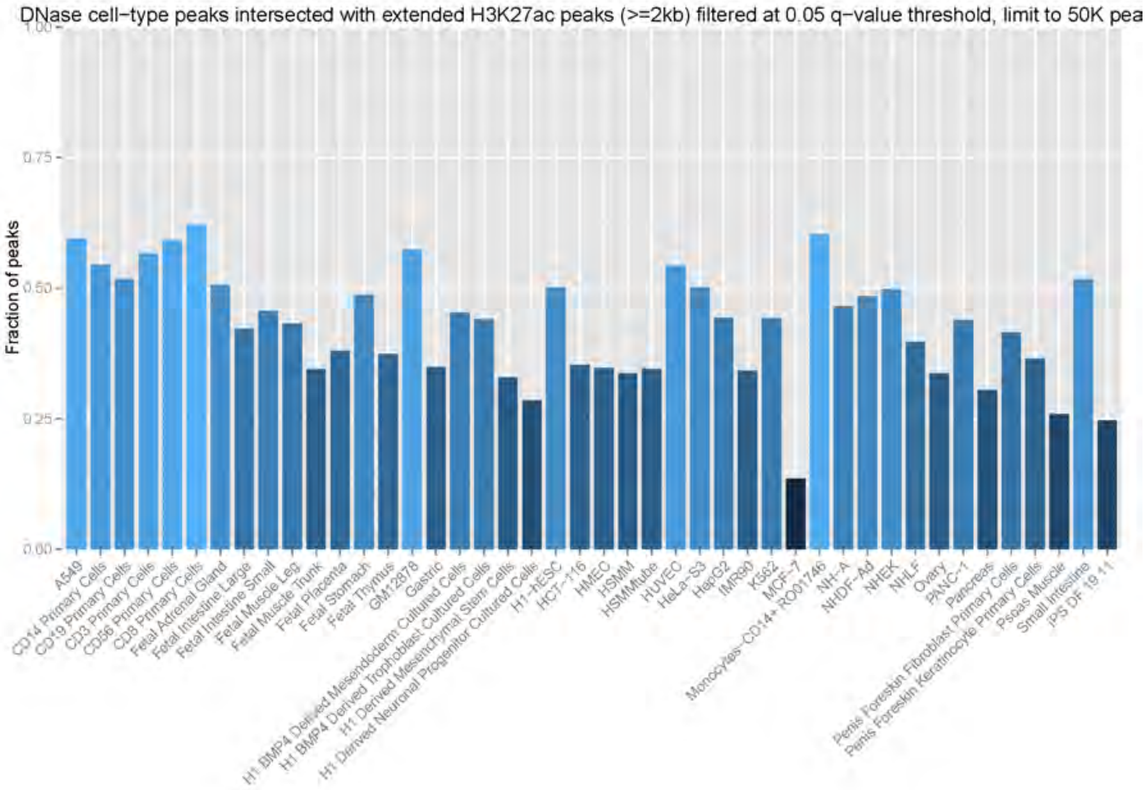


Figure VII-2 | Fraction of DNase peaks overlapping H3K27ac peaks (hg19)
DNase-seq "master peaks" indicate open regions of chromatin; many of these regions have H3K27ac marks, but many other epigenetic activities (TF binding, other histone modifications, etc.) may be taking place.

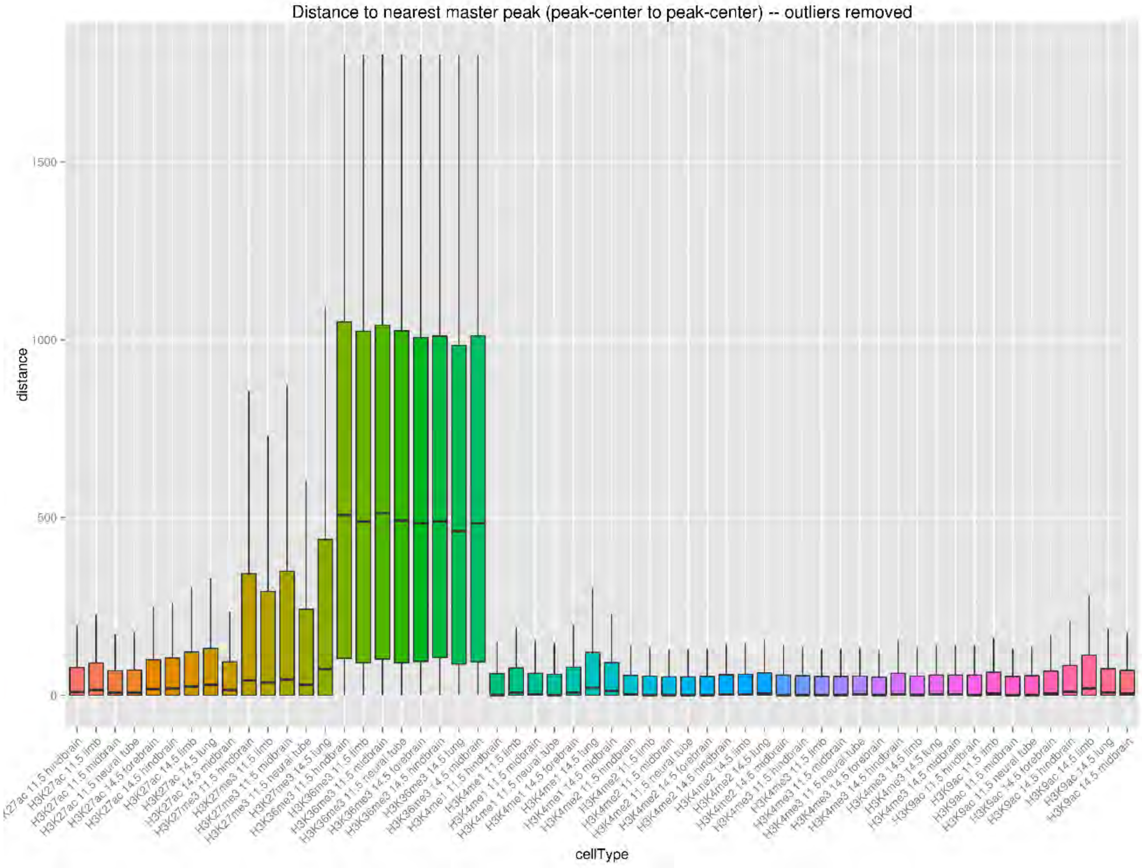


Figure VII-3 | Encyclopedia Master Peak Distances (mm10)
Distances to nearest master peak for different histone marks across different biosamples. Ideally, master peaks should have relatively similar peak distributions for the same histone marks across different biosamples. Very short distances may indicate large number of weak peaks.

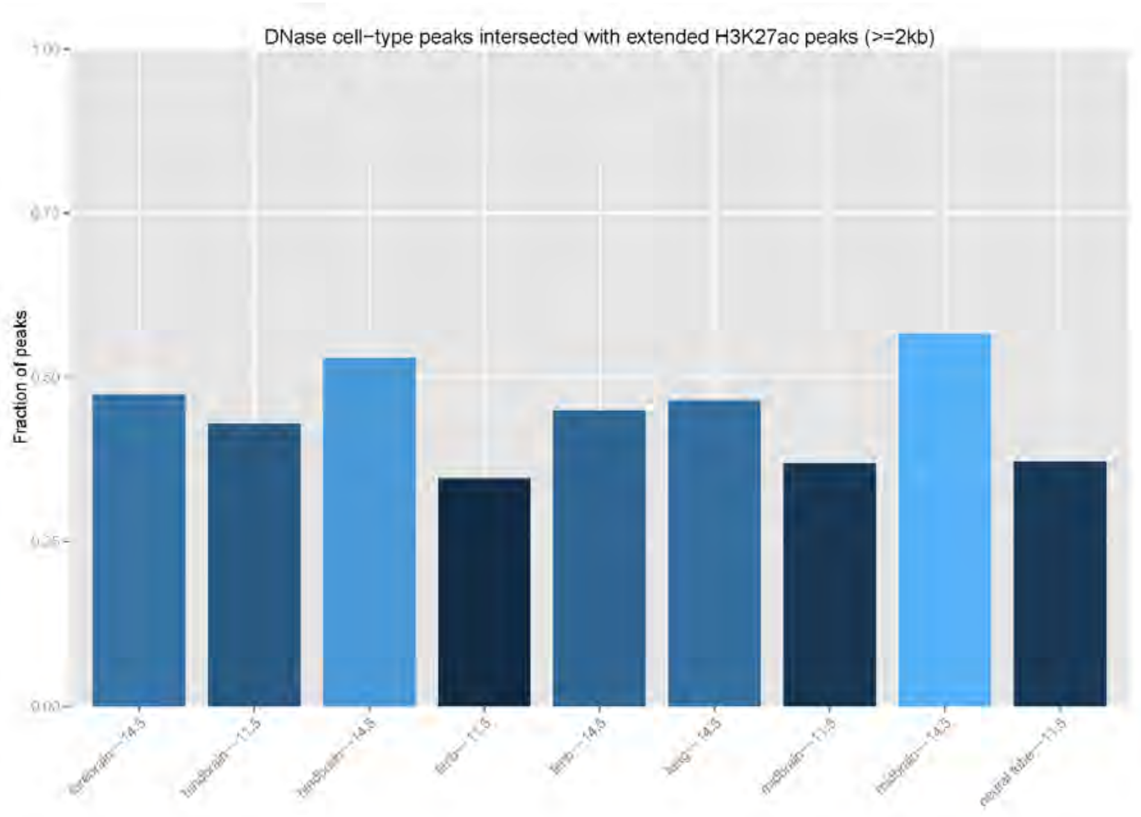


Figure VII-4 | Fraction of DNase peaks overlapping H3K27ac peaks (mm10)
DNase-seq “master peaks” indicate open regions of chromatin; many of these regions have H3K27ac marks, but many other epigenetic activities (TF binding, other histone modifications, etc.) may be taking place.

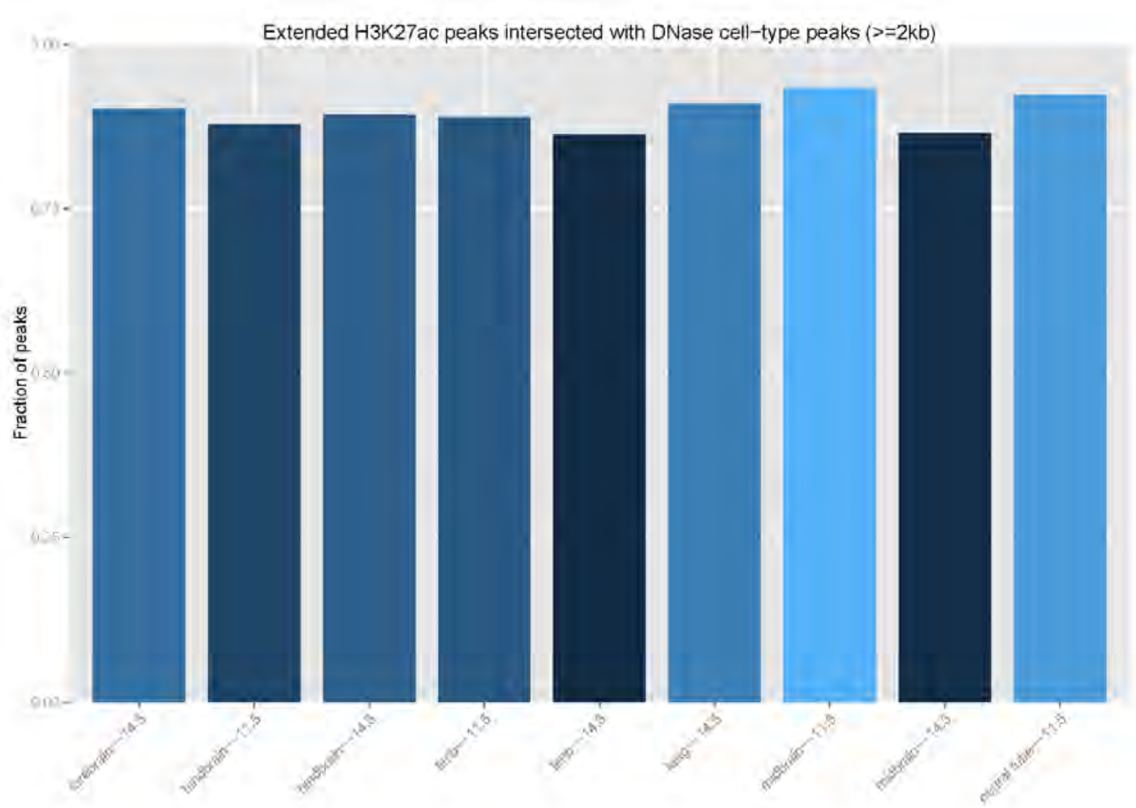


Figure VII-5 | Fraction of H3K27ac peaks that overlap DNase peaks (mm10)
DNase-seq “master peaks” overlap most (if not all) H3K27ac peaks, indicating we’re identifying the majority of sites with enhancer-like epigenetic signals based on H3K27ac.

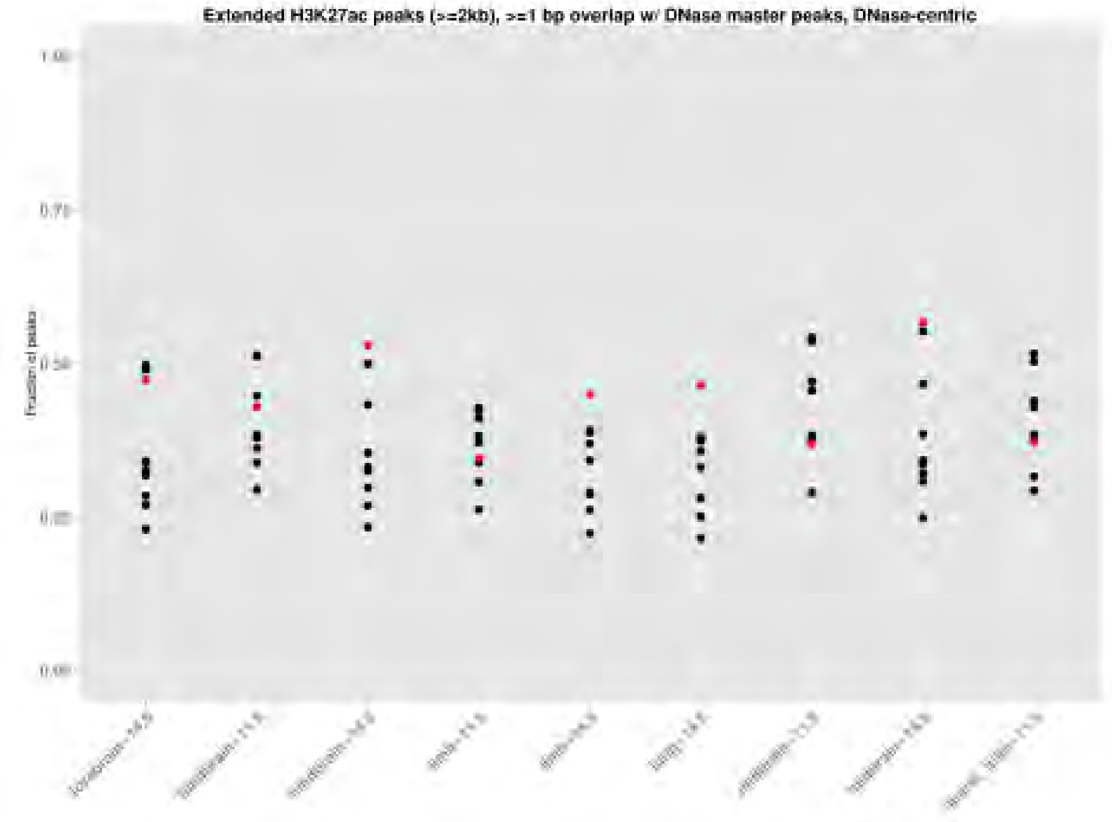


Figure VII-6 | | Cross-biosample master peak overlap
Compare intersection of master peaks in single cell type against peaks in all other cell types.

