2017-10-10

Defining a Registry of Candidate Regulatory Elements to Interpret Disease Associated Genetic Variation

Jill E. Moore

University of Massachusetts Medical School

DEFINING A REGISTRY OF CANDIDATE REGULATORY ELEMENTS TO INTERPRET DISEASE ASSOCIATED GENETIC VARIATION

A Dissertation Presented

By

Jill E. Moore

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

OCTOBER 10, 2017

BIOINFORMATICS & COMPUTATIONAL BIOLOGY

DEFINING A REGISTRY OF CANDIDATE REGULATORY ELEMENTS TO INTERPRET DISEASE ASSOCIATED GENETIC VARIATION

A Dissertation Presented
By

Jill E. Moore

This work was undertaken in the Graduate School of Biomedical Sciences

Bioinformatics and Computational Biology

Under the mentorship of

Zhiping Weng, Ph.D., Thesis Advisor

Manuel Garber, Ph.D., Member of Committee

Konstantin Zeldovich, Ph.D., Member of Committee

Elinor Karlsson, Ph.D., Member of Committee

Mark Gerstein, Ph.D., External Member of Committee

Jeffrey Bailey, MD, Ph.D., Chair of Committee

Anthony Carruthers, Ph.D., Dean of the Graduate School of Biomedical Science

October 10, 2017

DEDICATION

This thesis is dedicated to my husband Eric LaRocque, the EP300 to my RNA POLII, enhancing every aspect of life.

ACKNOWLEDGEMENTS

This thesis would not have been possible without guidance from my advisor, the hard work and support of my peers, and the patience and love of my friends and family. I first would like to acknowledge fellow members of the "SCREEN team": Michael Purcaro and Henry Pratt. Over the last year we have worked tirelessly together to assemble the Registry and SCREEN. I could not have asked for two better partners, though I will not miss our 9pm Skype calls. All our hard work has paid off I know we will all celebrate when the paper *eventually* comes out. I would also like to acknowledge Zlab members as a whole for their support and friendship. I have thoroughly enjoyed our time spent out of lab, whether at Armsby or exploring after conferences.

Third, I would like to acknowledge my "D4L" family. We started on a journey together in the fall of 2008 at the University of Massachusetts Amherst on a floor dedicated to ~~nerds~~ math and science. We have been inseparable ever since, which has enabled us to become the best architects, teachers, doctors, physician assistants, police officers, engineers, and computer scientists that we can be. I am forever thankful for the friendship that we all share.

Fourth, I would like to acknowledge my family. To my parents Donna and Robert, thank you for always supporting me. Thank you, Mom, for always lending an ear to vent my frustrations whether about the lab or life. Thank you, Dad, for instilling an appreciation for math and science in me at a young age and for the dimple chin. To my brothers, Mike and Ricky, thank you for preparing me for a

career in academia. Your playful (and sometimes not so playful) jabs growing up have readied me for the rejected grants, papers, and positions that I will undoubtedly face throughout my career. I am also very proud of what you have both accomplished at such a young age. A scientist, a lawyer, and a captain walk into a bar.... I am sure there is a joke somewhere in there. I also thank my husband Eric who has always encouraged me to be the best scientist and person that I can be. His unconditional patience and love throughout my PhD career have kept me sane and grounded, even while trying to plan a wedding.

Finally, I would like to acknowledge and thank my advisor over the last five years, Zhiping Weng. After meeting Zhiping during my prospective students weekend I knew that I wanted to work in her lab. Not only does she have a passion for research, but also for learning. She has instilled in me the importance of continually improving yourself, to reach your highest potential. I look forward to continuing to work together over the next several years.

ABSTRACT

Over the last decade there has been a great effort to annotate noncoding regions of the genome, particularly those that regulate gene expression. These regulatory elements contain binding sites for transcription factors (TF), which interact with one another and transcriptional machinery to initiate, enhance, or repress gene expression. The Encyclopedia of DNA Elements (ENCODE) consortium has generated thousands of epigenomic datasets, such as DNase-seq and ChIP-seq experiments, with the goal of defining such regions. By integrating these assays, we developed the Registry of candidate Regulatory Elements (cREs), a collection of putative regulatory regions across human and mouse. In total, we identified over 1.3M human and 400k mouse cREs each annotated with cell-type specific signatures (e.g. promoter-like, enhancer-like) in over 400 human and 100 mouse biosamples. We then demonstrated the biological utility of these regions by analyzing cell type enrichments for genetic variants reported by genome wide association studies (GWAS). To search and visualize these cREs, we developed the online database SCREEN (search candidate regulatory elements by ENCODE). After defining cREs, we next sought to determine their potential gene targets. To compare target gene prediction methods, we developed a comprehensive benchmark of enhancer-gene links by curating ChIA-PET, Hi-C and eQTL datasets. We then used this benchmark to evaluate unsupervised linking approaches such as the correlation of epigenomic signal. We determined that these methods have low overall performance and do not outperform simply selecting the closest gene. We then developed a supervised Random Forest model which had notably better performance than unsupervised methods. We demonstrated that this model can be applied across cell types and can be used to predict target genes for GWAS associated variants. Finally, we used the registry of cREs to annotate variants associated with psychiatric disorders. We found that these "psych SNPs" are enriched in cREs active in brain tissue and likely target genes involved in neural development pathways. We also demonstrated that psych SNPs overlap binding sites for TFs involved in neural and immune pathways. Finally, by identifying psych SNPs with allele imbalance in chromatin accessibility, we highlighted specific cases of psych SNPs altering TF binding motifs resulting in the disruption of TF binding. Overall, we demonstrated our collection of putative regulatory regions, the Registry of cREs, can be used to understand the potential biological function of noncoding variation and develop hypotheses for future testing.

TABLE OF CONTENTS

CHAPTER III: EVALUATING METHODS FOR LINKING ENHANCERS WITH TARGET GENES

CHAPTER V: CONCLUSIONS AND FUTURE DIRECTIONS

LIST OF FIGURES

LIST OF EXTERNAL FIGURES

These figure is too large to include in the PDF version of this thesis. It is available online at https://drive.google.com/drive/folders/0B07orkTYRj9pRy1IdE9JUVVYTzA?usp=sharing

LIST OF TABLES

LIST OF EXTERNAL TABLES

These tables are too large to include in the PDF version of this thesis. They are available online at https://drive.google.com/drive/folders/0B07orkTYRj9pRy1IdE9JUVVYTzA?usp=sharing

CHAPTER I: Introduction

INTRODUCTION

The human genome is comprised of three billion base pairs, of which less than 2% are protein-coding exons[1]. The remaining 98% of the genome contains introns, noncoding RNAs, pseudogenes, repeat sequences, transposons, and regulatory elements[1]. Properly annotating these regions is difficult and requires the integration of multiple genomic, transcriptomic, and epigenomic assays. Accurate characterization of these noncoding sequences has important biomedical applications. For example, mutations in regulatory elements are linked with monogenic disorders such as polydactyly[2], cleft palate[3], and congenital heart disease[4,5]. Additionally, 80% of common genetic variants associated with human disease are in noncoding regions with many overlapping potential regulatory elements[6]. Therefore, to better understand mutations linked with human disease and genome regulation as a whole, we aim to produce a comprehensive annotation of regulatory elements in the human genome.

PREDICTION AND VALIDATION OF REGULATORY ELEMENTS

Regulatory elements in the human genome

Regulatory elements are regions of DNA where proteins known as transcription factors (TFs) bind and interact with one another and transcriptional machinery to control gene expression[1,7]. While the genome likely has many different types of regulatory elements, with varying degrees of activity, the field often classifies them into four generalized categories: promoters, enhancers, repressors, and insulators. Promoter elements are proximal to transcriptional start sites (TSSs) of genes and are responsible for recruiting TFs and polymerase machinery to initiate transcription[7,8]. The core promoter is the minimal set of DNA elements required for transcription to occur. In humans, this includes the TSS, RNA polymerase binding site and general TF binding sites such as the TATA Box[8]. In addition to the core promoter, there are also proximal promoter elements where additional TFs bind to further promoter transcription[7].

Enhancers are regulatory elements that increase gene expression through TF interactions with a gene's promoter[7,9,10]. Initially discovered in viral genomes[11,12], enhancers were first characterized in a mammalian genome by Banerji *et al.*, who identified a lymphocyte specific enhancer within the Ig gene[13]. Enhancers can be located close to their target gene, like the Ig enhancer, or can be almost 1 Mb away such as the polydactyly linked ZRS element and its target SHH[2,14]. Generally, enhancers are more cell type specific than promoters and are responsible for regulating cell type specific gene expression[15].

Repressors (also referred to as silencers) are elements that suppress transcription[7,16]. Though repressors are not as well characterized as their enhancer counterparts, several examples have been identified in the human genome. For example, neuron-restrictive silencer elements (NRSEs) are regions that repress the transcription of neuronal genes in non-neuronal cells[17]. These elements have been reported near many neuronal genes such as *STMN2*

(SCG10), which is involved in neuronal growth[18] and *SCN2A*, which encodes a voltage-gated sodium channel[19]. NRSEs contain binding sites for the REST (NRSF), a repressive TF[20].

Insulators are elements that block transcriptional regulation between two genomic regions (often nearby genes)[7]. A well-studied example of insulator activity in the human genome is at the imprinting control region (IRC) located between the H19 and IGF2 genes[21,22]. On the paternal allele, the IRC is methylated, preventing TF binding, and IGF2 is expressed. However, on the maternal allele, the IRC is unmethylated, allowing the TF CTCF to bind and block IGF2 from interacting with its upstream enhancer[21,22]. While the exact biochemical mechanism of insulator activity is not well understood, it appears that CTCF is an important component of insulators[23].

The aforementioned examples of regulatory elements were characterized over a period of years using various biochemical and validation assays. The majority of these methods are low throughput and attempting to apply these methods on a genome wide scale is not practical. With the advent of sequencing technology and high throughput genomic assays, we can predict regions of regulatory activity in the human genome. The National Human Genome Research Institute established The Encyclopedia of DNA Elements (ENCODE) Consortium[24,25], a large collaborative effort, with the aim of identifying all of the functional elements in the human genome. Through technological development and collaborative analysis, the ENCODE project, along with other large consortia such as the Roadmap Epigenomics Project[26], has generated thousands of high throughput sequencing datasets which can be integrated to define an encyclopedia of DNA elements.

Assays used to characterize regulatory elements

The ENCODE and Roadmap Epigenomics projects have generated thousands of high throughput sequencing experiments that assay components of the genome, transcriptome, and epigenome. Using these datasets, we can identify genomic regions bound by TFs and interpret their chromatin context to identify putative regulatory elements.

ENCODE has generated over three thousand TF chromatin immunoprecipitation sequencing (ChIP-seq) experiments, identifying binding sites for hundreds of TFs in up to hundreds of cell types. By analyzing the binding sites of specific TFs, we can identify regions with potential regulatory function. For example, POLR2A is a subunit of RNA polymerase II (POLII) and co-localizes with regions of transcriptional activity such as promoters[27]. EP300 (p300), on the other hand, is known to bind at enhancers[28,29]. Other TFs, such as the previously mentioned CTCF, have multiple functions and are known to bind at enhancers, repressors, and insulators[30]. In addition to assaying specific TFs, the ENCODE and Roadmap projects have generated chromatin accessibility data, such as DNase-seq[31,32] and ATAC-seq (assay for transposase-accessible chromatin using sequencing)[33]**.** These assays identify regions of open chromatin (called DNase hypersensitivity sites (DHS) and ATAC peaks) presumably due to TF binding. While these methods do not report the specific TFs bound at these regions, they produce a genome-wide list of potential regulatory sites.

We can further annotate TF binding sites and DHSs using histone modification ChIP-seq data. Histone proteins are core components of nucleosomes, which condense and package DNA into higher order chromatin structures[34]. Chromatin structure is regulated by biochemical modifications to histones, particularly to their tails, causing chromatin to relax or tighten. Therefore, specific types of histone modifications are enriched at regulatory elements. For example, promoters tend to have high levels of H3K4me3 (histone 3, lysine 4, tri-methylation) and H3K27ac (histone 3, lysine 27, acetylation) and enhancers have high levels of H3K4me1 (histone 3, lysine 4, mono-methylation) and H3K27ac[15,28]. Repressed regions tend to have high levels of H3K27me3 (histone 3, lysine 27, tri-methylation) and H3K9me3 (histone 3, lysine 9, tri-methylation)[35]. Therefore, by integrating these different types of epigenomic datasets, we can begin to define regulatory elements in the human genome at a genome-wide scale.

Computational methods for predicting enhancers

Because of their association with human disease and estimated abundancy in the genome[15], labs have primarily focused on identifying enhancers. While EP300 binding sites have been shown to successfully predict functional enhancers[29], ENCODE has only produced EP300 ChIP-seq data for a small set of cell types. Therefore, to identify enhancers, we need to rely on computational methods that utilize other types of epigenomic datasets.

Over the last several years many labs have developed computational methods for identifying potential enhancers – a simple literature search for "enhancer prediction" generates hundreds of results. Most of these methods integrate histone modification ChIP-seq data to predict candidate enhancers, sometimes incorporating additional features such as DNA methylation[36] or conservation[37]. Most top performing methods use supervised machine learning algorithms and require a known set of enhancers for training. Because there are few large collections of experimentally validated enhancers, many of these methods use complementary epigenomic data (e.g., EP300[36,38] or H3K27ac ChIP-seq[37]) for their gold standard. For example, both RFECS (Random Forest based Enhancer identification from Chromatin States)[38] and REPTILE (Regulatory Element Prediction Based on Tissue-specific Local Epigenetic Marks)[36], two random forest based approaches, train on cell type specific EP300 binding sites. Therefore, predictions from these methods may be biased; they identify false positives due to spurious non-functional EP300 binding or ChIP-seq noise or fail to identify classes of enhancers lacking EP300 binding (false negatives). Additionally, while RFECS and REPTILE have high performance, they require data from many different experiments. For optimal performance, RFECS uses 24 histone modification ChIP-seq datasets as input; few cell types have this many assayed histone modifications. While the method can be modified to use just core marks H3K4me1, H3K27ac, and H3K4me3, this results in decreased performance. REPTILE also requires extensive types of data with cell type specific DNA methylation and six histone modification ChIP-seq datasets in addition to "comparative deviation values" for these assays calculated across multiple cell types. Once again, only a small percentage of cell types surveyed by the ENCODE and Roadmap Epigenomics projects have all of these assays. While He et al. demonstrated that REPTILE can be trained on H1-hESCs and validated on other cell types, REPTILEs performance drops dramatically. Therefore, while supervised methods have high performance, they do not train and test on functionally validated ennhancers and are not applicable to the majority of cell types.

An alternative approach to supervised enhancer prediction are chromatin segmentation methods such as ChromHMM[39,40] and Segway[41]. These methods integrate multiple genomic signals and assign every position in the genome to a chromatin state (e.g., TSS, enhancer, repressed). For example, ChromHMM takes a binarized matrix of histone modification signals for 200 bp bins and using a Hidden Markov model, assigns each bin to a state based on the combinations of histone modifications and emission probabilities from neighboring regions. Because these methods are unsupervised, the user must designate the number of states before running the model and then manually label each state using complementary data and genetic annotations. Chromatin segmentation methods are advantageous because they do not require a gold-standard and they annotate regions other than enhancers such as promoters and insulators. They can also identify sub-classes of elements such as strong and weak enhancers based on different combinations of signals. However, like some supervised methods, both ChromHMM and Segway require a lot of input data (at least eight histone modifications) so these methods are not applicable across all cell types.

In addition to using epigenomic datasets, enhancers can be identified by assaying transcriptional activity. In 2010, the Greenberg lab observed bidirectional transcription at cortical neuron enhancers that resulted in the production of small noncoding RNAs, which they called enhancer RNAs (eRNA)[42]. They hypothesized that these transcripts were the result of interactions between the enhancer and target promoter during the transfer of POLII. Because of these findings, groups have predicted enhancers by identifying regions of transcription that are distal to annotated TSSs. For example, the FANTOM5 consortium used cap analysis of gene expression (CAGE) data to identify distal regions of bidirectional transcription that they believed were candidate enhancers[43]. In total, they identified more than 43 thousand candidate enhancers for over 800 cell and tissue types. Similarly, the Siepal lab developed a computation method, discriminative regulatory-element detections from GRO-seq (dREG), to identify transcriptional regulatory elements (TREs) using GRO-seq data in eight cell types[44]. Overall, these methods tend to identify fewer candidate enhancers than epigenomic based methods, suggesting that either we currently only have the resolution to identify a subset of enhancers using transcriptional activity or that not all enhancers produce eRNA.

Methods for experimentally validating enhancer predictions

In order to properly validate computation methods for enhancer prediction, candidate regions should be experimentally tested for enhancer activity. Reporter assays, such as those used by Banerji et al. to identify the Ig enhancers[13], test enhancer activity by detecting the expression of a reporter gene[9]. Candidate enhancers are cloned into a plasmid upstream of a minimal promoter and a reporter gene such as luciferase or GFP. Then, these plasmids are transfected into cells; if the candidate region has enhancer activity, the cell will test positive for gene expression (e.g., via luminescence). While effective, this method can only test one candidate enhancer per experiment, so validating the thousands of predictions generated by most methods would be infeasible.

To solve this problem, researchers have modified these methods to test thousands of regions in one experiment[45-48]. In massively parallel reporter assays (MRPAs), each candidate enhancer is assigned a unique bar code. When these regions are cloned into the plasmid construct, the enhancer is positioned upstream of the TSS and the barcode is downstream of the reporter gene's open reading frame. After co-transfection, enhancer activity is quantified by sequencing the produced mRNAs and computing the relative abundance of each tag. These methods can be used to systematically investigate how genetic variation affects enhancer activity[45,46] or to identify sequences with enhancer activity from a synthetic library[47]. Similarly, STARR-seq (self-transcribing active regulatory region sequencing) developed by the Stark Lab, simultaneously tests thousands of genomic sequences for enhancer activity[48]. STARR-seq uses a gene construct where the candidate enhancer lies between the open reading frame and poly-A site of the reporter gene. Therefore, if the tested region has enhancer activity, the gene is transcribed and the resulting mRNA contains the enhancer sequence. Computationally, this sequence can be isolated and mapped directly to the genome. Enhancers are then identified by scanning the genome for peaks of mRNA signal.

While MPRA and STARR-seq methods allow one to test thousands of candidate enhancers simultaneously, each experiment is limited to testing in one cell type. Mouse transgenic assays, on the other hand, allow researchers to determine the tissue specificity of an enhancer[29,49,50]. In these experiments, a candidate enhancer is cloned into a minimal promoter construct containing the lacZ reporter gene. This construct is microinjected into a fertilized mouse egg which is then implanted in a female mouse. After 11.5 days of embryonic development (e11.5) embryos are harvested and stained to assay enhancer activity. Tissues in which the enhancer is active will stain blue. With these assays, the Pennacchio and Visel labs have tested hundreds of candidate enhancers, the results of which are in the VISTA enhancer database[51].

A major pitfall for all of these methods is that they do not assay enhancer activity in the candidate region's native chromatin context[9]. For reporter assays, MPRA, and STARR-seq, plasmids are transiently transfected into cells and therefore will have no chromatin organization. For the mouse transgenic assays, the constructs are stably transfected and will integrate with the genome. However, the location of integration differs from the enhancer's original position, and therefore local chromatin context and interactions may vary.

In order to study enhancer activity in its native environment, labs have utilized CRISPR-Cas9 genome editing methods. Using region specific guide RNAs, labs can delete candidate enhancers and perform RNA-seq to determine the effect of the deletion on gene expression[52-54]. Additionally, tiling approaches such as MERA (multiplexed editing regulatory assay) and CREST-seq (*cis*-regulatory element scan by tiling-deletion and sequencing) allow researchers to investigate large stretches of DNA to observe how systematically editing these regions effects gene expression[55,56]. With these methods, enhancer activity is measured by the decrease in target gene expression. This may not be a direct measure of enhancer activity due to regulatory elements that compensate for the deleted enhancer.

LINKING DISTAL ENHANCERS WITH TARGET GENES

Though labs have developed experimental and computational methods for identifying candidate enhancers, determining the genes they regulate (i.e., target genes) still remains a challenge. Some enhancers target their nearest gene, such as the intronic Ig enhancers. However other enhancers, such as the ZRS element, target genes up to 1 Mb away[2]. Currently, researchers utilize both experimental and computational approaches for predicting target genes.

Experimental methods for to assaying three-dimensional chromatin structure

Capture chromosome conformation, commonly referred to as 3C, is a method used to investigate interacting genomic loci[57]. Briefly, 3C cross-links interacting regions of chromatin then digests and ligates the ends of the DNA regions. To quantify the amount of ligated product, one performs qPCR using region specific primers. Though this method is limited by one knowing ahead of time which loci to test, it has been used to successfully characterize interactions between regulatory elements. For example, Tolhuis et al. used 3C to characterize enhancer-promoter interactions at the mouse Beta Globin Locus, concluding that DHSs in the locus control region interact with actively expressed genes[58].