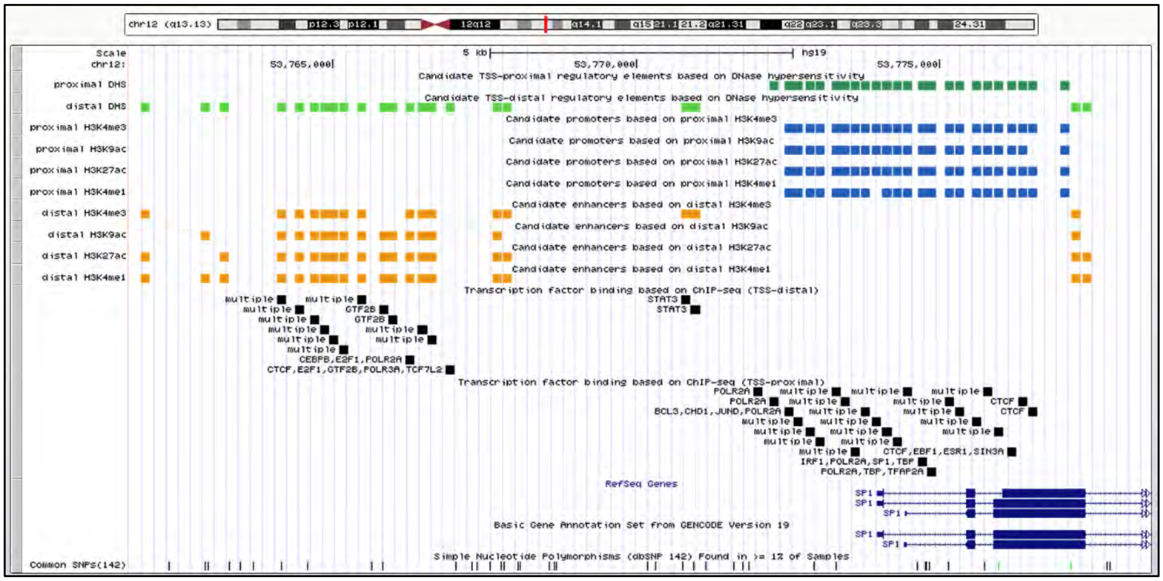
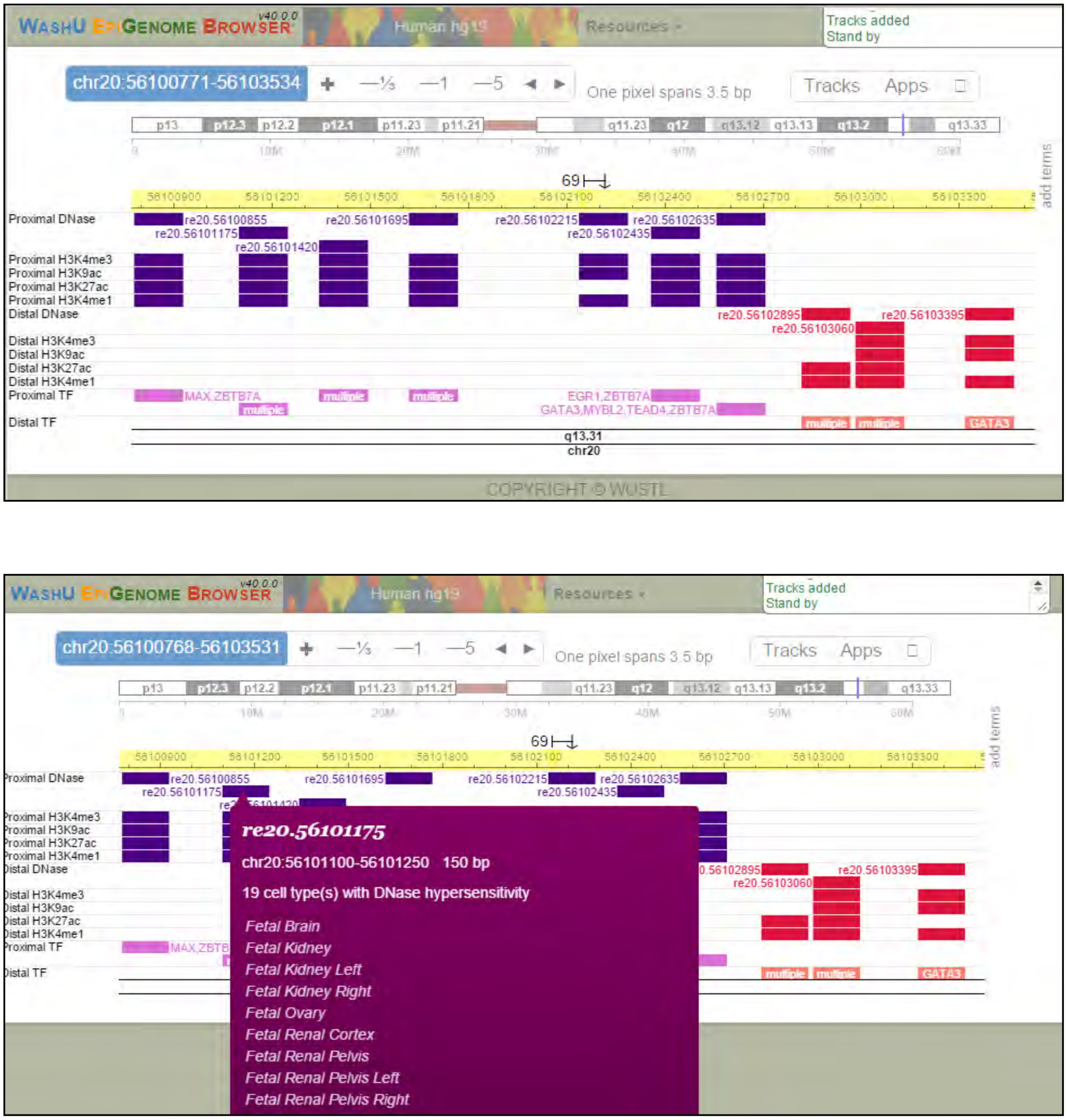

chr12 (q13.13)
p12.3
p12.1
12q12
q14.1
q15
21.1
21.2
q21.31
q22
q23.1
q23.3
24.31
Scale
5 kb
hg19
chr12:
53,765,000
53,770,000
53,775,000
proximal DHS
distal DHS
proximal H3K4me3
proximal H3K9ac
proximal H3K27ac
proximal H3K4me1
distal H3K4me3
distal H3K9ac
distal H3K27ac
distal H3K4me1
Candidate TSS-proximal regulatory elements based on DNase hypersensitivity
Candidate TSS-distal regulatory elements based on DNase hypersensitivity
Candidate promoters based on proximal H3K4me3
Candidate promoters based on proximal H3K9ac
Candidate promoters based on proximal H3K27ac
Candidate promoters based on proximal H3K4me1
Candidate enhancers based on distal H3K4me3
Candidate enhancers based on distal H3K9ac
Candidate enhancers based on distal H3K27ac
Candidate enhancers based on distal H3K4me1
Transcription factor binding based on ChIP-seq (TSS-distal)
multiple
multiple
multiple
GTF2B
multiple
GTF2B
multiple
multiple
multiple
multiple
multiple
CEBPB,E2F1,POLR2A
CTCF,E2F1,GTF2B,POLR3A,TCF7L2
STAT3
STAT3
Transcription factor binding based on ChIP-seq (TSS-proximal)
POLR2A
multiple
multiple
multiple
POLR2A
multiple
multiple
CTCF
BCL3,CHD1,JUND,POLR2A
multiple
multiple
CTCF
multiple
multiple
multiple
multiple
multiple
multiple
multiple
multiple
multiple
CTCF,EBF1,ESR1,SIN3A
IRF1,POLR2A,SP1,TBP
POLR2A,TBP,TFAP2A
RefSeq Genes
SP1
SP1
SP1
Basic Gene Annotation Set from GENCODE Version 19
SP1
SP1
Common SNPs(142)
Simple Nucleotide Polymorphisms (dbSNP 142) Found in >= 1% of Samples

Figure VII-7 | UCSC Visualization
Example region showing V2 candidate promoter and enhancer regions.



re20.56101175

chr20:56101100-56101250 150 bp

19 cell type(s) with DNase hypersensitivity

Fetal Brain
Fetal Kidney
Fetal Kidney Left
Fetal Kidney Right
Fetal Ovary
Fetal Renal Cortex
Fetal Renal Pelvis
Fetal Renal Pelvis Left
Fetal Renal Pelvis Right

Figure VII-8 | WashU Genome Browser
Example region showing V2 candidate promoter and enhancer regions

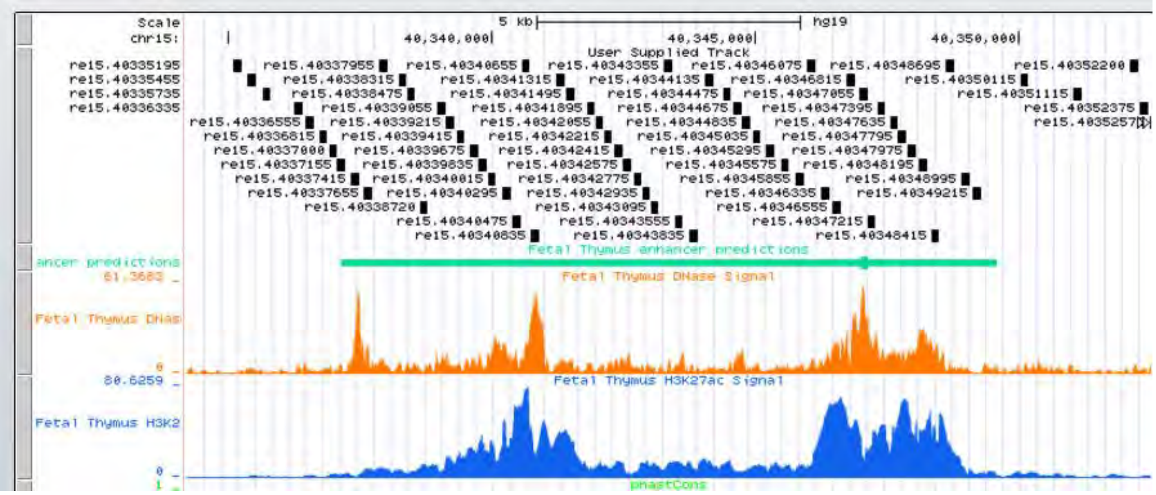


Figure VII-9 | Problems with early versions
Large number of “weak” DNase master peaks

VIII. Appendix B: Encyclopedia V3

VIII.1 Preface

The unpublished work described in this appendix was performed along with Jill Moore and Zhiping Weng; the methods and tracks are available through ENCODE[8].

VIII.2 Introduction

Version 3 of the Encyclopedia incorporated many suggestions based upon feedback from Version 2. This version selected candidate enhancer and promoter elements based upon ranking of DNase-seq, H3K27ac, and H3K4me3 signals. This approach was developed in the context of improving our enhancer predictions for VISTA regions (see IV.4 Enhancer Prediction on page 170). We also started development of our own visualizer, to better ease selecting and intersection of datasets of intersect.

Lastly, we started to formalize the overall structure of the ENCODE Encyclopedia, including what assays would and would not be incorporated, and how particular analysis products should be grouped. This hierarchy of data (Figure VIII-1 on page 245) placed data in the context of what information it provided. First, ground level annotations are typically derived directly from the experimental data. Next, middle level annotations integrate multiple types of experimental data and multiple ground level annotations. Finally, top level annotations integrate a broad range of experimental data and ground and middle level annotations. For instance, chromatin states in the Top Level of the Encyclopedia could be made from enhancer-like and promoter-like elements determined at the Middle Level which, themselves, depended upon peaks calls made in

[8] https://www.encodeproject.org/data/annotations/v3/

the Ground Level of the Encyclopedia. Files such as raw reads or alignments are too data and uncondensed to be formally incorporated in the Encyclopedia; they just form the input to the Ground Level.

VIII.3 Visualizer

The visualizer for Version 3 needed to fulfill certain requirements. First, it needed to provide an integrated interface to query annotations by gene, SNP, or genomic position. Second, it also needed a way to present the user with a list of available cell types (including developmental time points, ontology information, etc.). Lastly, it had to provide a gateway to view signal and annotations files in the UCSC Genome Browser. The main page of the visualizer is show in Figure VIII-2 on page 246, and demonstrates the integrated interface for search and cell type selection (Figure VIII-5 on page 249). This page ultimately produces a dynamically-made UCSC Trackhub, as shown in Figure VIII-4 on page 248.

This visualizer was the first to allow download of bed files containing the intersecting candidate elements, as well as the first to color predicted promoters near TSSs differently from enhancers. I experimented with best normalization techniques for overlay DNase with H3K27ac signals in the UCSC Genome Browser. This is also the first visualizer to exploit ENCODE’s ontology information (and to note how incomplete the ontology information was).

This visualizer ultimately proved to be a useful proof-of-concept for candidate enhancer visualization; many of the techniques and approaches used ended up as the basis for SCREEN. It also greatly helped stir discussion and feedback from the ENCODE community.