3C has been modified to identify interacting loci without knowing them *a priori*. 4C (circular chromosome conformation capture/chromosome conformation capture–on-chip)[59,60] for example, allows one to investigate all possible interactions with region of interest (referred to as "one to all" approach), while 5C reports all interactions within a given region (up to several Mb) (referred to as a "many to many" approach)[61]. Hi-C is the most high-throughput approach reporting chromatin interactions on a genome wide scale[62]. Until recently, the resolution of Hi-C was not precise enough to capture the majority of enhancer-promoter interactions. However, in 2014, the Aiden lab generated kilobase resolution *in situ* Hi-C data and demonstrated that they could identify CTCF anchored chromatin loops with a resolution of up to 5 kb[63]. In GM12878 they identified almost ten thousand loops and determined they were enriched for promoter-enhancer interactions. These loops have subsequently been used to predict enhancer-promoter interactions and train computational models (such as those mentioned below). However, recent preprint data from the Aiden lab suggest that disrupting these loops by knocking out a component of the cohesion complex has little effect on gene expression[64]. A variation on the Hi-C method, promoter capture Hi-C (CHi-C), enriches for promoter interactions by adding a hybridization step before sequencing, which enriches for known promoter sequences[65]. CHi-C results in a ten-fold increase for promoter reads and therefore can be used to link candidate enhancers with these regions.

Another widely used method for investigating genomic interactions is chromatin interaction analysis by paired-end tag sequencing (ChIA-PET)[66]. ChIA-PET follows a similar protocol to Hi-C except an antibody is used to select for interactions mediated by a specific protein such as CTCF, RAD21 or POLII. Using ChIA-PET, researchers have discovered distinct biological features associated with the interactions mediated by these proteins. For example, the Ruan lab reported significant biological differences between interactions generated by CTCF ChIA-PET data compared to POLII ChIA-PET data[67]. Enhancers and promoters in the CTCF interactions were much farther apart (as we also report in Chapter III) and genes anchored at these CTCF loops were more likely to be house-keeping genes, ubiquitously expressed across cell types. In contrast, most POLII interactions occurred within CTCF loops and overlapped cell type specific enhancers. From these results, the Ruan lab concluded that CTCF brings together distant regulatory elements to form ubiquitous large domains and within these domains enhancers and promoters interact to create local, context specific loops.

Computational methods for linking enhancers and genes

While three-dimensional chromatin assays have been demonstrated to successfully identify interacting enhancers and promoters, these experiments (Hi-C in particular) are extremely expensive and have only been performed in a small number of cell types. Therefore, labs have developed computational methods for linking enhancers with target genes. In several publications, the members of the ENCODE consortium have predicted target genes by correlating epigenomic and transcriptomic signals. This method is based on the hypothesis that enhancers are active in the same cell types in which their target genes are expressed. To link enhancers with genes, groups correlated a range of signals such as DNase[6,68,69], H3K4me1[70], POLII[70], and RNA-seq[69]. Each of these studies reported biologically relevant enhancer-promoter pairs with high correlation, but these methods have yet to be systematically evaluated using a gold standard. These methods are attractive because they are unsupervised and do not require cell-type specific experiments, but they have several major drawbacks. One, correlation methods cannot identify cell type interactions. Enhancers tend to be more cell type specific than promoters, and their activity will not correlate with their target gene if the gene is regulated by other enhancers in different cell types. Second, correlation based methods are dependent on the breadth of data selected for the analysis. For example, performance may change depending on the number and type of samples (e.g., tissues, primary cells, cell lines) used for correlation. To determine the severity these problems, we need to evaluate correlation based methods systematically.

Like signal correlation, PreSTIGE (Predicting Specific Tissue Interactions of Gene and Enhancers), developed by the Scacheri lab, is an unsupervised enhancer-gene linking method[71]. PreSTIGE predicts enhancer-promoter interactions by integrating ChIP-seq and RNA-seq data. First, PreSTIGE identifies cell type specific enhancers and genes by calculating the Shannon entropy, using H3K4me1 and RNA-seq respectively, across 12 cell types. PreSTIGE then links these cell type specific enhancers and genes using a linear domain model that considers both distance and CTCF binding sites. While this method predicts many enhancer-gene pairs with disease relevant applications, it only can identify links for cell-type specific genes. Genes that are ubiquitously expressed but are regulated by cell-type specific enhancers are not included in this analysis.

To address some of the problems with unsupervised prediction methods, labs have developed supervised learning approaches for linking enhancers with their target genes[72-74]. These methods train on enhancer-gene pairs defined using ChIA-PET or Hi-C datasets and use many different types of features (e.g., histone modification signal, conservation, distance, correlation, expression) to predict pairs. While each of these methods reported high performance in their original results, it is almost impossible to fairly compare them because they each use a different gold standard for training and testing. To illustrate this fact, we will compare three recently published methods: PETModule[73], IM-PET[72], and TargetFinder[74]. First, the methods used different definitions for enhancers. PETModule used EP300 peaks, IM-PET used the CSI-ANN algorithm with H3K4me1, H3K4me3, and H3K27ac data, and TargetFinder used consensus enhancer states called by ChromHMM and Segway. Second, each method used different three-dimensional chromatin data to generate enhancer-gene pairs. PETModule and IM-PET both used ChIA-PET data from Li et al.[66] but PETModule also included Hi-C data from Jin et al.[75] TargetFinder exclusively used Hi-C loop data from Rao et al.[63]. Finally, each method used a different scheme for generating negative enhancer-gene pairs. For each enhancer, PETModule labeled all genes within a 2 Mb window as negatives if they did not share a ChIA-PET or Hi-C link. IM-PET selected negative pairs by matching a contact frequency distribution, selecting one non-interacting promoter per enhancer. For its negative set, TargetFinder generated 20 non-interacting pairs by randomly selecting gene pairs to match the distance distribution between enhancer and promoter. Therefore, to accurately compare the performance of these methods, as well as unsupervised methods, we need to evaluated them using a consistent benchmark dataset.

NONCODING VARIATION ASSOCIATED WITH PSYCHIATRIC DISORDERS

Genetics of Psychiatric disorders

Schizophrenia, bipolar disorder, and major depressive disorder are prevalent, debilitating psychiatry disorders and little is known about their etiologies. Schizophrenia (SCZ) is characterized by a series of positive symptoms (e.g. delusions, hallucinations, and disorganized speech and behavior) and negative symptoms (e.g. diminished emotional expression and avolition)[76]. It affects between 0.3-0.7% of the adult population with age an age of onset between 18 and 30 years old. Bipolar disorder (BPD) is characterized by patterns of manic and depressive episodes in which an individual may experience states of inflated self-esteem, decreased sleep, and mood changes followed by periods of deep depression[76]. BPD affects about 1.4% of the US population and usually presents in patients in their mid-20s. Major depressive disorder is the most common of the three disorders affecting 7% of the adult population[76]. It is marked by a depressed mood and loss of interest or pleasure for an extended period.

Studying these disorders is challenging due to their imprecise diagnoses and treatment regimens. The fifth edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM 5)[76] characterizes each disorder at length, but most patients do not present with a clear set of symptoms. Additionally, these disorders are part of a spectrum of psychiatric disorders, many of which have overlapping symptoms. For example, patients with Schizoaffective disorder experience delusions or hallucinations, like patients with schizophrenia, but will also suffer from manic or depressive episodes, similar to patients with bipolar disorder[76]. Therefore, depending on the patient's presentation of symptoms, an exact clinical diagnosis is difficult. Treating these disorders is equally challenging. Treatment regimens may include therapy, mood stabilizers, antipsychotic drugs and electroshock therapy[76-79]. Finding a successful treatment is often a process of trial and error and varies greatly among patients. Therefore, learning more about genetic risk factors for these disorders will give the field a better understanding as to their etiologies and potential therapeutic targets.

Studies suggest that these disorders are highly heritable. With an estimated heritability of 60% for SCZ and BPD and 40% for MDD[76,80-82]. In addition to being highly heritable, these disorders share common genetic risk factors. By analyzing over 30 years of medical records from the Swedish national registry, Lichtenstein et al. estimated that over 50% of the genetic risk for SCZ and BPD is shared[83]. These results are concordant with findings by the Psychiatric Genomics Consortium (PGC) which estimated a SNP based correlation of 0.68 between SCZ and BPD, 0.47 between BPD and MDD, and 0.43 between SCZ and MDD[84]. These results suggest that these three psychiatric disorders share common genetic risk factors.

During the mid 2000's, groups started to investigate the role of common genetic variation in the onset of human disease using genome wide association studies (GWAS). GWAS operate under the hypothesis that many common variants (greater that 1% frequency in the population) each contribute a small amount to disease risk[85]. These studies analyzed large cohorts of affected and healthy individuals to determine if common genetic variants are enriched in the affected population and therefore associated with the disease. While GWAS were successful in identifying hundreds of variants associated with many different diseases, for psychiatric disorders they were only able to identify a small number of associations. For example, of the nine SCZ GWAS published between 2007 and 2011, only four studies reported variants meeting a genome wide significant threshold resulting in only 11 significant loci[86]. Many regions were just under genome-wide significance thresholds, and comparing across studies revealed common risk genes between the three disorders. Variants near ANK3[87-91], CACNA1C[88-90,92] and the majority histocompatibility complex[93-99], were all been reported as associated with schizophrenia, bipolar disorder, and major depressive, often by multiple studies. Therefore, associations with these regions are replicatable across disorders and presumably by increasing sample size, more loci will reach genome wide significance. This was the case for the Schizophrenia Working Group of Psychiatric Genomics Consortium, who reported over 100 novel genomic regions associated with schizophrenia after analyzing over 100 thousand individuals. Though this study and subsequent large-scale analyses were successful in identifying hundreds of variants associated with psychiatric disorders, the majority of these variants lie in noncoding regions of the genome and therefore understanding how they may contribute to disease is not well understood.

Interpreting GWAS variants with epigenomic data

Over the last decade, genome wide association studies (GWAS) have identified thousands of genetic variants associated with human disease. However, the majority of these variants are in noncoding regions of the genome and determining how they contribute to disease remains a challenge. To aid in interpreting genetic variants, researchers have integrated epigenomic datasets such as DNase-seq, TF ChIP-seq and histone modification ChIP-seq data to annotate surrounding noncoding regions.

One common application is to determine disease relevant cell and tissue types. Using epigenomic datasets and subsequent predicted regulatory regions, many groups have developed methods for analyzing enriched cell types for GWAS variants[6,26,39,100,101]. Group have repeatedly reported SNPs associated with blood cell traits are enriched in regulatory regions active in K562 (an erythroid leukemia cell line)[39], autoimmune SNPs are enriched in regulatory regions active in T cells and B cells[26,100,102], and SNPs associated with cholesterol levels are enriched in liver regions[26]. While cell type enrichments from these analyses may seem obvious, there are important biological applications. These results can give new insights for disease pathology and suggest cell types to consider for therapeutic targets. For example, in 2012, the Stamatoyannopoulos lab reported that SNPs associated with multiple sclerosis were enriched in DHSs active in CD3+ T cells and CD19$^+$/CD20$^+$ B cells[6]. At the time, the role of T cells in multiple sclerosis was well established, but recent evidence also suggested B cells may also play important role and have therapeutic applications[103]. Since, clinical trials have demonstrated that Ocrelizumab, an anti-CD20 antibody, can lower disease activity and progression compared to traditional treatments[104]. Therefore, analyzing cell type enrichments can give further insight into biological mechanisms of disease and highlight potential therapeutic strategies.

In the case of schizophrenia, the schizophrenia working group of the PGC previously reported that variants were enriched in cell type specific H3K27ac peaks from brain tissues and B cells[105]. While the enrichment in brain was not surprising, the enrichment in B cells was intriguing as it remained significant even after removing SNPs in the major histocompatibility complex (MHC). As the authors only performed this analysis using data from 35 cell types, reanalyzing enrichment for these SNPs as well as other psychiatric associated variants using a larger panel of cell types is of great interest.

Epigenomic datasets can also be used to predict the biological function of disease associated variants. The Kellis and Snyder labs have summarized intersections with epigenomic datasets in their variant annotation databases HaploReg[106] and RegulomeDB[107]. For each variant, users can explore overlapping histone modification, DHS, and TF peaks and TF motif sites. RegulomeDB even supplies a score of how likely the variant is to disrupt TF based on overlapping peaks and TF motifs.

Groups have also annotated the functional consequences of genetic variants by correlating changes in gene expression and the epigenomic landscape with genotypes. Now with the relatively low cost of RNA-seq, genotyping arrays, and whole genome sequencing, large consortia now have the ability to detect expression quantitative trait loci (eQTLs) for thousands of genes in many different cell types. One of the largest efforts to date, The Genotype-Tissue Expression (GTEx) project, aims to of study the relationship between genetic variation and gene expression across different human tissues[108]. During the second phase of the project, the consortium identified over 1 million eQTLs in 44 tissues. eQTL methodology can also be applied to correlate changes in histone modification signal or chromatin accessibility with genotype. Groups have identified hundreds of histone modification QTLs[109,110] and DNase QTLs[111] using data lymphoblastoid cell lines generated by the HapMap[112] and 1000 genomes projects[113,114].

While QTL based methods require hundreds of genotyped individuals to observe significant trends, we can gather similar types of information at a cell type specific level by looking for allelic imbalance at heterozygous loci. By calculating the ratio of alleles from high-throughput sequencing reads, one can test whether a variant results in an allelic imbalance. Groups have analyzed TF ChIP-seqn[115,116], DNase-seq[102] and RNA-seq data[117] for sites of allelic imbalance and in doing so have uncovered mapping biases. It has been demonstrated that there are biases towards the reference allele[118-121]. To correct this bias, groups have suggested using an allele sensitive mapping tool, such as GSNAP[122], or map reads to genomes containing the reference and alternative alleles, or masked positions[121].

Ultimately, results from both functional characterization, eQTL, and AS analysis can be validated experimentally to test whether the observed correlation or imbalance is directly caused by variant in question. Many groups have tested the effect of alleles on enhancer activity using luciferase assays and Vockley *et al.* even pioneered a MPRA version[123]. However, as previously mentioned, these types of assays do not account for local genomic context. Therefore, as CRISPR-Cas9 technology improves to edit individual nucleotides[124,125], more labs will be effect of genomic variants using single nucleotide genomic editing.

CHAPTER II: Creating a Registry of Candidate Regulatory Elements for Human and Mouse Genomes

PREFACE

Results from this chapter were adapted from

Moore*, Purcaro*, Pratt*, Epstein*, Shoresh*, Adrian*, Kawli*, Davis*, Dobin*, Kaul*, Halow*, Van Nostrand*, Freese*, Gorkin*, He*, Mackiewicz*, The ENCODE Consortium. Cherry, Myers, Bing Ren, Graveley, Stamatoyannopoulos, Gerstein, Pennacchio, Gingeras, Snyder, Bernstein, Wold, Hardison, and Weng. "ENCODE Phase III: Building an Encyclopaedia of candidate Regulatory Elements for Human and Mouse,"

which is currently under review at *Nature.*

Len Pennacchio's lab tested candidate enhancers using transgenic mouse assays, the results of which are in Figure 2.7 Tables 2.4-6. I performed all analysis and generated all the figures in the chapter. Michael Purcaro and Henry Pratt designed, engineered and implemented the online visualization tool SCREEN (Search Candidate Regulatory Elements by ENCODE).

ABSTRACT

Here we described the Registry of candidate Regulatory Elements (cREs), which we defined using 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 (http://screen.encodeproject.org), the Registry is a resource for exploring noncoding DNA elements and their variants.

INTRODUCTION

Over 98% of the human genome is noncoding; less than 2% composes protein coding exons. A portion of these noncoding bases contain regulatory elements where transcription factors (TF) interact to control gene expression[1,7]. There are several classes of regulatory elements including promoter, enhancers, repressors, and insulators each with their own regulatory roles and epigenomic features. Identifying regulatory elements has important biological implications for studying gene regulation, cell type differentiation, and human disease.

Because of their cell-type specificity and long distance from TSSs, many groups have aimed to annotate enhancers. Enhancers are genomic elements that regulate and increase gene expression through interacting transcription factors. Using massively parallel report assays, STARR-seq, and mouse transgenic assays labs have generated collections of experimentally validated enhancers. For example, The VISTA enhancer browser is a collection mouse and human genomic regions that were tested for enhancer activity using transgenic mouse assays (Figure 2.2)[51]. These regions were selected using conservation[49], EP300 ChIP-seq data[29], and/or H3K27ac ChIP-seq data[50]. However, this is still a relative small collection of validated enhancers and it far from annotating every enhancer in the human and mouse genomes.

Over the last few years many labs have developed enhancer prediction algorithms that integrate data from the ENCODE and Roadmap Epigenomics projects. Histone modifications H3K4me1 and H3K27ac and TF EP300 are enriched at enhancers so by integrating these datatypes, along with other epigenomic data, lab have generated genome wide lists of predicted enhancers. While some of these methods have high performance, the majority do not have practical biological applications. Some methods such as RFECS[38] and REPTILE[36], are supervised methods requiring a positive set of enhancers for training. These methods also require a large amount of input data in some cases requiring over 20 histone modifications. Additionally, these prediction methods are designed to only identify potential enhancers; they do not predict other types of regulatory elements. Unsupervised methods of genome segmentation, such as ChromHMM, generate chromatin state maps genome wide[40]. Using a Hidden Markov Model, ChromHMM labels the genome with regions having promoter, enhancer, transcription, and repressive features. While powerful, these methods also require a lot of input data (up to eight histone modifications) and are only able to segment regions at a resolution of 200bp making comparisons across cell types difficult.

In this chapter, we introduce the Registry of candidate Regulatory Elements (cREs), collection of putative regulatory regions in human in mouse. We assigned cREs functional annotations (i.e. promoter-like, enhancer-like) based on computational methods that we experimentally validated. We compared the Registry with other methods element prediction finding that our method is concordant with other epigenomic data based methods. Finally, we demonstrated that these cREs have biological implications as cell types cluster based on cRE activity and variants reported by genome wide association studies are enriched in cREs active in disease relevant cell types.

RESULTS

Cell Type Specific Enhancer Prediction

We began by developing a method for tissue-specific enhancer prediction that met the following criteria: One, our method needed to be unsupervised. While we had a collection of experimentally validated enhancers from reporter assays, mouse transgenic assays, and STARR-seq experiments, these data were only from a small number of cell types. Therefore, while we developed and tested models using these validated regions, our method did not require them. Two, we needed to use our method to predict enhancers across hundreds of cell types. While the ENCODE consortium generated hundreds of genomic assays in cell lines such as K562 and GM12878, the majority of cell types only have a handful of experiments such as DNase-seq and histone modification ChIP-seq. Though we may be able to develop a more accurate enhancer prediction model using multiple data types (e.g., RAMPAGE and TF ChIP-seq), this model could not be applied to the majority of cell types.

Mindful of these requirements, we decided to develop our enhancer prediction model using data and experimentally validated regions from embryonic mice. During the third phase of the ENCODE project, production labs generated a dense matrix of genomic data surveyed across twelve tissues in up to eight embryonic stages. The Ren lab assayed nine histone modifications across 72 embryonic tissue-time points (Figure 2.1) and the Wold and Ecker labs generated matching RNA-seq and whole genome bisulfite sequencing data (WGBS). Additionally, the Stamatoyannopoulos lab produced DNase-seq data for 18 of these tissue-time points including limb, midbrain, hindbrain, and neural tube at embryonic day 11.5 (e11.5). In addition to this expansive collection of genomic data, we established collaborations with the Pennacchio and Visel labs who maintain the VISTA enhancer browser. As of November 2015, there were 1,994 tested regions in the VISTA database; 228 were active in limb, 301 were active in midbrain, 271 were active in hindbrain, and 193 were active in neural tube. Therefore, we decided to develop our enhancer prediction method using genomic data and experimentally validated enhancers in mouse limb, midbrain, hindbrain, and neural tube tissues at e11.5.

To develop our unsupervised enhancer prediction method, we went through two rounds of testing. First, we tested which data type was best for anchoring predictions (i.e., which peaks we should center our predictions on). Second, we tested methods for ranking these peaks (i.e., what type of signal or combinations of signals are most predictive). To begin, we tested anchoring predictions on

H3K27ac, H3K4me3, H3K4me1, and DNase peaks (further referred to as DNase hypersensitivity sites – DHSs) (Table 2.2). To fairly compare performance across the different data types, we selected the top 20,000 peaks from each dataset and then set each peak to a uniform width of 300 bp. On average, we achieved the highest performance anchoring on DHSs with an average AUPR of 0.36, followed by H3K27ac peaks with an average AUPR of 0.33 (Table 2.2). When we analyzed individual tissues, DHSs had the best performance for limb, hindbrain, and neural tube whereas H3K27ac peaks had the best performance for midbrain (Figure 2.4, Table 2.2). Because DHSs had the highest average AUPR and were the top predictor for three of the four tissues, we decided to anchor our enhancer predictions on DHSs.