VIII.4 Figures

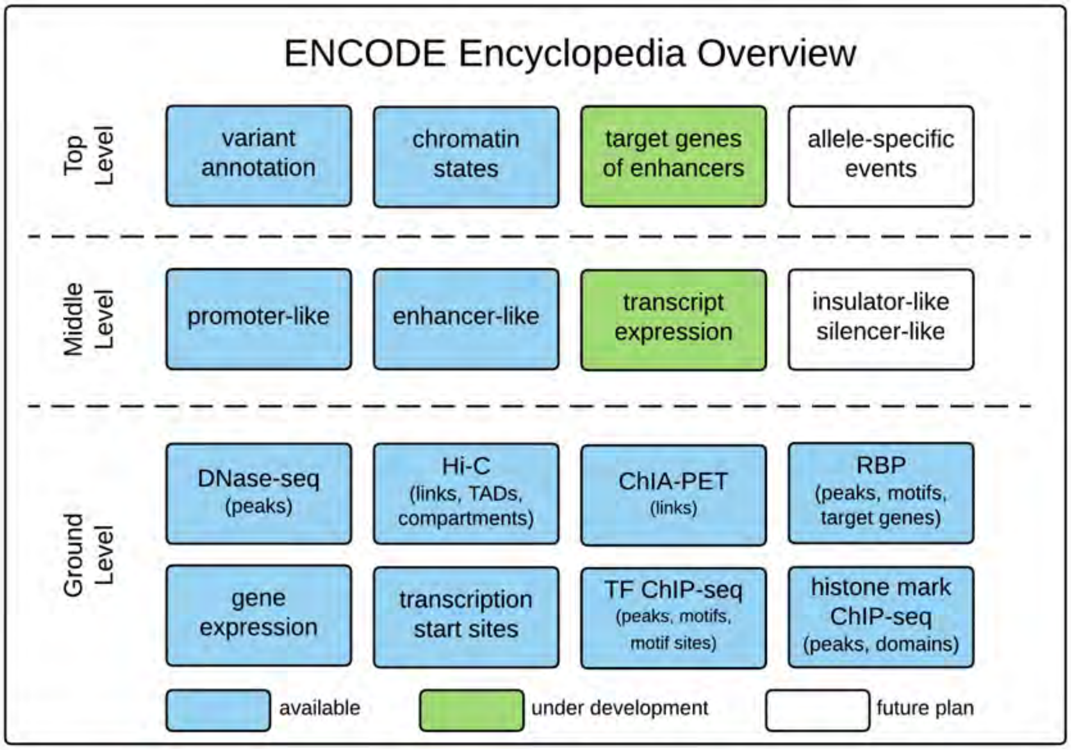
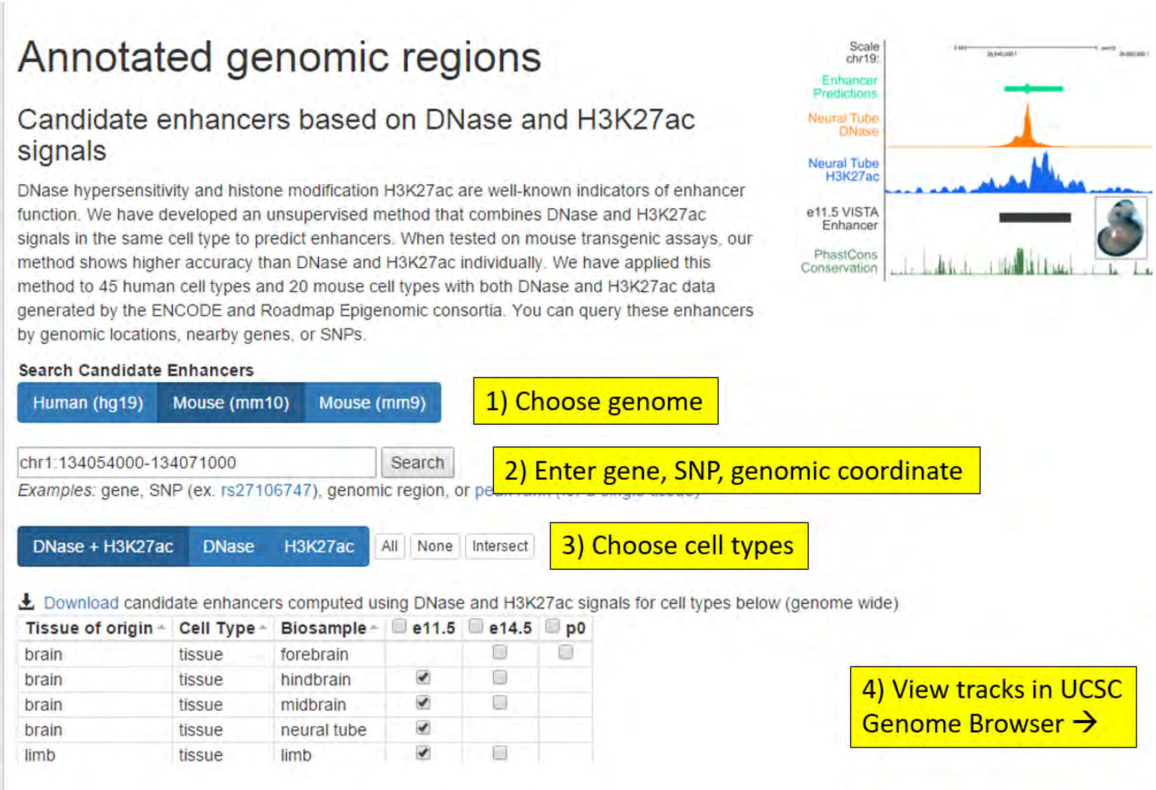

ENCODE Encyclopedia Overview
Top Level
Middle Level
Ground Level
variant annotation
chromatin states
target genes of enhancers
allele-specific events
promoter-like
enhancer-like
transcript expression
insulator-like silencer-like
DNase-seq (peaks)
Hi-C (links, TADs, compartments)
ChIA-PET (links)
RBP (peaks, motifs, target genes)
gene expression
transcription start sites
TF ChIP-seq (peaks, motifs, motif sites)
histone mark ChIP-seq (peaks, domains)
available
under development
future plan

Figure VIII-1 | Overview Of ENCODE Encyclopedia V3
Version 3 of the Encyclopedia incorporated 3 levels of annotations:

- Ground level annotations are typically derived directly from the experimental data.
- Middle level annotations integrate multiple types of experimental data and multiple ground level annotations.
- Top level annotations integrate a broad range of experimental data and ground and middle level annotations.



Annotated genomic regions

Candidate enhancers based on DNase and H3K27ac signals

DNase hypersensitivity and histone modification H3K27ac are well-known indicators of enhancer function. We have developed an unsupervised method that combines DNase and H3K27ac signals in the same cell type to predict enhancers. When tested on mouse transgenic assays, our method shows higher accuracy than DNase and H3K27ac individually. We have applied this method to 45 human cell types and 20 mouse cell types with both DNase and H3K27ac data generated by the ENCODE and Roadmap Epigenomic consortia. You can query these enhancers by genomic locations, nearby genes, or SNPs.

Search Candidate Enhancers

1) Choose genome

chr1:134054000-134071000

Examples: gene, SNP (ex. rs27106747), genomic region, or pe

2) Enter gene, SNP, genomic coordinate

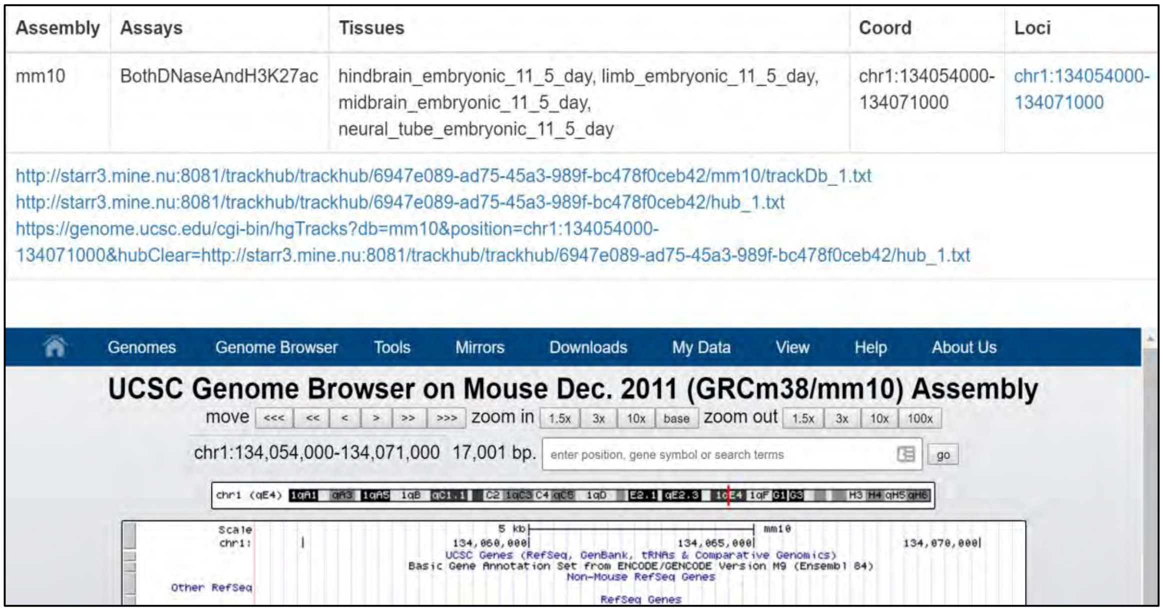
3) Choose cell types

Download candidate enhancers computed using DNase and H3K27ac signals for cell types below (genome wide)

Tissue of origin	Cell Type	Biosample	□ e11.5	□ e14.5	□ p0
brain	tissue	forebrain		□	□
brain	tissue	hindbrain	☑	□	
brain	tissue	midbrain	☑	□	
brain	tissue	neural tube	☑		
limb	tissue	limb	☑	□	

4) View tracks in UCSC Genome Browser →

Figure VIII-2 | Visualizer Main Page
This page provide an integrated interface to query annotations by gene, SNP, or genomic position, as well as show available cell types, and link out to the UCSC Genome Browser.



Assembly	Assays	Tissues	Coord	Loci
mm10	BothDNaseAndH3K27ac	hindbrain_embryonic_11_5_day, limb_embryonic_11_5_day, midbrain_embryonic_11_5_day, neural_tube_embryonic_11_5_day	chr1:134054000-134071000	chr1:134054000-134071000

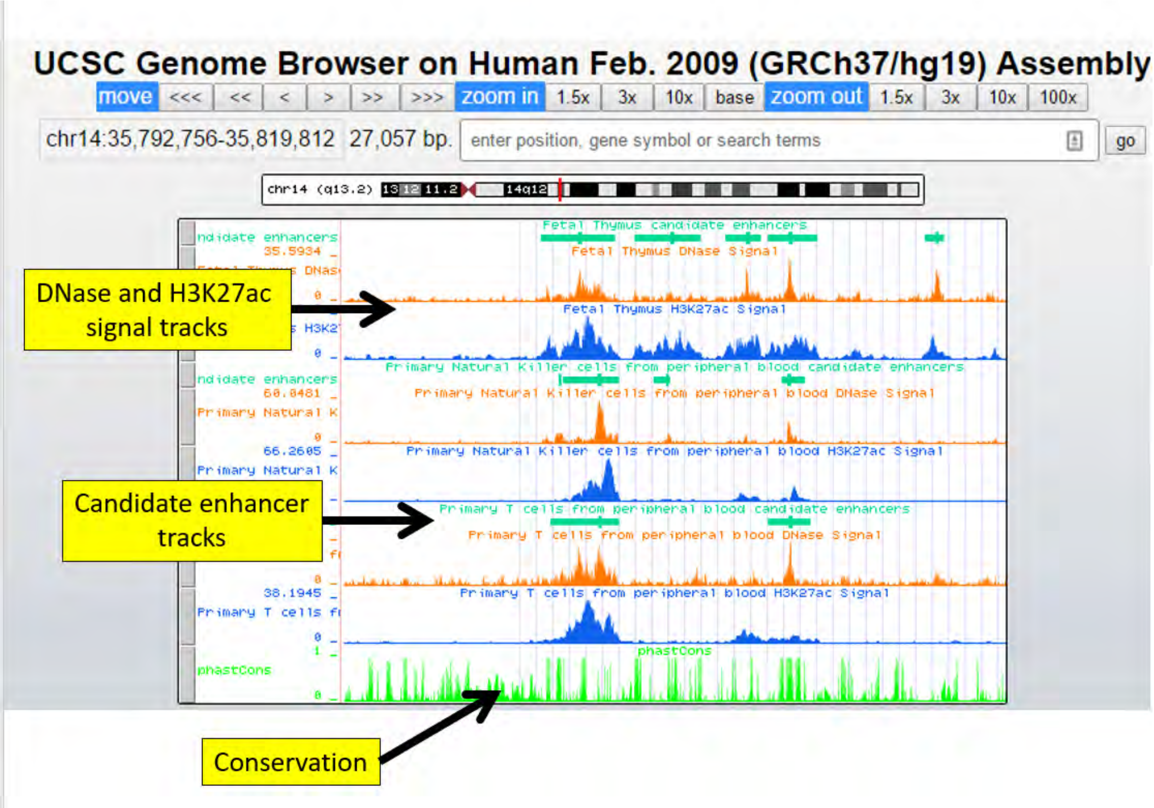
http://starr3.mine.nu:8081/trackhub/trackhub/6947e089-ad75-45a3-989f-bc478f0ceb42/mm10/trackDb_1.txt
http://starr3.mine.nu:8081/trackhub/trackhub/6947e089-ad75-45a3-989f-bc478f0ceb42/hub_1.txt
https://genome.ucsc.edu/cgi-bin/hgTracks?db=mm10&position=chr1:134054000-134071000&hubClear=http://starr3.mine.nu:8081/trackhub/trackhub/6947e089-ad75-45a3-989f-bc478f0ceb42/hub_1.txt

UCSC Genome Browser on Mouse Dec. 2011 (GRCm38/mm10) Assembly

chr1:134,054,000-134,071,000 17,001 bp.

Genomes
Genome Browser
Tools
Mirrors
Downloads
My Data
View
Help
About Us
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x
enter position, gene symbol or search terms
go
chr1 (qE4) 1qA1 qA3 1qA5 1qB qC1.1 C2 1qC3 C4 qC5 1qD E2.1 qE2.3 1qE4 1qF G1 G3 H3 H4 qH5 qH6
Scale
5 kb
mm10
chr1:
134,060,000
134,065,000
134,070,000
UCSC Genes (RefSeq, GenBank, tRNAs & Comparative Genomics)
Basic Gene Annotation Set from ENCODE/GENCODE Version M9 (Ensembl 84)
Non-Mouse RefSeq Genes
Other RefSeq
RefSeq Genes

Figure VIII-3 | Dynamic trackhubs
The visualizer created trackhubs on-the-fly, permitting dynamic reconfiguration of what the user was viewing in both UCSC and WashU Genome Browsers.