Our next step was to evaluate different methods of ranking DHSs. We tested ranking by different histone modification signals (H3K27ac, H3K4me3, H3K4me1, H3K9ac, H3K36me3, and H3K27me3) and DNA methylation signal. We also tested combining DNase with histone modification and methylation signals. We were unable to simple average these signals, due to differences in data processing. Therefore, to combine signals for region, we averaged its DNase signal rank and histone modification/methylation rank generating a metric we referred to as "average rank." In general, the best performing method was ranking DHSs by the average rank of DNase and H3K27ac signals (Figure 2.5, Table 2.3).

To validate our method, the Pennacchio and Visel labs tested our enhancer predictions using transgenic mouse assays. For limb, midbrain, and hindbrain, we curated a ranked list of predictions and selected three tiers of regions to test: ranks 1-20, ranks 1,500-1,515 and ranks 3,000-3,015 (Tables 2.4-2.6). In general, higher ranking regions were more likely to show enhancer activity than lower ranking regions in their predicted tissue (Figure 2.6). For example, in limb, 70% of the top ranked regions had enhancer activity in limb compared to 40% and 20% in the middle and bottom tiers respectively. In addition to the predicted tissues, we also observed activity in biologically similar tissues for some of the tested regions. For example, mm154, a high-ranking midbrain prediction, was active in midbrain, forebrain, hindbrain, neural tube and eye (Figure 2.7a, Table 2.4) and mm1489, a high-ranking hindbrain prediction, was active in hindbrain, midbrain and neural tube (Figure 2.7b, Table 2.5). High H3K27ac signal in these tissues supported this additional enhancer activity (right panels Figure 2.7). In contrast, mm1485, a high-ranking limb prediction, is only active in limb tissue (Figure 2.7c, Table 2.6) and we did not observe high H3K27ac signal in tissues other than limb. These results suggested that our unsupervised method of enhancer prediction, which combines DNase and H3K27ac data, was capable of successfully identifying tissue-specific enhancers.

Developing a Registry of Candidate Regulatory Elements

Since we developed a computational method that successfully identified active enhancers, our next step was to use this model to curate a collection of putative enhancers across human and mouse cell-types. Even though our method only required DNase and H3K27ac data, only a small percentage of ENCODE and Roadmap cell types had both of these assays. In humans, of the 540 cell types surveyed by DNase and H3K27ac, only 58 had both assays while in mouse only 23 of 131 cell types had both DNase and H3K27ac (Figure 2.8). Additionally, our prediction method was limited in that it only identified potential enhancers; it did not define other types of regulatory elements such as promoters, insulators, or repressors. Using DNase, H3K4me3, and gene expression data, we demonstrated that by using a method analogous to our enhancer prediction method, we could predict candidate promoters (Figure 2.9, Tables 2.7-2.8). Because of these concerns, we decided to adapt our method to make it applicable to more cell and tissues types as well as flexible enough to identify other types of regulatory elements. With our new method, we aimed to create a collection of putative regulatory regions across human and mouse that we called the registry of candidate regulatory elements (cREs).

For both our enhancer and promoter prediction methods, anchoring predictions on DHSs consistently resulted in the best overall performance. Because the boundaries of DHSs are generally consistent across cell types (Figure 2.10a), we decided to anchor cREs on a consensus set of DNase accessible regions, which we call representative DHSs (rDHSs). To create rDHSs in humans, we curated over 48 million high-quality DHSs (FDR < 0.1%) from 449 DNase experiments and grouped them into overlapping clusters (Figure 2.10b). For each cluster, we selected the DHS with the highest signal as the rDHS and discarded overlapping DHSs (Figure 2.10c). Using the selected rDHSs and the remaining non-overlapping DHSs we iteratively repeated the clustering and selection steps until all DHSs overlap at least one rDHS. For humans, we defined over 2 million rDHSs, and for mouse, with 8.6 million high-quality DHSs from 62 datasets, we defined about 1 million rDHSs.

To classify an rDHS as a cRE, we integrated DNase data with H3K4me3, H3K27ac, and CTCF data, adopting similar approaches to our enhancer and promoter prediction schemes. However, comparing these datasets and choosing consistent cutoffs is complicated due to differences in data processing pipelines, sequencing depth, and assay protocols. For example, DNase-seq, CTCF-ChIP-seq, and histone modification ChIP-seq experiments are all processed using different pipelines at the ENCODE data coordination center (DCC). Signal from ChIP-seq experiments is normalized using input data while signal from DNase is dependent on sequencing depth. Additionally, even for data that has been processed using the same pipeline, such as H3K27ac experiments, signal files are not comparable across cells types (Figure 2.11a). Therefore, to normalize signal across assays and cell types, we took the log of the average signal across an rDHS and converted it to a Z-score (Figure 2.11). This resulted in an approximately normal distribution of signals for each cell type. For each assay and cell type, we assigned an rDHS a signal Z-score. We refer to a max Z-score as the maximum Z-score for the rDHS across all cell types for the particular assay (DNase, H3K4me3, H3K27ac, or CTCF). While this method may bias against cell types, such as embryonic stem cells that have higher numbers of active cREs, we decided error on the side of being more conservative rather than over call elements.

Using this signal normalization scheme, we required rDHSs to meet two criteria to be classified as a cRE. One, the rDHS must have a max DNase Z-score greater than 1.64 (95 percentile using one-tailed test). This filters out low signal rDHSs. Two, the rDHS must also have a max Z-score > 1.64 for either H3K4me3, H3K27ac, or CTCF. This step filters out rDHSs that may be due to spurious, nonfunctional transcription factor binding. These two requirements (DNase max Z > 1.64 and ChIP-seq max Z > 1.64) do not need to occur in the same cell type since only 3% of cell types have all four assays. However, cREs that have high DNase and one of the three ChIP-seq signals in the same cell-type are given a special designation as concordant cREs (approximately 55% of human cREs and 52% of mouse cREs). Using the classification trees in Figures 2.12 and 2.13, we classified cREs into three groups (cREs with promoter-like signatures (PLS), cREs with enhancer-like signatures (ELS), and CTCF-only cREs) based on max Z of ChIP-seq signals and distance from TSS. We refer to these groups as cell-type agnostic classifications. In total, we curated 1.3 million cREs in human, and 432 thousand cREs in mouse which comprise 20.8% and 8.8% of their respective genomes (Figure 2.14). We assigned each of these cREs a unique accession, with prefixes EH37E for human and EM10E for mouse.

After defining cREs, we next sought to determine the cell-type specific activity of each cRE across hundreds of human and mouse cell types. For cell types with DNase and all three ChIP-seq assays, there were 16 possible combinations of these four signals. Since we only considered a cRE as active in a cell type if it has a DNase Z-score > 1.64, we condense these 16 possible states into eight active states and one inactive state (2.15a). To simplify this classification scheme, we aimed to further condense these states into five groups: PLS-cREs, ELS-cREs, CTCF-only cREs, DNase-only cREs, and inactive cREs. To aid in classification, we used GM12878 as a test case (Figures 15,16). We separated cREs in each of the nine states based on TSS proximity, resulting in 18 sub-states (Figure 15a). For cRES in each of these sub-states, we calculated the average POLII (Figure 2.15b), EP300 (Figure 2.15c) and RAD21 ChIP-seq signal using experiments in GM12878. We predicted POLII, which is present at active transcription start sites, should be highest at PLS elements. EP300, a TF known to bind at active ELS elements[28], should be the highest at enhancer-like elements, and RAD21, part of the cohesion complex that is known to localize with CTCF, should be the highest at CTCF-only elements. Using the median signal values for each TF, we compared the 18 sub-states and observed that like sub-states formed clusters (Figure 16a,b). For example, distal cREs with H3K27ac z-scores > 1.64 have EP300 high signal and low POLII and RAD21 signals (Figure 16a). Based on these plots, we assigned each of the 18 sub-states to one of the five groups (Figure 16c) which in the case of GM12878 results in 36,022 PLS, 27,739 ELS, 10,913

CTCF-only and 16,085 DNase-only cREs (Figure 16d). As groups PLS, ELS, and CTCF-only cRES are enriched in POLII, EP300, and RAD21 signals respectively (Figure 16e). Using this method, we classified cREs into cell-type specific groups for the 21 cell types with all four assays. For PLS, ELS, and CTCF-only groups, we plotted the saturation curve across the 21 cells types (Figure 2.17). There are more ELS cREs than PLS cREs, which we would expect since enhancers are more likely to be cell-type specific than promoters[15]. For the majority of cell types, we do not have all four genomic assays, but we are still able to combine sub-states into simplified groups. We listed all possible state-group classification schemes in Figure 2.18. In total, we assigned each human cRE to a group in 620 cell types and each mouse cRE to a group in 138 cell types.

After generating cREs for human and mouse and determining their cell-type specificity, we wanted to compare cREs between species. We mapped mouse cREs to the human genome using UCSC's liftover tool. 20% of human cREs overlapped a mouse cRE (52% of total mouse cREs), which we refer to as orthologous cREs. We noticed that larger percentage of orthologous cREs were PLS cREs compared to either human or mouse cREs as a whole.

Comprehensiveness of the Registry of cREs

First, we examined how many GENCODE-annotated TSSs (V19 for human and M4 for mouse) were covered by the current version of the Registry of cREs. For human, 67% of all annotated TSSs and 72% of protein-coding TSSs overlap a cRE. When we searched +/- 2 kb around each TSS, to account for possible misannotation, 92% of all TSSs were proximal to at least one cRE. Coverage is similarly high in mouse, despite having far fewer cREs, with 61% of all annotated TSSs overlapping a cRE and 80% having at least one proximal cRE.

Second, we analyzed how rapidly the total number of unique rDHSs saturated when we increase the number of covered cell types. In ENCODE Phase II, Steven Wilder and Ian Dunham modelled DHS saturation using a Weibull distribution and estimated that they had discovered around half of the total DHSs[25]. We replicated this analysis using the 460 DNase datasets that we used to create the rDHSs. The saturation curves of rDHSs follow Weibull distributions, reaching a plateau at 1.66 M rDHSs with FDR < 0.1% and Z-score > 1.64 (Figure 2.17). 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 estimate that we have identified at least 78.9% of total 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 and H3K4me3 peaks (FDR<0.01) in cell types with ChIP-seq data but no DNase data. The Registry covered 90 ± 8% of H3K4me3 peaks (74 cell types), and 87 ± 5% of H3K27ac peaks (54 cell types) (Figure 2.20) The coverage was equally high for mouse, despite a smaller number of DNase 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 (Figure 2.21) 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 (Figure 2.22) When we visually inspected the two datasets with the lowest coverage (CD-1 megakaryocyte and GR1-ER4 in mouse), we confirmed that the peaks that were not covered by the Registry had low signals and were likely false positives by the peak calling algorithm MACS2.