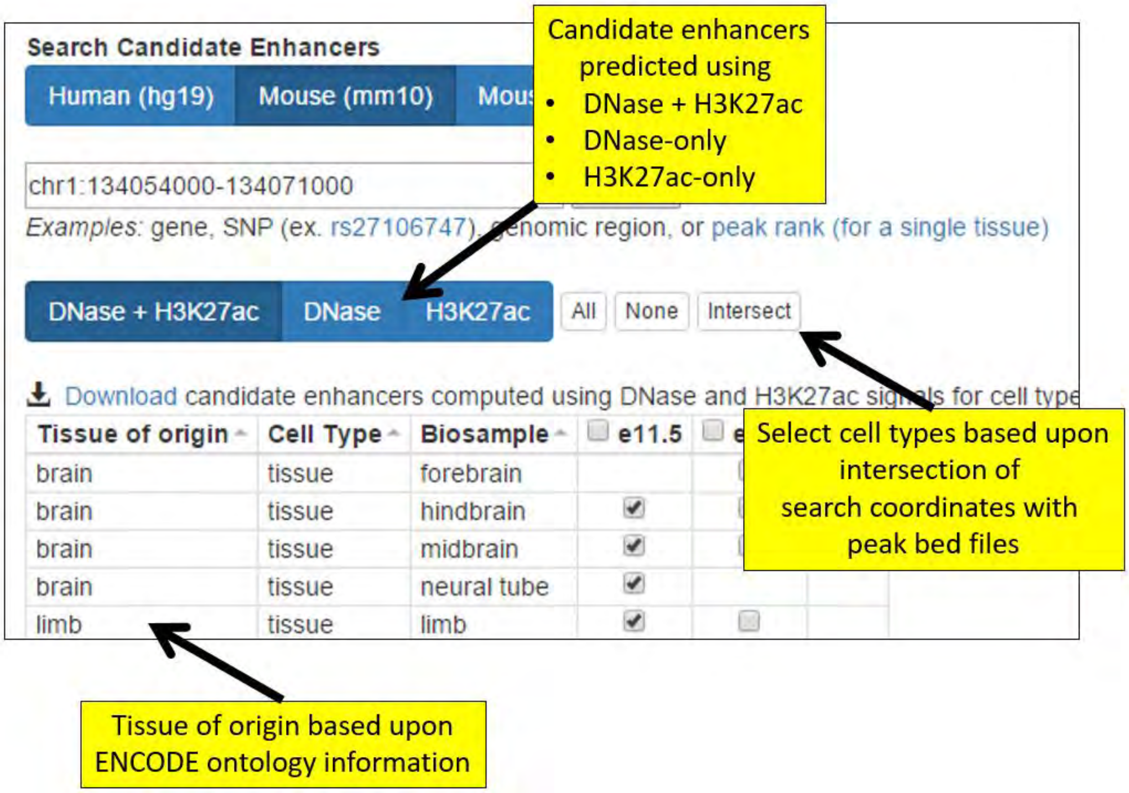


UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

chr14:35,792,756-35,819,812 27,057 bp.

move
<<<
<<
<
>
>>
>>>
zoom in
1.5x
3x
10x
base
zoom out
1.5x
3x
10x
100x
enter position, gene symbol or search terms
go
chr14 (q13.2)
13
12
11.2
14q12
Fetal Thymus candidate enhancers
Fetal Thymus DNase Signal
Fetal Thymus H3K27ac Signal
Primary Natural Killer cells from peripheral blood candidate enhancers
Primary Natural Killer cells from peripheral blood DNase Signal
Primary Natural Killer cells from peripheral blood H3K27ac Signal
Primary T cells from peripheral blood candidate enhancers
Primary T cells from peripheral blood DNase Signal
Primary T cells from peripheral blood H3K27ac Signal
phastCons
35.5934
68.8481
66.2605
38.1945
1
0
DNase and H3K27ac signal tracks
Candidate enhancer tracks
Conservation

Figure VIII-4 | UCSC Genome Browser
Example trackhub for UCSC



Search Candidate Enhancers

chr1:134054000-134071000

Examples: gene, SNP (ex. rs27106747), genomic region, or peak rank (for a single tissue)

Download candidate enhancers computed using DNase and H3K27ac signals for cell type

Tissue of origin	Cell Type	Biosample	□ e11.5	□ e
brain	tissue	forebrain		
brain	tissue	hindbrain	☑	
brain	tissue	midbrain	☑	
brain	tissue	neural tube	☑	
limb	tissue	limb	☑	□

Candidate enhancers predicted using

- DNase + H3K27ac
- DNase-only
- H3K27ac-only

Select cell types based upon intersection of search coordinates with peak bed files

Tissue of origin based upon ENCODE ontology information

Figure VIII-5 | Visualizer Search Interface
Users could select cell types to display genome tracks based upon intersection of their coordinate region with our database of hundreds of millions of peaks from ENCODE bed files.

IX. Chapter VII: Bibliography

Aken, Bronwen L., Sarah Ayling, Daniel Barrell, Laura Clarke, Valery Curwen, Susan Fairley, Julio Fernandez Banet, Konstantinos Billis, Carlos García Girón, Thibaut Hourlier, Kevin Howe, Andreas Kähäri, Felix Kokocinski, Fergal J. Martin, Daniel N. Murphy, Rishi Nag, Magali Ruffier, Michael Schuster, Y. Amy Tang, Jan-Hinnerk Vogel, Simon White, Amonida Zadissa, Paul Flicek, and Stephen M. J. Searle. 2016. 'The Ensembl gene annotation system', *Database: The Journal of Biological Databases and Curation*, 2016: baw093.

Allen, E. K., A. G. Randolph, T. Bhangale, P. Dogra, M. Ohlson, C. M. Oshansky, A. E. Zamora, J. P. Shannon, D. Finkelstein, A. Dressen, J. DeVincenzo, M. Caniza, B. Youngblood, C. M. Rosenberger, and P. G. Thomas. 2017. 'SNP-mediated disruption of CTCF binding at the IFITM3 promoter is associated with risk of severe influenza in humans', *Nat Med*, 23: 975-83.

Andersson, R., C. Gebhard, I. Miguel-Escalada, I. Hoof, J. Bornholdt, M. Boyd, Y. Chen, X. Zhao, C. Schmidl, T. Suzuki, E. Ntini, E. Arner, E. Valen, K. Li, L. Schwarzfischer, D. Glatz, J. Raithel, B. Lilje, N. Rapin, F. O. Bagger, M. Jorgensen, P. R. Andersen, N. Bertin, O. Rackham, A. M. Burroughs, J. K. Baillie, Y. Ishizu, Y. Shimizu, E. Furuhata, S. Maeda, Y. Negishi, C. J. Mungall, T. F. Meehan, T. Lassmann, M. Itoh, H. Kawaji, N. Kondo, J. Kawai, A. Lennartsson, C. O. Daub, P. Heutink, D. A. Hume, T. H. Jensen, H. Suzuki, Y. Hayashizaki, F. Muller, A. R. R. Forrest, P. Carninci, M. Rehli, and A. Sandelin. 2014. 'An atlas of active enhancers across human cell types and tissues', *Nature*, 507: 455-61.

Ashoor, H., D. Kleftogiannis, A. Radovanovic, and V. B. Bajic. 2015. 'DENdb: database of integrated human enhancers', *Database (Oxford)*, 2015.

Bailey, T. L., M. Boden, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li, and W. S. Noble. 2009. 'MEME SUITE: tools for motif discovery and searching', *Nucleic Acids Res*, 37: W202-8.

Bailey, T. L., and C. Elkan. 1994. 'Fitting a mixture model by expectation maximization to discover motifs in biopolymers', *Proc Int Conf Intell Syst Mol Biol*, 2: 28-36.

Batut, P., A. Dobin, C. Plessy, P. Carninci, and T. R. Gingeras. 2013. 'High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression', *Genome Res*, 23: 169-80.

Bejerano, G., M. Pheasant, I. Makunin, S. Stephen, W. J. Kent, J. S. Mattick, and D. Haussler. 2004. 'Ultraconserved elements in the human genome', *Science*, 304: 1321-5.

Bergen, S. E., C. T. O'Dushlaine, S. Ripke, P. H. Lee, D. M. Ruderfer, S. Akterin, J. L. Moran, K. D. Chambert, R. E. Handsaker, L. Backlund, U. Osby, S. McCarroll, M. Landen, E. M. Scolnick, P. K. Magnusson, P. Lichtenstein, C. M. Hultman, S. M. Purcell, P. Sklar, and P. F. Sullivan. 2012. 'Genome-wide association study in a Swedish population yields support for greater CNV and MHC involvement in schizophrenia compared with bipolar disorder', *Mol Psychiatry*, 17: 880-6.

Berman, Helen M., John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. 2000. 'The Protein Data Bank', *Nucleic Acids Res*, 28: 235-42.

Birney, E., T. D. Andrews, P. Bevan, M. Caccamo, Y. Chen, L. Clarke, G. Coates, J. Cuff, V. Curwen, T. Cutts, T. Down, E. Eyras, X. M. Fernandez-Suarez, P. Gane, B. Gibbins, J. Gilbert, M. Hammond, H. R. Hotz, V. Iyer, K. Jekosch, A. Kahari, A. Kasprzyk, D. Keefe, S. Keenan, H. Lehvaslaiho, G. McVicker, C. Melsopp, P. Meidl, E. Mongin, R. Pettett, S. Potter, G. Proctor, M. Rae, S. Searle, G. Slater, D. Smedley, J. Smith, W. Spooner, A. Stabenau, J. Stalker, R. Storey, A. Ureta-Vidal, K. C. Woodwark, G. Cameron, R. Durbin, A. Cox, T. Hubbard, and M. Clamp. 2004. 'An overview of Ensembl', *Genome Res*, 14: 925-8.

Birney, E., J. A. Stamatoyannopoulos, A. Dutta, R. Guigo, T. R. Gingeras, E. H. Margulies, Z. Weng, M. Snyder, E. T. Dermitzakis, R. E. Thurman, M. S. Kuehn, C. M. Taylor, S. Neph, C. M. Koch, S. Asthana, A. Malhotra, I. Adzhubei, J. A. Greenbaum, R. M. Andrews, P. Flicek, P. J. Boyle, H. Cao, N. P. Carter, G. K. Clelland, S. Davis, N. Day, P. Dhami, S. C. Dillon, M. O. Dorschner, H. Fiegler, P. G. Giresi, J. Goldy, M. Hawrylycz, A. Haydock, R. Humbert, K. D. James, B. E. Johnson, E. M. Johnson, T. T. Frum, E. R. Rosenzweig, N. Karnani, K. Lee, G. C. Lefebvre, P. A. Navas, F. Neri, S. C. Parker, P. J. Sabo, R. Sandstrom, A. Shafer, D. Vetrie, M. Weaver, S. Wilcox, M. Yu, F. S. Collins, J. Dekker, J. D. Lieb, T. D. Tullius, G. E. Crawford, S. Sunyaev, W. S. Noble, I. Dunham, F. Denoeud, A. Reymond, P. Kapranov, J. Rozowsky, D. Zheng, R. Castelo, A. Frankish, J. Harrow, S. Ghosh, A. Sandelin, I. L. Hofacker, R. Baertsch, D. Keefe, S. Dike, J. Cheng, H. A. Hirsch, E. A. Sekinger, J. Lagarde, J. F. Abril, A. Shahab, C. Flamm, C. Fried, J. Hackermuller, J. Hertel, M. Lindemeyer, K. Missal, A. Tanzer, S. Washietl, J. Korbel, O. Emanuelsson, J. S. Pedersen, N. Holroyd, R. Taylor, D. Swarbreck, N. Matthews, M. C. Dickson, D. J. Thomas, M. T. Weirauch, J. Gilbert, J. Drenkow, I. Bell, X. Zhao, K. G. Srinivasan, W. K. Sung, H. S. Ooi, K. P. Chiu, S. Foissac, T. Alioto, M. Brent, L. Pachter, M. L. Tress, A. Valencia, S. W. Choo, C. Y. Choo, C. Ucla, C. Manzano, C. Wyss, E. Cheung, T. G. Clark, J. B. Brown, M. Ganesh, S. Patel, H. Tammana, J. Chrast, C. N. Henrichsen, C. Kai, J. Kawai, U. Nagalakshmi, J. Wu, Z. Lian, J. Lian, P. Newburger, X. Zhang, P. Bickel, J. S. Mattick, P. Carninci, Y. Hayashizaki, S. Weissman, T. Hubbard, R. M. Myers, J. Rogers, P. F. Stadler, T. M. Lowe, C. L. Wei, Y. Ruan, K. Struhl, M. Gerstein, S. E. Antonarakis, Y. Fu, E. D. Green, U. Karaoz, A. Siepel, J. Taylor, L. A. Liefer, K. A. Wetterstrand, P. J. Good, E. A. Feingold, M. S. Guyer, G. M. Cooper, G. Asimenos, C. N. Dewey, M. Hou, S. Nikolaev, J. I. Montoya-Burgos, A. Loytynoja, S. Whelan, F. Pardi, T. Massingham, H. Huang, N. R. Zhang, I. Holmes, J. C. Mullikin, A. Ureta-Vidal, B. Paten, M. Seringhaus, D. Church, K. Rosenbloom, W. J. Kent, E. A. Stone, S. Batzoglou, N. Goldman, R. C. Hardison, D. Haussler, W. Miller, A. Sidow, N. D. Trinklein, Z. D. Zhang, L. Barrera, R. Stuart, D. C. King, A. Ameur, S. Enroth,

M. C. Bieda, J. Kim, A. A. Bhinge, N. Jiang, J. Liu, F. Yao, V. B. Vega, C. W. Lee, P. Ng, A. Shahab, A. Yang, Z. Moqtaderi, Z. Zhu, X. Xu, S. Squazzo, M. J. Oberley, D. Inman, M. A. Singer, T. A. Richmond, K. J. Munn, A. Rada-Iglesias, O. Wallerman, J. Komorowski, J. C. Fowler, P. Couttet, A. W. Bruce, O. M. Dovey, P. D. Ellis, C. F. Langford, D. A. Nix, G. Euskirchen, S. Hartman, A. E. Urban, P. Kraus, S. Van Calcar, N. Heintzman, T. H. Kim, K. Wang, C. Qu, G. Hon, R. Luna, C. K. Glass, M. G. Rosenfeld, S. F. Aldred, S. J. Cooper, A. Halees, J. M. Lin, H. P. Shulha, X. Zhang, M. Xu, J. N. Haidar, Y. Yu, Y. Ruan, V. R. Iyer, R. D. Green, C. Wadelius, P. J. Farnham, B. Ren, R. A. Harte, A. S. Hinrichs, H. Trumbower, H. Clawson, J. Hillman-Jackson, A. S. Zweig, K. Smith, A. Thakkapallayil, G. Barber, R. M. Kuhn, D. Karolchik, L. Armengol, C. P. Bird, P. I. de Bakker, A. D. Kern, N. Lopez-Bigas, J. D. Martin, B. E. Stranger, A. Woodroffe, E. Davydov, A. Dimas, E. Eyras, I. B. Hallgrimsdottir, J. Huppert, M. C. Zody, G. R. Abecasis, X. Estivill, G. G. Bouffard, X. Guan, N. F. Hansen, J. R. Idol, V. V. Maduro, B. Maskeri, J. C. McDowell, M. Park, P. J. Thomas, A. C. Young, R. W. Blakesley, D. M. Muzny, E. Sodergren, D. A. Wheeler, K. C. Worley, H. Jiang, G. M. Weinstock, R. A. Gibbs, T. Graves, R. Fulton, E. R. Mardis, R. K. Wilson, M. Clamp, J. Cuff, S. Gnerre, D. B. Jaffe, J. L. Chang, K. Lindblad-Toh, E. S. Lander, M. Koriabine, M. Nefedov, K. Osoegawa, Y. Yoshinaga, B. Zhu, and P. J. de Jong. 2007. 'Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project', *Nature*, 447: 799-816.

Blackwood, E. M., and J. T. Kadonaga. 1998. 'Going the distance: a current view of enhancer action', *Science*, 281: 60-3.

Blow, M. J., D. J. McCulley, Z. Li, T. Zhang, J. A. Akiyama, A. Holt, I. Plajzer-Frick, M. Shoukry, C. Wright, F. Chen, V. Afzal, J. Bristow, B. Ren, B. L. Black, E. M. Rubin, A. Visel, and L. A. Pennacchio. 2010. 'ChIP-Seq identification of weakly conserved heart enhancers', *Nat Genet*, 42: 806-10.

Bostock, Michael, Vadim Ogievetsky, and Jeffrey Heer. 2011. 'D<sup>3</sup> Data-Driven Documents', *IEEE Transactions on Visualization and Computer Graphics*, 17: 2301-09.

Boyle, A. P., E. L. Hong, M. Hariharan, Y. Cheng, M. A. Schaub, M. Kasowski, K. J. Karczewski, J. Park, B. C. Hitz, S. Weng, J. M. Cherry, and M. Snyder. 2012. 'Annotation of functional variation in personal genomes using RegulomeDB', *Genome Res*, 22: 1790-7.

Buenrostro, Jason D., Paul G. Giresi, Lisa C. Zaba, Howard Y. Chang, and William J. Greenleaf. 2013. 'Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position', *Nature Methods*, 10: 1213.

Calo, Eliezer, and Joanna Wysocka. 2013. 'Modification of enhancer chromatin: what, how and why?', *Mol Cell*, 49: 10.1016/j.molcel.2013.01.038.

Chen, Tianqi, and Carlos Guestrin. 2016. "XGBoost: A Scalable Tree Boosting System." In *Proceedings of the 22nd ACM SIGKDD International Conference on*

Knowledge Discovery and Data Mining, 785-94. San Francisco, California, USA: ACM.

Chu, Cheng T., Sang K. Kim, Yi A. Lin, Yuanyuan Yu, Gary R. Bradski, Andrew Y. Ng, and Kunle Olukotun. 2006. "Map-Reduce for Machine Learning on Multicore." In *NIPS*, edited by Bernhard Schölkopf, John C. Platt and Thomas Hoffman, 281--88. MIT Press.

Cloutier, J., S. Kpodjedo, and G. El Boussaidi. 2016. "WAVI: A reverse engineering tool for web applications." In *2016 IEEE 24th International Conference on Program Comprehension (ICPC)*, 1-3.

Consortium, Encode Project. 2012. 'An integrated encyclopedia of DNA elements in the human genome', *Nature*, 489: 57-74.

Creyghton, M. P., A. W. Cheng, G. G. Welstead, T. Kooistra, B. W. Carey, E. J. Steine, J. Hanna, M. A. Lodato, G. M. Frampton, P. A. Sharp, L. A. Boyer, R. A. Young, and R. Jaenisch. 2010. 'Histone H3K27ac separates active from poised enhancers and predicts developmental state', *Proc Natl Acad Sci U S A*, 107: 21931-6.

Crick, F. H. 1958. 'On protein synthesis', *Symp Soc Exp Biol*, 12: 138-63.

Cuellar-Partida, G., F. A. Buske, R. C. McLeay, T. Whitington, W. S. Noble, and T. L. Bailey. 2012. 'Epigenetic priors for identifying active transcription factor binding sites', *Bioinformatics*, 28: 56-62.

Davis, Carrie A., Benjamin C. Hitz, Cricket A. Sloan, Esther T. Chan, Jean M. Davidson, Idan Gabdank, Jason A. Hilton, Kriti Jain, Ulugbek K. Baymuradov, Aditi K. Narayanan, Kathrina C. Onate, Keenan Graham, Stuart R. Miyasato, Timothy R. Dreszer, J. Seth Strattan, Otto Jolanki, Forrest Y. Tanaka, and J. Michael Cherry. 2017. 'The Encyclopedia of DNA elements (ENCODE): data portal update', *Nucleic Acids Res*: gkx1081-gkx81.

Davis, Jesse, and Mark Goadrich. 2006. "The relationship between Precision-Recall and ROC curves." In *Proceedings of the 23rd international conference on Machine learning*, 233-40. Pittsburgh, Pennsylvania: ACM.

De Jager, P. L., X. Jia, J. Wang, P. I. de Bakker, L. Ottoboni, N. T. Aggarwal, L. Piccio, S. Raychaudhuri, D. Tran, C. Aubin, R. Briskin, S. Romano, S. E. Baranzini, J. L. McCauley, M. A. Pericak-Vance, J. L. Haines, R. A. Gibson, Y. Naeglin, B. Uitdehaag, P. M. Matthews, L. Kappos, C. Polman, W. L. McArdle, D. P. Strachan, D. Evans, A. H. Cross, M. J. Daly, A. Compston, S. J. Sawcer, H. L. Weiner, S. L. Hauser, D. A. Hafler, and J. R. Oksenberg. 2009. 'Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci', *Nat Genet*, 41: 776-82.

Dekker, Job, Marc A. Marti-Renom, and Leonid A. Mirny. 2013. 'Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data', *Nat Rev Genet*, 14: 390-403.

Diao, Y., R. Fang, B. Li, Z. Meng, J. Yu, Y. Qiu, K. C. Lin, H. Huang, T. Liu, R. J. Marina, I. Jung, Y. Shen, K. L. Guan, and B. Ren. 2017. 'A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells', *Nat Methods*, 14: 629-35.

Dickel, D. E., I. Barozzi, Y. Zhu, Y. Fukuda-Yuzawa, M. Osterwalder, B. J. Mannion, D. May, C. H. Spurrell, I. Plajzer-Frick, C. S. Pickle, E. Lee, T. H. Garvin, M. Kato, J. A. Akiyama, V. Afzal, A. Y. Lee, D. U. Gorkin, B. Ren, E. M. Rubin, A. Visel, and L. A. Pennacchio. 2016. 'Genome-wide compendium and functional assessment of in vivo heart enhancers', *Nat Commun*, 7: 12923.

Dorschner, Michael O., Michael Hawrylycz, Richard Humbert, James C. Wallace, Anthony Shafer, Janelle Kawamoto, Joshua Mack, Robert Hall, Jeff Goldy, Peter J. Sabo, Ajay Kohli, Qiliang Li, Michael McArthur, and John A. Stamatoyannopoulos. 2004. 'High-throughput localization of functional elements by quantitative chromatin profiling', *Nature Methods*, 1: 219.

Eberharter, Anton, and Peter B. Becker. 2002. 'Histone acetylation: a switch between repressive and permissive chromatin: Second in review series on chromatin dynamics', *EMBO Rep*, 3: 224-29.

Ebert, Peter, and Christoph Bock. 2015. 'Improving reference epigenome catalogs by computational prediction', *Nat Biotechnol*, 33: 354.

Eichenberger, Alexandre, John Mellor-Crummey, Martin Schulz, Nawal Copty, Jim Cownie, Robert Dietrich, Xu Liu, Eugene Loh, and Daniel Lorenz. 2014. 'OMPT: An OpenMP Tools Application Programming Interface for Performance Analysis'.

Elemento, O., and S. Tavazoie. 2005. 'Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach'.

Erceg, J., T. Pakozdi, R. Marco-Ferreres, Y. Ghavi-Helm, C. Girardot, A. P. Bracken, and E. E. Furlong. 2017. 'Dual functionality of cis-regulatory elements as developmental enhancers and Polycomb response elements', *Genes Dev*, 31: 590-602.

Ernst, J., and M. Kellis. 2012. 'ChromHMM: automating chromatin-state discovery and characterization', *Nat Methods*, 9: 215-6.

Ernst, J., P. Kheradpour, T. S. Mikkelsen, N. Shoresh, L. D. Ward, C. B. Epstein, X. Zhang, L. Wang, R. Issner, M. Coyne, M. Ku, T. Durham, M. Kellis, and B. E. Bernstein. 2011. 'Mapping and analysis of chromatin state dynamics in nine human cell types', *Nature*, 473: 43-9.

Ernst, Jason, and Manolis Kellis. 2010. 'Discovery and characterization of chromatin states for systematic annotation of the human genome', *Nat Biotechnol*, 28: 817-25.

———. 2015a. 'Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues', *Nat Biotechnol*, 33: 364.

———. 2015b. 'Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues', *Nat Biotech*, 33: 364-76.

Erwin, G. D., N. Oksenberg, R. M. Truty, D. Kostka, K. K. Murphy, N. Ahituv, K. S. Pollard, and J. A. Capra. 2014. 'Integrating diverse datasets improves developmental enhancer prediction', *PLoS Comput Biol*, 10: e1003677.

Ezkurdia, Iakes, David Juan, Jose Manuel Rodriguez, Adam Frankish, Mark Diekhans, Jennifer Harrow, Jesus Vazquez, Alfonso Valencia, and Michael L. Tress. 2014.

'Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes', *Human Molecular Genetics*, 23: 5866-78.

Fan, Rong-En, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. 'LIBLINEAR: A Library for Large Linear Classification', *J. Mach. Learn. Res.*, 9: 1871-74.

Farh, Kyle Kai-How, Alexander Marson, Jiang Zhu, Markus Kleinewietfeld, William J. Housley, Samantha Beik, Noam Shoresh, Holly Whitton, Russell J. H. Ryan, Alexander A. Shishkin, Meital Hatan, Marlene J. Carrasco-Alfonso, Dita Mayer, C. John Luckey, Nikolaos A. Patsopoulos, Philip L. De Jager, Vijay K. Kuchroo, Charles B. Epstein, Mark J. Daly, David A. Hafler, and Bradley E. Bernstein. 2015. 'Genetic and epigenetic fine mapping of causal autoimmune disease variants', *Nature*, 518: 337-43.

Farnham, P. J. 2009. 'Insights from genomic profiling of transcription factors', *Nat Rev Genet*, 10: 605-16.

Fewings, N. L., P. N. Gatt, F. C. McKay, G. P. Parnell, S. D. Schibeci, J. Edwards, M. A. Basuki, A. Goldinger, M. J. Fabis-Pedrini, A. G. Kermode, C. P. Manrique, J. L. McCauley, D. Nickles, S. E. Baranzini, T. Burke, S. Vucic, G. J. Stewart, and D. R. Booth. 2017. 'The autoimmune risk gene ZMIZ1 is a vitamin D responsive marker of a molecular phenotype of multiple sclerosis', *J Autoimmun*, 78: 57-69.

Fishilevich, S., R. Nudel, N. Rappaport, R. Hadar, I. Plaschkes, T. Iny Stein, N. Rosen, A. Kohn, M. Twik, M. Safran, D. Lancet, and D. Cohen. 2017. 'GeneHancer: genome-wide integration of enhancers and target genes in GeneCards', *Database (Oxford)*, 2017.

Forte, M., B. G. Gold, G. Marracci, P. Chaudhary, E. Basso, D. Johnsen, X. Yu, J. Fowlkes, M. Rahder, K. Stem, P. Bernardi, and D. Bourdette. 2007. 'Cyclophilin D inactivation protects axons in experimental autoimmune encephalomyelitis, an animal model of multiple sclerosis', *Proc Natl Acad Sci U S A*, 104: 7558-63.

Frankish, Adam, Barbara Uszczynska, Graham RS Ritchie, Jose M. Gonzalez, Dmitri Pervouchine, Robert Petryszak, Jonathan M. Mudge, Nuno Fonseca, Alvis Brazma, Roderic Guigo, and Jennifer Harrow. 2015. 'Comparison of GENCODE and RefSeq gene annotation and the impact of reference geneset on variant effect prediction', *BMC Genomics*, 16: S2.

Gallo, P., S. Pagni, M. G. Piccinno, B. Giometto, V. Argentiero, M. Chiusole, F. Bozza, and B. Tavolato. 1992. 'On the role of interleukin-2 (IL-2) in multiple sclerosis (MS). IL-2-mediated endothelial cell activation', *Ital J Neurol Sci*, 13: 65-8.

Gao, T., B. He, S. Liu, H. Zhu, K. Tan, and J. Qian. 2016. 'EnhancerAtlas: a resource for enhancer annotation and analysis in 105 human cell/tissue types', *Bioinformatics*, 32: 3543-51.

Gerstein, M. B., Z. J. Lu, E. L. Van Nostrand, C. Cheng, B. I. Arshinoff, T. Liu, K. Y. Yip, R. Robilotto, A. Rechtsteiner, K. Ikegami, P. Alves, A. Chateigner, M. Perry, M. Morris, R. K. Auerbach, X. Feng, J. Leng, A. Vielle, W. Niu, K. Rhrissorrakrai, A. Agarwal, R. P. Alexander, G. Barber, C. M. Brdlik, J. Brennan, J. J. Brouillet, A. Carr, M. S. Cheung, H. Clawson, S. Contrino, L. O. Dannenberg, A. F. Dernburg, A. Desai, L. Dick, A. C. Dose, J. Du, T. Egelhofer,

S. Ercan, G. Euskirchen, B. Ewing, E. A. Feingold, R. Gassmann, P. J. Good, P. Green, F. Gullier, M. Gutwein, M. S. Guyer, L. Habegger, T. Han, J. G. Henikoff, S. R. Henz, A. Hinrichs, H. Holster, T. Hyman, A. L. Iniguez, J. Janette, M. Jensen, M. Kato, W. J. Kent, E. Kephart, V. Khivansara, E. Khurana, J. K. Kim, P. Kolasinska-Zwierz, E. C. Lai, I. Latorre, A. Leahey, S. Lewis, P. Lloyd, L. Lochovsky, R. F. Lowdon, Y. Lubling, R. Lyne, M. MacCoss, S. D. Mackowiak, M. Mangone, S. McKay, D. Mecenas, G. Merrihew, D. M. Miller, 3rd, A. Muroyama, J. I. Murray, S. L. Ooi, H. Pham, T. Phippen, E. A. Preston, N. Rajewsky, G. Ratsch, H. Rosenbaum, J. Rozowsky, K. Rutherford, P. Ruzanov, M. Sarov, R. Sasidharan, A. Sboner, P. Scheid, E. Segal, H. Shin, C. Shou, F. J. Slack, C. Slightam, R. Smith, W. C. Spencer, E. O. Stinson, S. Taing, T. Takasaki, D. Vafeados, K. Voronina, G. Wang, N. L. Washington, C. M. Whittle, B. Wu, K. K. Yan, G. Zeller, Z. Zha, M. Zhong, X. Zhou, J. Ahringer, S. Strome, K. C. Gunsalus, G. Micklem, X. S. Liu, V. Reinke, S. K. Kim, L. W. Hillier, S. Henikoff, F. Piano, M. Snyder, L. Stein, J. D. Lieb, and R. H. Waterston. 2010. 'Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project', *Science*, 330: 1775-87.