Therefore, when judged against gene annotations and epigenomic datasets, the human Registry appears to be comprehensive. It covers almost 80% of all cREs and 85% of elements marked by H3K4me3 or H3K27ac. 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.

Comparison with Previous Defined Regulatory Regions

To further validate our approach, we compared the cREs to previously defined regulatory elements. Like cREs, ChromHMM regions are defined using epigenomic datasets generated by the Roadmap and ENCODE projects. Using states from a ChromHMM model, which was implemented using eight histone modifications and CTCF[40], we found that the overlap between the model and our cREs was highly concordant. Of the top 10,000 ranked PLS cREs (ranked by H3K4me3 Z-scores), 90% overlapped ChromHMM TSS states while 85% of the top 10,000 ranked ELS cREs (ranked by H3K27ac Z-scores) overlapped

ChromHMM high-signal enhancers states. The overlap decreased for lower ranking ELS cREs, but the overlap with ChromHMM low-signal enhancers increased; 82% of the ELS cREs ranked above 20,000 overlap with ChromHMM enhancers or low-signal enhancers (Figure 2.26a,b). We also compared the cREs for five e11.5 and six e14.5 mouse tissues (tissues with DNase data) with the ChromHMM states called using eight histone modifications in the corresponding tissues. We observed that 95 ± 2% of PLS cREs overlapped ChromHMM TSS states and 78 ± 3% of ELS cREs overlapped ChromHMM enhancers states in the corresponding tissue and time point (Figure 2.26c,d). This suggests that while our method was able to identify similar putative regulatory regions as the ChromHMM model using less input data.

We also compared our ELS cREs with enhancers annotated by transcription data and STARR-seq peaks. We intersected our ELS cREs with FANTOM defined enhancers in GM12878, astrocyte, hepatocyte and keratinocyte cells[43]. While the overall percentage of overlapping ELS cREs was low (2%), we observed the largest percentage of overlap for highly ranked enhancers (Figure 2.27a). When we overlapped the FANTOM enhancers with all cREs we found that 74% overlapped, with 70% overlapping cREs active in the cell type (Figure 2.28a, Table 2.10). Of the active cREs the majority were ELS (66%) followed by PLS (28%). We observed similar results transcriptional regulatory elements (TREs) defined using GRO-seq data[44], except that these elements overlapped a higher percentage of PLS cREs (57% of active cREs) because they encompass all types of regulator elements (Figure 2.27b, and 2.28b, Table 2.10). STARR-seq peaks in HeLa cells overlapped even fewer cREs (31%) (Figure 2.27c and 2.28c, Table 2.10) but even by their own internal annotations, they only annotated 20% of their peaks with chromHMM enhancer states.

These results suggest that while the Registry had similar performance to other epigenomic data based methods, such as chromHMM, it did not identify all potentially functional regions in the human and mouse genomes.

Cell and tissue type clustering using cRE activity

To examine whether the Registry of cREs captured biologically relevant regulatory patterns, we clustered primary cells and tissues based on the number of overlapping active cREs defined using DNase, H3K27ac, or H3K4me3 signal. We first compared the clustering schemes in mouse using the 72 embryonic tissue-time points with H3K27ac and H3K4me3 data. Using H3K27ac, we observed almost perfect clustering of tissues by their organs of origin (Figure 2.23). When we clustered by H3K4me3 activity, the tissues do not segregate as cleanly . We believe this is because H3K4me3 signal is enriched at promoters, which are more consistent across cell types (Figure 2.17). This is reflected in average Jaccard coefficients; the average H3K27ac coefficient for mouse is 0.36 while for H3K4me3 it is 0.79. Higher similarity between cell types will make them more difficult to cluster.

In humans we decided to cluster primary cells and tissues by cRE DNase and H3K27ac activities since these assays cover different tissues (Figure 2.8). Using H3K27ac signal, we observed tissues from different regions of the same organ cluster together (Figure 2.24a). For example, brain regions form a distinct group. In some cases, we also observed fetal and adult tissues clustering such fetal and adult adrenal gland tissues. Interestingly, samples from the gastrointestinal tract form two clusters, one for smooth muscle tissues (the purple and maroon samples at the top) and the other for mucosa tissue (the maroon samples at the center). This suggests that while these tissues are located in proximity to one another in the human body, they have different regulatory landscapes. When we analyzed primary cells, we observed three perfectly segregated groups colored by their embryonic origins: blood, non-blood mesoderm, and ectoderm (Figure 2.24b). Even the endothelial cells of the umbilical vein, which are derived from the extraembryonic mesoderm, clustered with the cell types derived from the embryonic mesoderm such as fibroblasts, myoblasts, osteoblasts, and astrocytes.

When we clustered using DNase signal, we observed similar results. For DNase, we have multiple donors for the same types of tissues and in the majority of cases these donor samples clustered together (e.g., kidney, stomach, lung, and muscle tissues) (Figure 2.25a). We did observe some noticeable outliers such as one fetal thoracic segment muscle sample clustering with lung fetal lung tissue. This could be due to a possible sample swamp (i.e. samples were submitted to the

ENCODE DCC with incorrect labels) and therefore should be further analyzed. When we clustered primary cells, we observed two large clusters, with the one cluster composed entirely of blood cells, subdivided into to the myeloid and lymphoid lineages (Figure 2.25b). The second cluster, was comprised of several smaller "subclusters". The bottom subcluster contained of four trophoblast samples (in black), thus reflecting their extraembryonic fate. The topmost subcluster contained mostly fibroblasts, and the middle subcluster contained endothelial cells, epithelial cells, keratinocytes, and melanocytes. The fibroblasts aggregated together regardless of their anatomical locations, as did most of the endothelial cells, in agreement with their common mesodermal origin. Most of the epithelial cells also clustered together, despite their different embryonic germ layers. Overall, these results demonstrated that like tissues and cell types clustered together when we compared their cRE activity, suggesting that the registry of cREs is able to capture biologically relevant regulatory patterns.

Applications to Genome Wide Association Studies (GWAS)

Previous studies have repeatedly demonstrated that most GWAS variants reside in noncoding regions of the genome. Annotation of these noncoding regions can be used to guide the interpretation of GWAS variants by predicting disease-relevant cell types and regulatory factors[6,26,39,100,101]. 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.

We curated variants from over 50 studies in the NHGRI-EBI GWAS. For each phenotype-cell type comparison, we tested whether active cREs (H3K27ac or DNase z-score > 1.64) were significantly enriched in the GWAS SNPs. Overall, we observed enrichment in disease related cell types. Like previous studies[6,26,39,100,101], we observed enrichments in immune cells such T and B cells for autoimmune disorders multiple sclerosis, type 1 diabetes, inflammatory bowel disease and Crohn's disease (Figure 26). Additionally, in blood cells we observed an enrichment for platelet count, platelet volume and red blood cell traits. We also observed enrichment for variants linked with cholesterols, metabolite, and fibrinogen levels in liver cREs. Finally, thyroid hormone level variants were enriched in thyroid tissue cREs, schizophrenia variants were enriched in brain cREs, and breast cancer variants were enriched in cREs active in MCF-7, a breast cancer cell line.

SCREEN: A Web Based Visualization Tool for the Registry of cRES

To search and visualize the Registry of cREs, we built a web-based tool called SCREEN (Search Candidate Regulatory Elements by ENCODE). SCREEN hosts the 1.3 million human and 430 thousand mouse cREs and connects them with underlying ENCODE data and annotations. The first version of SCREEN is divided into three “apps”, each of which provide a different perspective on the cREs. The core app is a cRE-centric search, where users can retrieve a subset of cREs using genomic coordinates, a gene name, or SNP accession (Figure 2.27a). SCREEN returns a list cREs, annotated with their location, nearest genes, and max Z-scores for H3K4me3, H3K27ac, and CTCF signals (Figure 2.27b). Users can filter this list by selecting a cell type interest; SCREEN will then filter out cREs that are not active in that cell type. Users can also filter results using Z-score cutoffs by choosing stricter or more permissive thresholds compared to the default (Z-scores > 1.64). If a user selects a cRE, SCREEN brings him to a cREs details page. Here the user can browse the cRE's H3K4me3, H3K27ac, CTCF, and DNase Z-scores in every cell type (Figure 2.30c), search for overlapping genomic datasets and genetic features such as topologically associated domains and SNPs.

The gene-centric app, which opens when a user searches for a gene name, plots RNA-seq and RAMPAGE TSS expression data. Within this app, we developed a differential gene expression tool to analyze the relationship between changes in cRE activity and gene expression across mouse embryonic development. Users can selected two tissue-time point combinations and a region of interest. SCREEN will display differentially expressed genes in this region and ELS and PLS cREs in this region. For example, we observed that *Ogn,* a protein involved in bone formation, dramatically increases in expression between e11.5 and e15.5 in limb tissue. This increase in gene expression corresponds to bone development which occurs around e12.0[41]. SCREEN's differential gene expression tool displays the expression fold change of *Ogn* and nearby differentially expressed genes as bars (Figure 2.31a). PLS and ELS cREs are shown as red and yellow dots with the y-axis indicated difference in H3K4me3 and H3K27ac Z-scores between the time points. This large-scale view helps users identify cREs that might account for the increase in *Ogn* expression by looking for corresponding changes in cRE activity and gene expression. Using this approach, we identified an ELS cRE (EM10E0113220) that increases in activity between e11.5 and e15.5. When we analyzed EM10E0113220's H3K27ac activity across limb development (Figure 2.31b,c), we found that it is highly correlated with PLS cRE H3K4me3 activity and Ogn expression. Therefore we hypothesize that increase in EM10E0113220 activity leads to increased Ogn expression during limb development.

Finally, the SNP-centric GWAS app intersects cREs with SNPs we curated from the aforementioned GWAS studies. Users can selected a study of interest and SCREEN will return a list of cell types and cREs that are enriched for GWAS SNPs. All three apps show links to the UCSC genome browser, thus facilitating visualization of the epigenetic signals at a cRE’s or a gene’s locus, such as the Ogn example describe above. We have also set up a trackhub for visualizing all available signal tracks at the UCSC browser, organized by cell type.

DISCUSSION

By integrating DNase-seq and ChIP-seq datasets, we generated a collection of putative regulatory regions in human and mouse, which we referred to as the Registry of cREs. Adapting our unsupervised approach for enhancer prediction, which we used to successfully predict enhancers active in embryonic mouse tissues, we developed a cRE identification and classification scheme. We classified cREs into groups (PLS, ELS, CTCF-only, DNase-only, inactive) across 600 human and 100 mouse cell types, generating the most comprehensive collection of cell-type specific regulatory elements. We demonstrated that the registry covers the majority of H3K4me3 and H3K27ac peaks in cell types without DNase data and that its classifications were concordant with ChromHMM genome segmentations.

We also determined that our registry is biologically consistent with our current biological understanding of cell type relationships. With our clustering analysis, we determined that these cREs have biologically relevant activity patterns. While human tissues and primary cells generally clustered by their organ and embryonic tissues of origin, the clustering was not nearly as clean as it was in mouse. This could be due to a number of reasons. The human samples were collected from a number of individuals who had different genetic backgrounds and had experienced different environmental effects. These samples also may have been collected using different collection methods depending the on if the samples were biopsies or postmortem tissues. Finally, in some cases the ChIP-seq experiments were run by different production labs during different phases of the ENCODE and Roadmap projects. The means that this experiments were run with different quality control standards, on different sequencing machines, and analyzing using different analysis pipeline. For, mouse, the embryonic tissues were collected from mice with identical genetic backgrounds using identical tissue collection procedures. The assays were all conducted during the third phase of the ENCODE project by the same production labs. While experimental biases may still exist, the nature of these datasets controls for many common sources of bias. This also presents one of the advantageous of using mouse data such as the application of mouse cREs to (see Chapter IV).

While we have identified over 1 million cREs in human and 400 thousand in mouse, we are aware that our registry is far from complete. Currently, in order for an rDHS to be classified as a cRE, it must have high DNase signal and high H3K4me3, H3K27ac or CTCF signal. Therefore, we are filtering our elements such as poised/weak enhancers, which have high H3K4me1 and H3K27me3 signals and low H3K27ac signal[39]. Additionally, our current classification scheme only identifies three types of elements. Realistically, the human genome is composed of many types of cREs with a spectrum of activity patterns. For example, recent studies have reported promoters having enhancer-like activity and regulating the activity of distal genes[56,126]. By analyzing additional histone modifications, we may be able to identify PLS cREs with these long-range activities. Additionally, CTCF has many roles in the genome: It is associated with the cohesion complex, insulators, and repressors[30]. Therefore, we maybe be able to further separate CTCF-only cREs into different sub-classifications. During phase IV of the ENCODE project, we hope to not only increase the number of cell types for these assays, but continue to analyze different combinations of signal patterns, binding of transcription factors and transcription data play a role, particularly in cell types such as K562 and GM12878 that have a large number of datasets. We also hope to integrate in publicly available datasets, from databases such as CISTROME[127] to cover a wider range of cell types and further annotate cREs.

Finally, while the registry of cREs had high coverage across the epigenomic landscape, its overlap with regulatory elements defined by other methods such as GRO-seq, CAGE and STARR-seq was far less. One possibility for the poor overlap is that our method is filtering out low signal enhancer that test positive in these assays or have strong transcriptional activity. For transcription based methods, such as CAGE and GRO-seq, this poor overlap could also be due to random transcription events in the genome. These regions, despite not having hallmark histone modifications, may have sequences with transcription initiation potential. These regions may have no biological function, so transcription is a result of the noise of genomic regulation, or they could be a different class of regulatory elements. In the future, it will be worth investigating the transcriptionally active loci to see what features are predictive of these regions.

For methods that test specific genomic regions for enhancer activity, such as STARR-seq, MPRA, and transgenic mouse assays, this poor overlap could be due to differences in genomic context. For example, the Stark lab reported in their original STARR-seq paper that 31% of STARR-seq peaks did not overlap DHSs in the same cell type[128]. They hypothesized that this is due genomic context of the region. For example, a sequence may innately have enhancer-like activity causing it to test positive in STARR-seq, luciferase, or transgenic mouse assays. However, in the genome, the region may be silenced and inaccessible to transcription factors and therefore will not overlap a DHS. Moving forward, we can isolate these cases to learn more about how sequence features contribute to enhancer activity. We can apply the same analysis to analyze human regulatory elements and use these features to further refine our enhancer prediction models.

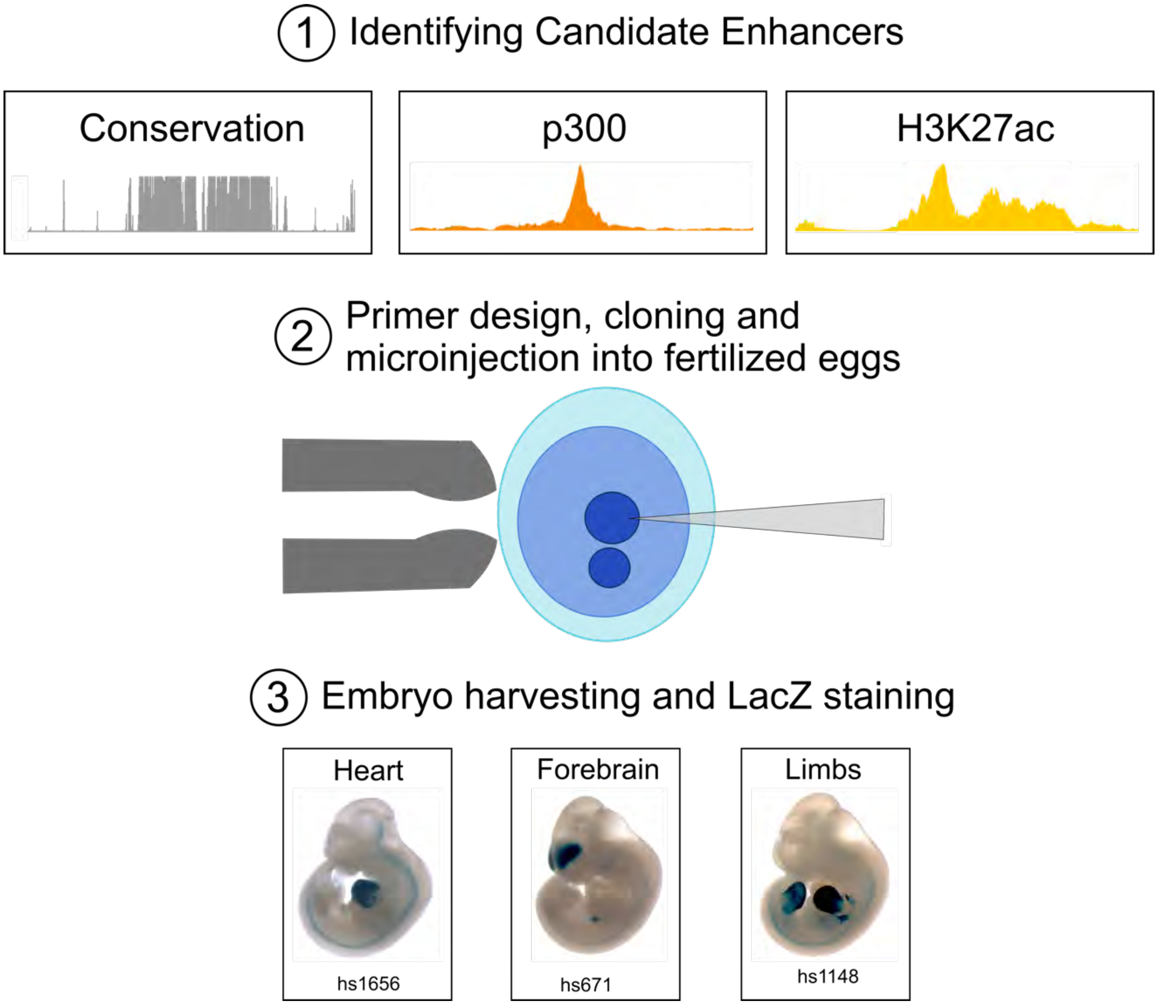
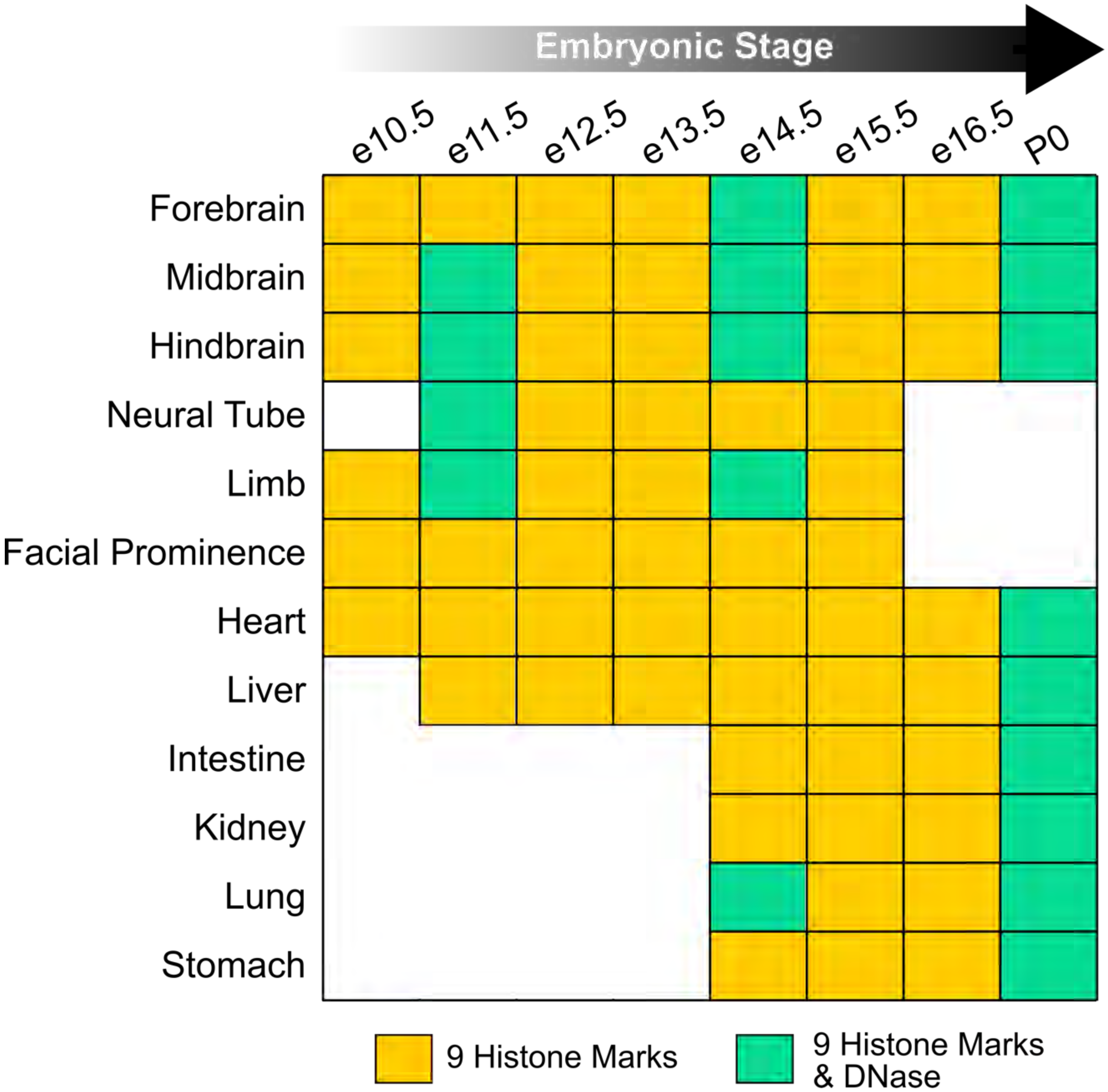