Gilbert, D. M. 2002. 'Replication timing and transcriptional control: beyond cause and effect', *Curr Opin Cell Biol*, 14: 377-83.

Gillies, Stephen D., Sherie L. Morrison, Vernon T. Oi, and Susumu Tonegawa. 1983. 'A tissue-specific transcription enhancer element is located in the major intron of a rearranged immunoglobulin heavy chain gene', *Cell*, 33: 717-28.

Grant, C. E., T. L. Bailey, and W. S. Noble. 2011. 'FIMO: scanning for occurrences of a given motif', *Bioinformatics*, 27: 1017-8.

Hawkins, D. M. 2004. 'The problem of overfitting', *J Chem Inf Comput Sci*, 44: 1-12.

Hayden, E. C. 2014. 'Technology: The $1,000 genome', *Nature*, 507: 294-5.

Hayes, J. E., G. Trynka, J. Vijai, K. Offit, S. Raychaudhuri, and R. J. Klein. 2015. 'Tissue-Specific Enrichment of Lymphoma Risk Loci in Regulatory Elements', *PLoS ONE*, 10: e0139360.

Heintzman, N. D., R. K. Stuart, G. Hon, Y. Fu, C. W. Ching, R. D. Hawkins, L. O. Barrera, S. Van Calcar, C. Qu, K. A. Ching, W. Wang, Z. Weng, R. D. Green, G. E. Crawford, and B. Ren. 2007. 'Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome', *Nat Genet*, 39: 311-8.

Hesselberth, Jay R., Xiaoyu Chen, Zhihong Zhang, Peter J. Sabo, Richard Sandstrom, Alex P. Reynolds, Robert E. Thurman, Shane Neph, Michael S. Kuehn, William S. Noble, Stanley Fields, and John A. Stamatoyannopoulos. 2009. 'Global mapping of protein-DNA interactions in vivo by digital genomic footprinting', *Nature Methods*, 6: 283.

Hindorff, L. A., P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins, and T. A. Manolio. 2009. 'Potential etiologic and functional implications of genome-wide association loci for human diseases and traits', *Proc Natl Acad Sci U S A*, 106: 9362-7.

Hoffman, Michael M., Orion J. Buske, Jie Wang, Zhiping Weng, Jeff A. Bilmes, and William Stafford Noble. 2012. 'Unsupervised pattern discovery in human chromatin structure through genomic segmentation', *Nature Methods*, 9: 473-76.

Holwerda, Sjoerd Johannes Bastiaan, and Wouter de Laat. 2013. 'CTCF: the protein, the binding partners, the binding sites and their chromatin loops', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368: 20120369.

Hong, Eurie L., Cricket A. Sloan, Esther T. Chan, Jean M. Davidson, Venkat S. Malladi, J. Seth Strattan, Benjamin C. Hitz, Idan Gabdank, Aditi K. Narayanan, Marcus Ho, Brian T. Lee, Laurence D. Rowe, Timothy R. Dreszer, Greg R. Roe, Nikhil R. Podduturi, Forrest Tanaka, Jason A. Hilton, and J. Michael Cherry. 2016. 'Principles of metadata organization at the ENCODE data coordination center', *Database: The Journal of Biological Databases and Curation*, 2016: baw001.

Hrdlickova, Barbara, Rodrigo Coutinho de Almeida, Zuzanna Borek, and Sebo Withoff. 2014. 'Genetic variation in the non-coding genome: Involvement of micro-RNAs and long non-coding RNAs in disease', *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, 1842: 1910-22.

International Human Genome Sequencing, Consortium. 2001. 'Initial sequencing and analysis of the human genome', *Nature*, 409: 860.

Iyer, Sowmya. 2015. 'Analysis of genomic data to derive biological conclusions on (1) transcriptional regulation in the human genome and (2) antibody resistance in hepatitis C virus'.

Janky, Rekin's, Annelien Verfaillie, Hana Imrichová, Bram Van de Sande, Laura Standaert, Valerie Christiaens, Gert Hulselmans, Koen Herten, Marina Naval Sanchez, Delphine Potier, Dmitry Svetlichnyy, Zeynep Kalender Atak, Mark Fiers, Jean-Christophe Marine, and Stein Aerts. 2014. 'iRegulon: From a Gene List to a Gene Regulatory Network Using Large Motif and Track Collections', *PLoS Comput Biol*, 10: e1003731.

Johnson, D. S., A. Mortazavi, R. M. Myers, and B. Wold. 2007. 'Genome-wide mapping of in vivo protein-DNA interactions', *Science*, 316: 1497-502.

Jostins, L., S. Ripke, R. K. Weersma, R. H. Duerr, D. P. McGovern, K. Y. Hui, J. C. Lee, L. P. Schumm, Y. Sharma, C. A. Anderson, J. Essers, M. Mitrovic, K. Ning, I. Cleynen, E. Theatre, S. L. Spain, S. Raychaudhuri, P. Goyette, Z. Wei, C. Abraham, J. P. Achkar, T. Ahmad, L. Amininejad, A. N. Ananthakrishnan, V. Andersen, J. M. Andrews, L. Baidoo, T. Balschun, P. A. Bampton, A. Bitton, G. Boucher, S. Brand, C. Buning, A. Cohain, S. Cichon, M. D'Amato, D. De Jong, K. L. Devaney, M. Dubinsky, C. Edwards, D. Ellinghaus, L. R. Ferguson, D. Franchimont, K. Fransen, R. Gearry, M. Georges, C. Gieger, J. Glas, T. Haritunians, A. Hart, C. Hawkey, M. Hedl, X. Hu, T. H. Karlsen, L. Kupcinskas, S. Kugathasan, A. Latiano, D. Laukens, I. C. Lawrance, C. W. Lees, E. Louis, G. Mahy, J. Mansfield, A. R. Morgan, C. Mowat, W. Newman, O. Palmieri, C. Y. Ponsioen, U. Potocnik, N. J. Prescott, M. Regueiro, J. I. Rotter, R. K. Russell, J. D. Sanderson, M. Sans, J. Satsangi, S. Schreiber, L. A. Simms, J. Sventoraityte, S. R. Targan, K. D. Taylor, M. Tremelling, H. W. Verspaget, M. De Vos, C. Wijmenga, D. C. Wilson, J. Winkelmann, R. J. Xavier, S. Zeissig, B. Zhang, C.

K. Zhang, H. Zhao, M. S. Silverberg, V. Annese, H. Hakonarson, S. R. Brant, G. Radford-Smith, C. G. Mathew, J. D. Rioux, E. E. Schadt, M. J. Daly, A. Franke, M. Parkes, S. Vermeire, J. C. Barrett, and J. H. Cho. 2012. 'Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease', *Nature*, 491: 119-24.

Keilwagen, Jens, Stefan Posch, and Jan Grau. 2017. 'Learning from mistakes: Accurate prediction of cell type-specific transcription factor binding', *bioRxiv*.

Keim, D. 2010. "Mastering the Information Age: Solving Problems with Visual Analytics." In, edited by D. Keim, J. Kohlhammer, G. Ellis and F. Mansmann. VisMaster, http://www.vismaster.eu/book/.

Kent, W. J., C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler. 2002. 'The human genome browser at UCSC', *Genome Res*, 12: 996-1006.

Kent, W. J., A. S. Zweig, G. Barber, A. S. Hinrichs, and D. Karolchik. 2010. 'BigWig and BigBed: enabling browsing of large distributed datasets', *Bioinformatics*, 26: 2204-07.

Kim, T. H., Z. K. Abdullaev, A. D. Smith, K. A. Ching, D. I. Loukinov, R. D. Green, M. Q. Zhang, V. V. Lobanenkov, and B. Ren. 2007. 'Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome', *Cell*, 128: 1231-45.

Koprinarova, M., M. Schnekenburger, and M. Diederich. 2016. 'Role of Histone Acetylation in Cell Cycle Regulation', *Curr Top Med Chem*, 16: 732-44.

Kothary, R., S. Clapoff, A. Brown, R. Campbell, A. Peterson, and J. Rossant. 1988. 'A transgene containing lacZ inserted into the dystonia locus is expressed in neural tube', *Nature*, 335: 435.

Kundaje, A., W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z. Zhang, J. Wang, M. J. Ziller, V. Amin, J. W. Whitaker, M. D. Schultz, L. D. Ward, A. Sarkar, G. Quon, R. S. Sandstrom, M. L. Eaton, Y. C. Wu, A. R. Pfenning, X. Wang, M. Claussnitzer, Y. Liu, C. Coarfa, R. A. Harris, N. Shoresh, C. B. Epstein, E. Gjoneska, D. Leung, W. Xie, R. D. Hawkins, R. Lister, C. Hong, P. Gascard, A. J. Mungall, R. Moore, E. Chuah, A. Tam, T. K. Canfield, R. S. Hansen, R. Kaul, P. J. Sabo, M. S. Bansal, A. Carles, J. R. Dixon, K. H. Farh, S. Feizi, R. Karlic, A. R. Kim, A. Kulkarni, D. Li, R. Lowdon, G. Elliott, T. R. Mercer, S. J. Neph, V. Onuchic, P. Polak, N. Rajagopal, P. Ray, R. C. Sallari, K. T. Siebenthall, N. A. Sinnott-Armstrong, M. Stevens, R. E. Thurman, J. Wu, B. Zhang, X. Zhou, A. E. Beaudet, L. A. Boyer, P. L. De Jager, P. J. Farnham, S. J. Fisher, D. Haussler, S. J. Jones, W. Li, M. A. Marra, M. T. McManus, S. Sunyaev, J. A. Thomson, T. D. Tlsty, L. H. Tsai, W. Wang, R. A. Waterland, M. Q. Zhang, L. H. Chadwick, B. E. Bernstein, J. F. Costello, J. R. Ecker, M. Hirst, A. Meissner, A. Milosavljevic, B. Ren, J. A. Stamatoyannopoulos, T. Wang, and M. Kellis. 2015. 'Integrative analysis of 111 reference human epigenomes', *Nature*, 518: 317-30.

Lambert, N., A. Robertson, M. Jangi, S. McGeary, P. A. Sharp, and C. B. Burge. 2014. 'RNA Bind-n-Seq: quantitative assessment of the sequence and structural binding specificity of RNA binding proteins', *Mol Cell*, 54: 887-900.

Law, Philip J., Sonja I. Berndt, Helen E. Speedy, Nicola J. Camp, Georgina P. Sava, Christine F. Skibola, Amy Holroyd, Vijai Joseph, Nicola J. Sunter, Alexandra Nieters, Silvia Bea, Alain Monnereau, David Martin-Garcia, Lynn R. Goldin, Guillem Clot, Lauren R. Teras, Inés Quintela, Brenda M. Birmann, Sandrine Jayne, Wendy Cozen, Aneela Majid, Karin E. Smedby, Qing Lan, Claire Dearden, Angela R. Brooks-Wilson, Andrew G. Hall, Mark P. Purdue, Tryfonia Mainou-Fowler, Claire M. Vajdic, Graham H. Jackson, Pierluigi Cocco, Helen Marr, Yawei Zhang, Tongzhang Zheng, Graham G. Giles, Charles Lawrence, Timothy G. Call, Mark Liebow, Mads Melbye, Bengt Glimelius, Larry Mansouri, Martha Glenn, Karen Curtin, W. Ryan Diver, Brian K. Link, Lucia Conde, Paige M. Bracci, Elizabeth A. Holly, Rebecca D. Jackson, Lesley F. Tinker, Yolanda Benavente, Paolo Boffetta, Paul Brennan, Marc Maynadie, James McKay, Demetrius Albanes, Stephanie Weinstein, Zhaoming Wang, Neil E. Caporaso, Lindsay M. Morton, Richard K. Severson, Elio Riboli, Paolo Vineis, Roel C. H. Vermeulen, Melissa C. Southey, Roger L. Milne, Jacqueline Clavel, Sabine Topka, John J. Spinelli, Peter Kraft, Maria Grazia Ennas, Geoffrey Summerfield, Giovanni M. Ferri, Robert J. Harris, Lucia Miligi, Andrew R. Pettitt, Kari E. North, David J. Allsup, Joseph F. Fraumeni, James R. Bailey, Kenneth Offit, Guy Pratt, Henrik Hjalgrim, Chris Pepper, Stephen J. Chanock, Chris Fegan, Richard Rosenquist, Silvia de Sanjose, Angel Carracedo, Martin J. S. Dyer, Daniel Catovsky, Elias Campo, James R. Cerhan, James M. Allan, Nathanial Rothman, Richard Houlston, and Susan Slager. 2017. 'Genome-wide association analysis implicates dysregulation of immunity genes in chronic lymphocytic leukaemia', *Nat Commun*, 8: 14175.

Lee, D., D. U. Gorkin, M. Baker, B. J. Strober, A. L. Asoni, A. S. McCallion, and M. A. Beer. 2015. 'A method to predict the impact of regulatory variants from DNA sequence', *Nat Genet*, 47: 955-61.

Lee, T. I., and R. A. Young. 2013. 'Transcriptional regulation and its misregulation in disease', *Cell*, 152: 1237-51.

Lek, M., K. J. Karczewski, E. V. Minikel, K. E. Samocha, E. Banks, T. Fennell, A. H. O'Donnell-Luria, J. S. Ware, A. J. Hill, B. B. Cummings, T. Tukiainen, D. P. Birnbaum, J. A. Kosmicki, L. E. Duncan, K. Estrada, F. Zhao, J. Zou, E. Pierce-Hoffman, J. Berghout, D. N. Cooper, N. Deflaux, M. DePristo, R. Do, J. Flannick, M. Fromer, L. Gauthier, J. Goldstein, N. Gupta, D. Howrigan, A. Kiezun, M. I. Kurki, A. L. Moonshine, P. Natarajan, L. Orozco, G. M. Peloso, R. Poplin, M. A. Rivas, V. Ruano-Rubio, S. A. Rose, D. M. Ruderfer, K. Shakir, P. D. Stenson, C. Stevens, B. P. Thomas, G. Tiao, M. T. Tusie-Luna, B. Weisburd, H. H. Won, D. Yu, D. M. Altshuler, D. Ardissino, M. Boehnke, J. Danesh, S. Donnelly, R. Elosua, J. C. Florez, S. B. Gabriel, G. Getz, S. J. Glatt, C. M. Hultman, S. Kathiresan, M. Laakso, S. McCarroll, M. I. McCarthy, D. McGovern, R. McPherson, B. M. Neale, A. Palotie, S. M. Purcell, D. Saleheen, J. M. Scharf, P. Sklar, P. F. Sullivan, J. Tuomilehto, M. T. Tsuang, H. C. Watkins, J. G. Wilson, M. J. Daly, and D. G. MacArthur. 2016. 'Analysis of protein-coding genetic variation in 60,706 humans', *Nature*, 536: 285-91.

Li, H., and R. Durbin. 2009. 'Fast and accurate short read alignment with Burrows-Wheeler transform', *Bioinformatics*, 25: 1754-60.

Li, Qunhua, James B. Brown, Haiyan Huang, and Peter J. Bickel. 2011. 'Measuring reproducibility of high-throughput experiments', *Ann. Appl. Stat.*, 5: 1752-79.

Libbrecht, Maxwell Wing, Oscar Rodriguez, Zhiping Weng, Michael Hoffman, Jeffrey A Bilmes, and William Stafford Noble. 2016. 'A unified encyclopedia of human functional DNA elements through fully automated annotation of 164 human cell types', *bioRxiv*.

Liu, T., J. A. Ortiz, L. Taing, C. A. Meyer, B. Lee, Y. Zhang, H. Shin, S. S. Wong, J. Ma, Y. Lei, U. J. Pape, M. Poidinger, Y. Chen, K. Yeung, M. Brown, Y. Turpaz, and X. S. Liu. 2011. 'Cistrome: an integrative platform for transcriptional regulation studies', *Genome Biol*, 12: R83.

Liu, Yunxian, Ninad M. Walavalkar, Mikhail G. Dozmorov, Stephen S. Rich, Mete Civelek, and Michael J. Guertin. 2017. 'Identification of breast cancer associated variants that modulate transcription factor binding', *PLOS Genetics*, 13: e1006761.

Love, M. I., W. Huber, and S. Anders. 2014. 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2', *Genome Biol*, 15: 550.

MacArthur, J., E. Bowler, M. Cerezo, L. Gil, P. Hall, E. Hastings, H. Junkins, A. McMahon, A. Milano, J. Morales, Z. M. Pendlington, D. Welter, T. Burdett, L. Hindorff, P. Flicek, F. Cunningham, and H. Parkinson. 2017. 'The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog)', *Nucleic Acids Res*, 45: D896-d901.

Machanick, P., and T. L. Bailey. 2011. 'MEME-ChIP: motif analysis of large DNA datasets', *Bioinformatics*, 27: 1696-7.

Marsland, Stephan. 2009. 'Machine Learning: An Algorithmic Perspective'.

Maston, G. A., S. K. Evans, and M. R. Green. 2006. 'Transcriptional regulatory elements in the human genome', *Annu Rev Genomics Hum Genet*, 7: 29-59.

Maurano, M. T., E. Haugen, R. Sandstrom, J. Vierstra, A. Shafer, R. Kaul, and J. A. Stamatoyannopoulos. 2015. 'Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo', *Nat Genet*, 47: 1393-401.

Maurano, M. T., R. Humbert, E. Rynes, R. E. Thurman, E. Haugen, H. Wang, A. P. Reynolds, R. Sandstrom, H. Qu, J. Brody, A. Shafer, F. Neri, K. Lee, T. Kutyavin, S. Stehling-Sun, A. K. Johnson, T. K. Canfield, E. Giste, M. Diegel, D. Bates, R. S. Hansen, S. Neph, P. J. Sabo, S. Heimfeld, A. Raubitschek, S. Ziegler, C. Cotsapas, N. Sotoodehnia, I. Glass, S. R. Sunyaev, R. Kaul, and J. A. Stamatoyannopoulos. 2012. 'Systematic localization of common disease-associated variation in regulatory DNA', *Science*, 337: 1190-5.

Mirabella, Anne C., Benjamin M. Foster, and Till Bartke. 2016. 'Chromatin deregulation in disease', *Chromosoma*, 125: 75-93.

Mohan, Ananth, Zheng Chen, and Kilian Q. Weinberger. 2011. 'Web-Search Ranking with Initialized Gradient Boosted Regression Trees', *Journal of Machine Learning Research, Workshop and Conference Proceedings*, 14: 77-89.

Mohrs, M., C. M. Blankespoor, Z. E. Wang, G. G. Loots, V. Afzal, H. Hadeiba, K. Shinkai, E. M. Rubin, and R. M. Locksley. 2001. 'Deletion of a coordinate regulator of type 2 cytokine expression in mice', *Nat Immunol*, 2: 842-7.

Nguyen, T. A., R. D. Jones, A. R. Snavely, A. R. Pfenning, R. Kirchner, M. Hemberg, and J. M. Gray. 2016. 'High-throughput functional comparison of promoter and enhancer activities', *Genome Res*, 26: 1023-33.

Nierenberg, J., D. F. Salisbury, J. J. Levitt, E. A. David, R. W. McCarley, and M. E. Shenton. 2005. 'Reduced left angular gyrus volume in first-episode schizophrenia', *Am J Psychiatry*, 162: 1539-41.

Niznikiewicz, M., R. Donnino, R. W. McCarley, P. G. Nestor, D. V. Iosifescu, B. O'Donnell, J. Levitt, and M. E. Shenton. 2000. 'Abnormal angular gyrus asymmetry in schizophrenia', *Am J Psychiatry*, 157: 428-37.

Ong, C. T., and V. G. Corces. 2014. 'CTCF: an architectural protein bridging genome topology and function', *Nat Rev Genet*, 15: 234-46.

Park, Daechan, Yaelim Lee, Gurvani Bhupindersingh, and Vishwanath R. Iyer. 2013. 'Widespread Misinterpretable ChIP-seq Bias in Yeast', *PLoS ONE*, 8: e83506.

Patsopoulos, N. A., F. Esposito, J. Reischl, S. Lehr, D. Bauer, J. Heubach, R. Sandbrink, C. Pohl, G. Edan, L. Kappos, D. Miller, J. Montalban, C. H. Polman, M. S. Freedman, H. P. Hartung, B. G. Arnason, G. Comi, S. Cook, M. Filippi, D. S. Goodin, D. Jeffery, P. O'Connor, G. C. Ebers, D. Langdon, A. T. Reder, A. Traboulsee, F. Zipp, S. Schimrigk, J. Hillert, M. Bahlo, D. R. Booth, S. Broadley, M. A. Brown, B. L. Browning, S. R. Browning, H. Butzkueven, W. M. Carroll, C. Chapman, S. J. Foote, L. Griffiths, A. G. Kermode, T. J. Kilpatrick, J. Lechner-Scott, M. Marriott, D. Mason, P. Moscato, R. N. Heard, M. P. Pender, V. M. Perreau, D. Perera, J. P. Rubio, R. J. Scott, M. Slee, J. Stankovich, G. J. Stewart, B. V. Taylor, N. Tubridy, E. Willoughby, J. Wiley, P. Matthews, F. M. Boneschi, A. Compston, J. Haines, S. L. Hauser, J. McCauley, A. Ivinson, J. R. Oksenberg, M. Pericak-Vance, S. J. Sawcer, P. L. De Jager, D. A. Hafler, and P. I. de Bakker. 2011. 'Genome-wide meta-analysis identifies novel multiple sclerosis susceptibility loci', *Ann Neurol*, 70: 897-912.

Pekowska, A., T. Benoukraf, J. Zacarias-Cabeza, M. Belhocine, F. Koch, H. Holota, J. Imbert, J. C. Andrau, P. Ferrier, and S. Spicuglia. 2011. 'H3K4 tri-methylation provides an epigenetic signature of active enhancers', *Embo j*, 30: 4198-210.

Pennacchio, L. A., W. Bickmore, A. Dean, M. A. Nobrega, and G. Bejerano. 2013. 'Enhancers: five essential questions', *Nat Rev Genet*, 14: 288-95.

Pennacchio, Len A., Nadav Ahituv, Alan M. Moses, Shyam Prabhakar, Marcelo A. Nobrega, Malak Shoukry, Simon Minovitsky, Inna Dubchak, Amy Holt, Keith D. Lewis, Ingrid Plajzer-Frick, Jennifer Akiyama, Sarah De Val, Veena Afzal, Brian L. Black, Olivier Couronne, Michael B. Eisen, Axel Visel, and Edward M. Rubin. 2006. 'In vivo enhancer analysis of human conserved non-coding sequences', *Nature*, 444: 499.

Perkel, J. M. 2017. 'Plot a course through the genome', *Nature*, 549: 117-18.

Petretto, Enrico, Rizwan Sarwar, Ian Grieve, Han Lu, Mande K. Kumaran, Phillip J. Muckett, Jonathan Mangion, Blanche Schroen, Matthew Benson, Prakash P.

Punjabi, Sanjay K. Prasad, Dudley J. Pennell, Chris Kiesewetter, Elena S. Tasheva, Lolita M. Corpuz, Megan D. Webb, Gary W. Conrad, Theodore W. Kurtz, Vladimir Kren, Judith Fischer, Norbert Hubner, Yigal M. Pinto, Michal Pravenec, Timothy J. Aitman, and Stuart A. Cook. 2008. 'Integrated genomic approaches implicate osteoglycin (Ogn) in the regulation of left ventricular mass', *Nature genetics*, 40: 546-52.

Phillips, T.; Hoopes L. 2008. 'Transcription factors and transcriptional control in eukaryotic cells', *Nature Education*, 1: 119.

Pihur, Vasyl, Susmita Datta, and Somnath Datta. 2009. 'RankAggreg, an R package for weighted rank aggregation', *BMC Bioinformatics*, 10: 62.

Piper, J., M. C. Elze, P. Cauchy, P. N. Cockerill, C. Bonifer, and S. Ott. 2013. 'Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data', *Nucleic Acids Res*, 41: e201.

Pique-Regi, R., J. F. Degner, A. A. Pai, D. J. Gaffney, Y. Gilad, and J. K. Pritchard. 2011. 'Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data', *Genome Res*, 21: 447-55.

Pruitt, Kim D., Tatiana Tatusova, and Donna R. Maglott. 2005. 'NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins', *Nucleic Acids Res*, 33: D501-D04.

Quinlan, Aaron R. 2014. 'BEDTools: the Swiss-army tool for genome feature analysis', *Curr Protoc Bioinformatics*, 47: 11.12.1-11.12.34.

Rada-Iglesias, A., R. Bajpai, T. Swigut, S. A. Brugmann, R. A. Flynn, and J. Wysocka. 2011. 'A unique chromatin signature uncovers early developmental enhancers in humans', *Nature*, 470: 279-83.

Rajagopal, Nisha, Wei Xie, Yan Li, Uli Wagner, Wei Wang, John Stamatoyannopoulos, Jason Ernst, Manolis Kellis, and Bing Ren. 2013. 'RFECS: A Random-Forest Based Algorithm for Enhancer Identification from Chromatin State', *PLoS Comput Biol*, 9: e1002968.

Rao, S. S., M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander, and E. L. Aiden. 2014. 'A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping', *Cell*, 159: 1665-80.

Rieck, S., and C. Wright. 2014. 'PIQ-ing into chromatin architecture', *Nat Biotechnol*, 32: 138-40.

Rivera-Mulia, J. C., Q. Buckley, T. Sasaki, J. Zimmerman, R. A. Didier, K. Nazor, J. F. Loring, Z. Lian, S. Weissman, A. J. Robins, T. C. Schulz, L. Menendez, M. J. Kulik, S. Dalton, H. Gabr, T. Kahveci, and D. M. Gilbert. 2015. 'Dynamic changes in replication timing and gene expression during lineage specification of human pluripotent stem cells', *Genome Res*, 25: 1091-103.

Romer, K. A., G. R. Kayombya, and E. Fraenkel. 2007. 'WebMOTIFS: automated discovery, filtering and scoring of DNA sequence motifs using multiple programs and Bayesian approaches', *Nucleic Acids Res*, 35: W217-20.

Rosenbloom, Kate R., Joel Armstrong, Galt P. Barber, Jonathan Casper, Hiram Clawson, Mark Diekhans, Timothy R. Dreszer, Pauline A. Fujita, Luvina Guruvadoo, Maximilian Haeussler, Rachel A. Harte, Steve Heitner, Glenn Hickey, Angie S. Hinrichs, Robert Hubley, Donna Karolchik, Katrina Learned, Brian T. Lee, Chin H. Li, Karen H. Miga, Ngan Nguyen, Benedict Paten, Brian J. Raney, Arian F. A. Smit, Matthew L. Speir, Ann S. Zweig, David Haussler, Robert M. Kuhn, and W. James Kent. 2015. 'The UCSC Genome Browser database: 2015 update', *Nucleic Acids Res*, 43: D670-D81.

Roy, S., J. Ernst, P. V. Kharchenko, P. Kheradpour, N. Negre, M. L. Eaton, J. M. Landolin, C. A. Bristow, L. Ma, M. F. Lin, S. Washietl, B. I. Arshinoff, F. Ay, P. E. Meyer, N. Robine, N. L. Washington, L. Di Stefano, E. Berezikov, C. D. Brown, R. Candeias, J. W. Carlson, A. Carr, I. Jungreis, D. Marbach, R. Sealfon, M. Y. Tolstorukov, S. Will, A. A. Alekseyenko, C. Artieri, B. W. Booth, A. N. Brooks, Q. Dai, C. A. Davis, M. O. Duff, X. Feng, A. A. Gorchakov, T. Gu, J. G. Henikoff, P. Kapranov, R. Li, H. K. MacAlpine, J. Malone, A. Minoda, J. Nordman, K. Okamura, M. Perry, S. K. Powell, N. C. Riddle, A. Sakai, A. Samsonova, J. E. Sandler, Y. B. Schwartz, N. Sher, R. Spokony, D. Sturgill, M. van Baren, K. H. Wan, L. Yang, C. Yu, E. Feingold, P. Good, M. Guyer, R. Lowdon, K. Ahmad, J. Andrews, B. Berger, S. E. Brenner, M. R. Brent, L. Cherbas, S. C. Elgin, T. R. Gingeras, R. Grossman, R. A. Hoskins, T. C. Kaufman, W. Kent, M. I. Kuroda, T. Orr-Weaver, N. Perrimon, V. Pirrotta, J. W. Posakony, B. Ren, S. Russell, P. Cherbas, B. R. Graveley, S. Lewis, G. Micklem, B. Oliver, P. J. Park, S. E. Celniker, S. Henikoff, G. H. Karpen, E. C. Lai, D. M. MacAlpine, L. D. Stein, K. P. White, and M. Kellis. 2010. 'Identification of functional elements and regulatory circuits by Drosophila modENCODE', *Science*, 330: 1787-97.

Sanyal, A., B. R. Lajoie, G. Jain, and J. Dekker. 2012. 'The long-range interaction landscape of gene promoters', *Nature*, 489: 109-13.

Savic, D., H. Ye, I. Aneas, S. Y. Park, G. I. Bell, and M. A. Nobrega. 2011. 'Alterations in TCF7L2 expression define its role as a key regulator of glucose metabolism', *Genome Res*, 21: 1417-25.

Schizophrenia Psychiatric Genome-Wide Association Study, Consortium. 2011. 'Genome-wide association study identifies five new schizophrenia loci', *Nat Genet*, 43: 969-76.

Schmidt, D., M. D. Wilson, B. Ballester, P. C. Schwalie, G. D. Brown, A. Marshall, C. Kutter, S. Watt, C. P. Martinez-Jimenez, S. Mackay, I. Talianidis, P. Flicek, and D. T. Odom. 2010. 'Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding', *Science*, 328: 1036-40.

Shen, Y., F. Yue, D. F. McCleary, Z. Ye, L. Edsall, S. Kuan, U. Wagner, J. Dixon, L. Lee, V. V. Lobanenkov, and B. Ren. 2012. 'A map of the cis-regulatory sequences in the mouse genome', *Nature*, 488: 116-20.

Sherwood, Richard I., Tatsunori Hashimoto, Charles W. O'Donnell, Sophia Lewis, Amira A. Barkal, John Peter van Hoff, Vivek Karun, Tommi Jaakkola, and David K. Gifford. 2014. 'Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape', *Nat Biotech*, 32: 171-78.

Sindhwani, Vikas, and S. Sathiya Keerthi. 2006. "Large scale semi-supervised linear SVMs." In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, 477-84. Seattle, Washington, USA: ACM.

Sing, Tobias, Oliver Sander, Niko Beerenwinkel, and Thomas Lengauer. 2005. 'ROCR: visualizing classifier performance in R', *Bioinformatics*, 21: 3940-41.

Sloan, C. A., E. T. Chan, J. M. Davidson, V. S. Malladi, J. S. Strattan, B. C. Hitz, I. Gabdank, A. K. Narayanan, M. Ho, B. T. Lee, L. D. Rowe, T. R. Dreszer, G. Roe, N. R. Podduturi, F. Tanaka, E. L. Hong, and J. M. Cherry. 2016. 'ENCODE data at the ENCODE portal', *Nucleic Acids Res*, 44: D726-32.

Stelzer, Gil, Naomi Rosen, Inbar Plaschkes, Shahar Zimmerman, Michal Twik, Simon Fishilevich, Tsippi Iny Stein, Ron Nudel, Iris Lieder, Yaron Mazor, Sergey Kaplan, Dvir Dahary, David Warshawsky, Yaron Guan-Golan, Asher Kohn, Noa Rappaport, Marilyn Safran, and Doron Lancet. 2002. 'The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses.' in, *Current Protocols in Bioinformatics* (John Wiley & Sons, Inc.).

Stergachis, A. B., S. Neph, A. Reynolds, R. Humbert, B. Miller, S. L. Paige, B. Vernot, J. B. Cheng, R. E. Thurman, R. Sandstrom, E. Haugen, S. Heimfeld, C. E. Murry, J. M. Akey, and J. A. Stamatoyannopoulos. 2013. 'Developmental fate and cellular maturity encoded in human regulatory DNA landscapes', *Cell*, 154: 888-903.

Stolovitzky, G., D. Monroe, and A. Califano. 2007. 'Dialogue on reverse-engineering assessment and methods: the DREAM of high-throughput pathway inference', *Ann N Y Acad Sci*, 1115: 1-22.

Streit, Marc, Alexander Lex, Michael Kalkusch, Kurt Zatloukal, and Dieter Schmalstieg. 2009. 'Caleydo: connecting pathways and gene expression', *Bioinformatics*, 25: 2760-61.

Stunnenberg, H. G., Consortium International Human Epigenome, and M. Hirst. 2016. 'The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery', *Cell*, 167: 1145-49.

Taher, L., N. M. Collette, D. Murugesh, E. Maxwell, I. Ovcharenko, and G. G. Loots. 2011. 'Global gene expression analysis of murine limb development', *PLoS ONE*, 6: e28358.

Teytelman, L., Jasper Thurtle Dm Fau - Rine, Alexander Rine J Fau - van Oudenaarden, and A. van Oudenaarden. 2013. 'Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins'.

Thanos, D., and T. Maniatis. 1995. 'Virus induction of human IFN beta gene expression requires the assembly of an enhanceosome', *Cell*, 83: 1091-100.

Thurman, R. E., E. Rynes, R. Humbert, J. Vierstra, M. T. Maurano, E. Haugen, N. C. Sheffield, A. B. Stergachis, H. Wang, B. Vernot, K. Garg, S. John, R. Sandstrom, D. Bates, L. Boatman, T. K. Canfield, M. Diegel, D. Dunn, A. K. Ebersol, T. Frum, E. Giste, A. K. Johnson, E. M. Johnson, T. Kutyavin, B. Lajoie, B. K. Lee, K. Lee, D. London, D. Lotakis, S. Neph, F. Neri, E. D. Nguyen, H. Qu, A. P. Reynolds, V. Roach, A. Safi, M. E. Sanchez, A. Sanyal, A. Shafer, J. M. Simon, L. Song, S. Vong, M. Weaver, Y. Yan, Z. Zhang, Z. Zhang, B. Lenhard, M.

Tewari, M. O. Dorschner, R. S. Hansen, P. A. Navas, G. Stamatoyannopoulos, V. R. Iyer, J. D. Lieb, S. R. Sunyaev, J. M. Akey, P. J. Sabo, R. Kaul, T. S. Furey, J. Dekker, G. E. Crawford, and J. A. Stamatoyannopoulos. 2012. 'The accessible chromatin landscape of the human genome', *Nature*, 489: 75-82.

Vaknin-Dembinsky, A., K. Balashov, and H. L. Weiner. 2006. 'IL-23 is increased in dendritic cells in multiple sclerosis and down-regulation of IL-23 by antisense oligos increases dendritic cell IL-10 production', *J Immunol*, 176: 7768-74.

Van Nostrand, E. L., G. A. Pratt, A. A. Shishkin, C. Gelboin-Burkhart, M. Y. Fang, B. Sundararaman, S. M. Blue, T. B. Nguyen, C. Surka, K. Elkins, R. Stanton, F. Rigo, M. Guttman, and G. W. Yeo. 2016. 'Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP)', *Nat Methods*, 13: 508-14.

Visel, A., M. J. Blow, Z. Li, T. Zhang, J. A. Akiyama, A. Holt, I. Plajzer-Frick, M. Shoukry, C. Wright, F. Chen, V. Afzal, B. Ren, E. M. Rubin, and L. A. Pennacchio. 2009. 'ChIP-seq accurately predicts tissue-specific activity of enhancers', *Nature*, 457: 854-8.

Visel, A., S. Minovitsky, I. Dubchak, and L. A. Pennacchio. 2007. 'VISTA Enhancer Browser--a database of tissue-specific human enhancers', *Nucleic Acids Res*, 35: D88-92.

Visel, A., S. Prabhakar, J. A. Akiyama, M. Shoukry, K. D. Lewis, A. Holt, I. Plajzer-Frick, V. Afzal, E. M. Rubin, and L. A. Pennacchio. 2008. 'Ultraconservation identifies a small subset of extremely constrained developmental enhancers', *Nat Genet*, 40: 158-60.

Visel, A., E. M. Rubin, and L. A. Pennacchio. 2009. 'Genomic views of distant-acting enhancers', *Nature*, 461: 199-205.

Wang, C., M. Fu, S. Mani, S. Wadler, A. M. Senderowicz, and R. G. Pestell. 2001. 'Histone acetylation and the cell-cycle in cancer', *Front Biosci*, 6: D610-29.

Wang, J., J. Zhuang, S. Iyer, X. Lin, T. W. Whitfield, M. C. Greven, B. G. Pierce, X. Dong, A. Kundaje, Y. Cheng, O. J. Rando, E. Birney, R. M. Myers, W. S. Noble, M. Snyder, and Z. Weng. 2012. 'Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors', *Genome Res*, 22: 1798-812.

Wang, J., J. Zhuang, S. Iyer, X. Y. Lin, M. C. Greven, B. H. Kim, J. Moore, B. G. Pierce, X. Dong, D. Virgil, E. Birney, J. H. Hung, and Z. Weng. 2013. 'Factorbook.org: a Wiki-based database for transcription factor-binding data generated by the ENCODE consortium', *Nucleic Acids Res*, 41: D171-6.

Ward, L. D., and M. Kellis. 2012. 'HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants', *Nucleic Acids Res*, 40: D930-4.

Warne, J., G. Pryce, J. M. Hill, X. Shi, F. Lenneras, F. Puentes, M. Kip, L. Hilditch, P. Walker, M. I. Simone, A. W. Chan, G. J. Towers, A. R. Coker, M. R. Duchen, G. Szabadkai, D. Baker, and D. L. Selwood. 2016. 'Selective Inhibition of the Mitochondrial Permeability Transition Pore Protects against Neurodegeneration in Experimental Multiple Sclerosis', *J Biol Chem*, 291: 4356-73.

Weinberger, D. R., K. F. Berman, and R. F. Zec. 1986. 'Physiologic dysfunction of dorsolateral prefrontal cortex in schizophrenia. I. Regional cerebral blood flow evidence', *Arch Gen Psychiatry*, 43: 114-24.

Wilson, D., V. Charoensawan, S. K. Kummerfeld, and S. A. Teichmann. 2008. 'DBD--taxonomically broad transcription factor predictions: new content and functionality', *Nucleic Acids Res*, 36: D88-92.

Yates, Andrew, Wasiu Akanni, M. Ridwan Amode, Daniel Barrell, Konstantinos Billis, Denise Carvalho-Silva, Carla Cummins, Peter Clapham, Stephen Fitzgerald, Laurent Gil, Carlos García Girón, Leo Gordon, Thibaut Hourlier, Sarah E. Hunt, Sophie H. Janacek, Nathan Johnson, Thomas Juettemann, Stephen Keenan, Ilias Lavidas, Fergal J. Martin, Thomas Maurel, William McLaren, Daniel N. Murphy, Rishi Nag, Michael Nuhn, Anne Parker, Mateus Patricio, Miguel Pignatelli, Matthew Rahtz, Harpreet Singh Riat, Daniel Sheppard, Kieron Taylor, Anja Thormann, Alessandro Vullo, Steven P. Wilder, Amonida Zadissa, Ewan Birney, Jennifer Harrow, Matthieu Muffato, Emily Perry, Magali Ruffier, Giulietta Spudich, Stephen J. Trevanion, Fiona Cunningham, Bronwen L. Aken, Daniel R. Zerbino, and Paul Flicek. 2016. 'Ensembl 2016', *Nucleic Acids Res*, 44: D710-D16.

Yue, F., Y. Cheng, A. Breschi, J. Vierstra, W. Wu, T. Ryba, R. Sandstrom, Z. Ma, C. Davis, B. D. Pope, Y. Shen, D. D. Pervouchine, S. Djebali, R. E. Thurman, R. Kaul, E. Rynes, A. Kirilusha, G. K. Marinov, B. A. Williams, D. Trout, H. Amrhein, K. Fisher-Aylor, I. Antoshechkin, G. DeSalvo, L. H. See, M. Fastuca, J. Drenkow, C. Zaleski, A. Dobin, P. Prieto, J. Lagarde, G. Bussotti, A. Tanzer, O. Denas, K. Li, M. A. Bender, M. Zhang, R. Byron, M. T. Groudine, D. McCleary, L. Pham, Z. Ye, S. Kuan, L. Edsall, Y. C. Wu, M. D. Rasmussen, M. S. Bansal, M. Kellis, C. A. Keller, C. S. Morrissey, T. Mishra, D. Jain, N. Dogan, R. S. Harris, P. Cayting, T. Kawli, A. P. Boyle, G. Euskirchen, A. Kundaje, S. Lin, Y. Lin, C. Jansen, V. S. Malladi, M. S. Cline, D. T. Erickson, V. M. Kirkup, K. Learned, C. A. Sloan, K. R. Rosenbloom, B. Lacerda de Sousa, K. Beal, M. Pignatelli, P. Flicek, J. Lian, T. Kahveci, D. Lee, W. J. Kent, M. Ramalho Santos, J. Herrero, C. Notredame, A. Johnson, S. Vong, K. Lee, D. Bates, F. Neri, M. Diegel, T. Canfield, P. J. Sabo, M. S. Wilken, T. A. Reh, E. Giste, A. Shafer, T. Kutyavin, E. Haugen, D. Dunn, A. P. Reynolds, S. Neph, R. Humbert, R. S. Hansen, M. De Bruijn, L. Selleri, A. Rudensky, S. Josefowicz, R. Samstein, E. E. Eichler, S. H. Orkin, D. Levasseur, T. Papayannopoulou, K. H. Chang, A. Skoultchi, S. Gosh, C. Disteche, P. Treuting, Y. Wang, M. J. Weiss, G. A. Blobel, X. Cao, S. Zhong, T. Wang, P. J. Good, R. F. Lowdon, L. B. Adams, X. Q. Zhou, M. J. Pazin, E. A. Feingold, B. Wold, J. Taylor, A. Mortazavi, S. M. Weissman, J. A. Stamatoyannopoulos, M. P. Snyder, R. Guigo, T. R. Gingeras, D. M. Gilbert, R. C. Hardison, M. A. Beer, and B. Ren. 2014. 'A comparative encyclopedia of DNA elements in the mouse genome', *Nature*, 515: 355-64.

Zacher, Benedikt, Margaux Michel, Björn Schwalb, Patrick Cramer, Achim Tresch, and Julien Gagneur. 2017. 'Accurate Promoter and Enhancer Identification in 127

ENCODE and Roadmap Epigenomics Cell Types and Tissues by GenoSTAN', *PLoS ONE*, 12: e0169249.

Zerbino, D. R., S. P. Wilder, N. Johnson, T. Juettemann, and P. R. Flicek. 2015. 'The ensembl regulatory build', *Genome Biol*, 16: 56.

Zhang, Y., T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li, and X. S. Liu. 2008. 'Model-based analysis of ChIP-Seq (MACS)', *Genome Biol*, 9: R137.

Zhou, X., B. Maricque, M. Xie, D. Li, V. Sundaram, E. A. Martin, B. C. Koebbe, C. Nielsen, M. Hirst, P. Farnham, R. M. Kuhn, J. Zhu, I. Smirnov, W. J. Kent, D. Haussler, P. A. Madden, J. F. Costello, and T. Wang. 2011. 'The Human Epigenome Browser at Washington University', *Nat Methods*, 8: 989-90.

X. Chapter VIII: Fin

Never trust to general impressions…
but concentrate yourself upon details.

–Sherlock Holmes, A Case of Identity