1 Identifying Candidate Enhancers
Conservation
p300
H3K27ac
2 Primer design, cloning and microinjection into fertilized eggs
3 Embryo harvesting and LacZ staining
Heart
hs1656
Forebrain
hs671
Limbs
hs1148

Figure 2.1 | Protocol used to validate enhancers included in the VISTA database. Regions were selected based on conservation, EP300 ChIP-seq, or H3K27ac ChIP-seq signals. These regions were then cloned and then injected into mouse embryos. Embryos were harvested and enhancer activity was measured using lacZ staining (blue regions). This figure was adapted from the home page figure on https://enhancer.lbl.gov/.



	e10.5	e11.5	e12.5	e13.5	e14.5	e15.5	e16.5	P0
Forebrain	🟨	🟨	🟨	🟨	🟩	🟨	🟨	🟩
Midbrain	🟨	🟩	🟨	🟨	🟩	🟨	🟨	🟩
Hindbrain	🟨	🟩	🟨	🟨	🟩	🟨	🟨	🟩
Neural Tube		🟩	🟨	🟨	🟨	🟨		
Limb	🟨	🟩	🟨	🟨	🟩	🟨		
Facial Prominence	🟨	🟨	🟨	🟨	🟨	🟨		
Heart	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟩
Liver		🟨	🟨	🟨	🟨	🟨	🟨	🟩
Intestine					🟨	🟨	🟨	🟩
Kidney					🟨	🟨	🟨	🟩
Lung					🟩	🟨	🟨	🟩
Stomach					🟨	🟨	🟨	🟩

🟨 9 Histone Marks 🟩 9 Histone Marks & DNase

Figure 2.2 | ENCODE3 mouse embryonic time series data. During phase III of the ENCODE project the Ren lab generated histone modification ChIP-seq data for twelve tissues across eight embryonic time points (orange and green boxes, 72 unique tissue-time point combinations). The Stamatoyannopoulos lab generated DNase-seq data for 18 of these tissue-time points (green boxes).

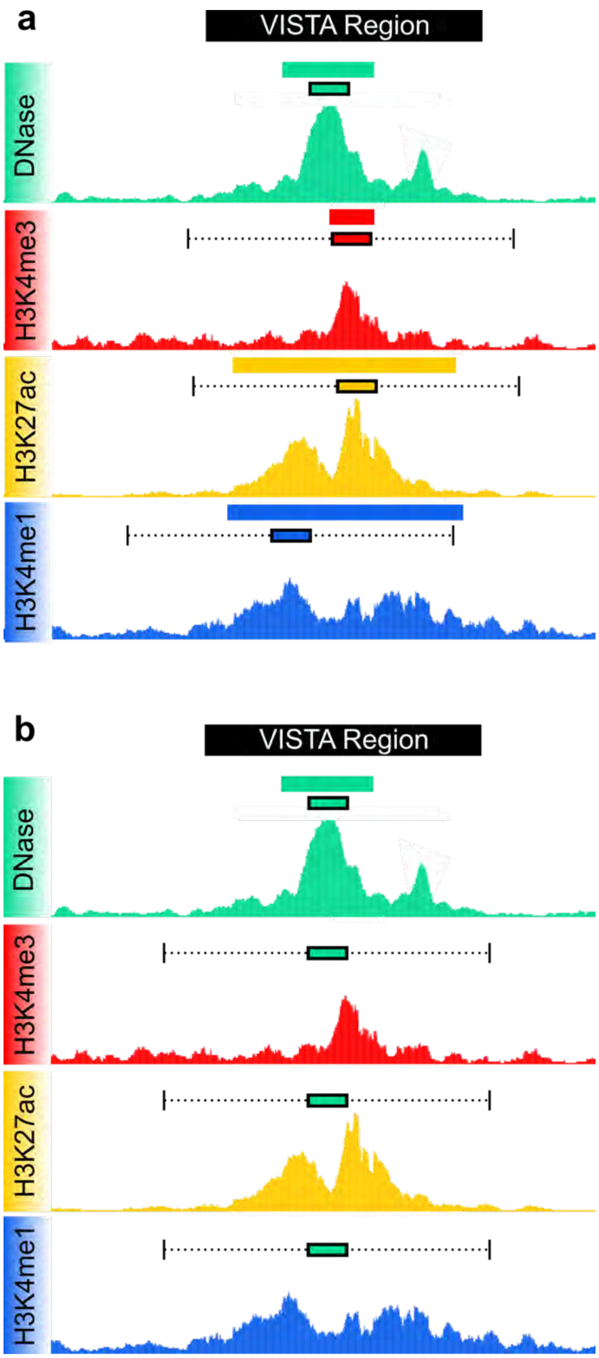


Figure 2.3 | Unsupervised enhancer prediction methods. a, Testing which peaks to anchor predictions. To control for genome coverage, all peaks are set to a uniform 300bp in width. For DHSs, we use signal across the 300bp region. For histone peaks we use +/- 1kb from the summit of each peak. **b,** After we determined anchoring on DHSs result in the best performance, we tested ranking schemes using signal. For DHSs, we use signal across the 300bp region. For histone peaks we use +/- 1kb from the summit of each DHS.

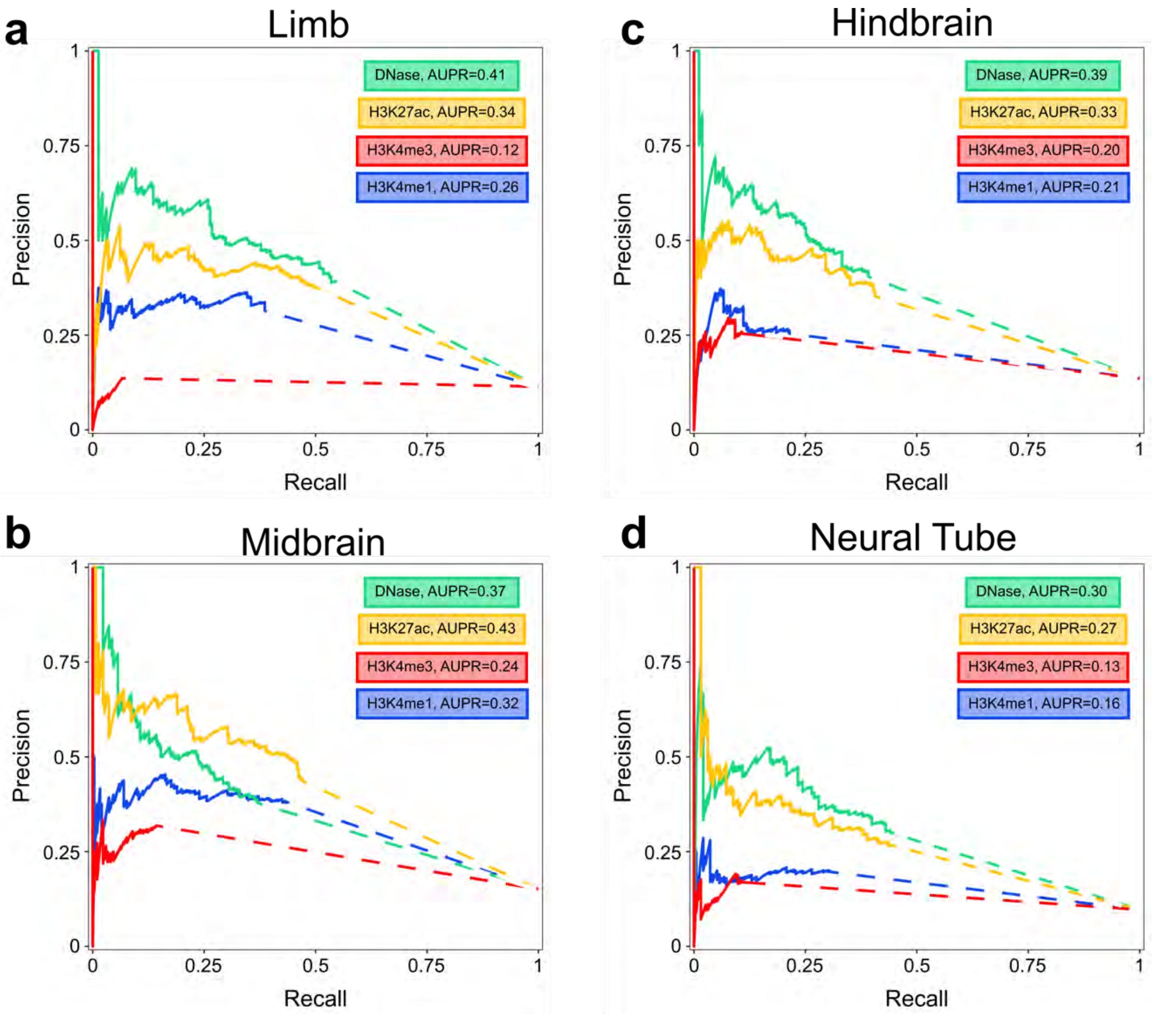


Figure 2.4 | Precision-Recall (PR) curves for VISTA Enhancer prediction. PR curves for **a,** limb, **b,** midbrain, **c,** hindbrain, **c,** midbrain, and **d,** neural tube enhancers at e11.5. Colors indicate peaks and signals used for anchoring and ranking the enhancer predictions. All peaks were set to 300 bp centered on their summits and the 20k top-ranked peaks were used for each tissue to ensure consistent genome coverage.

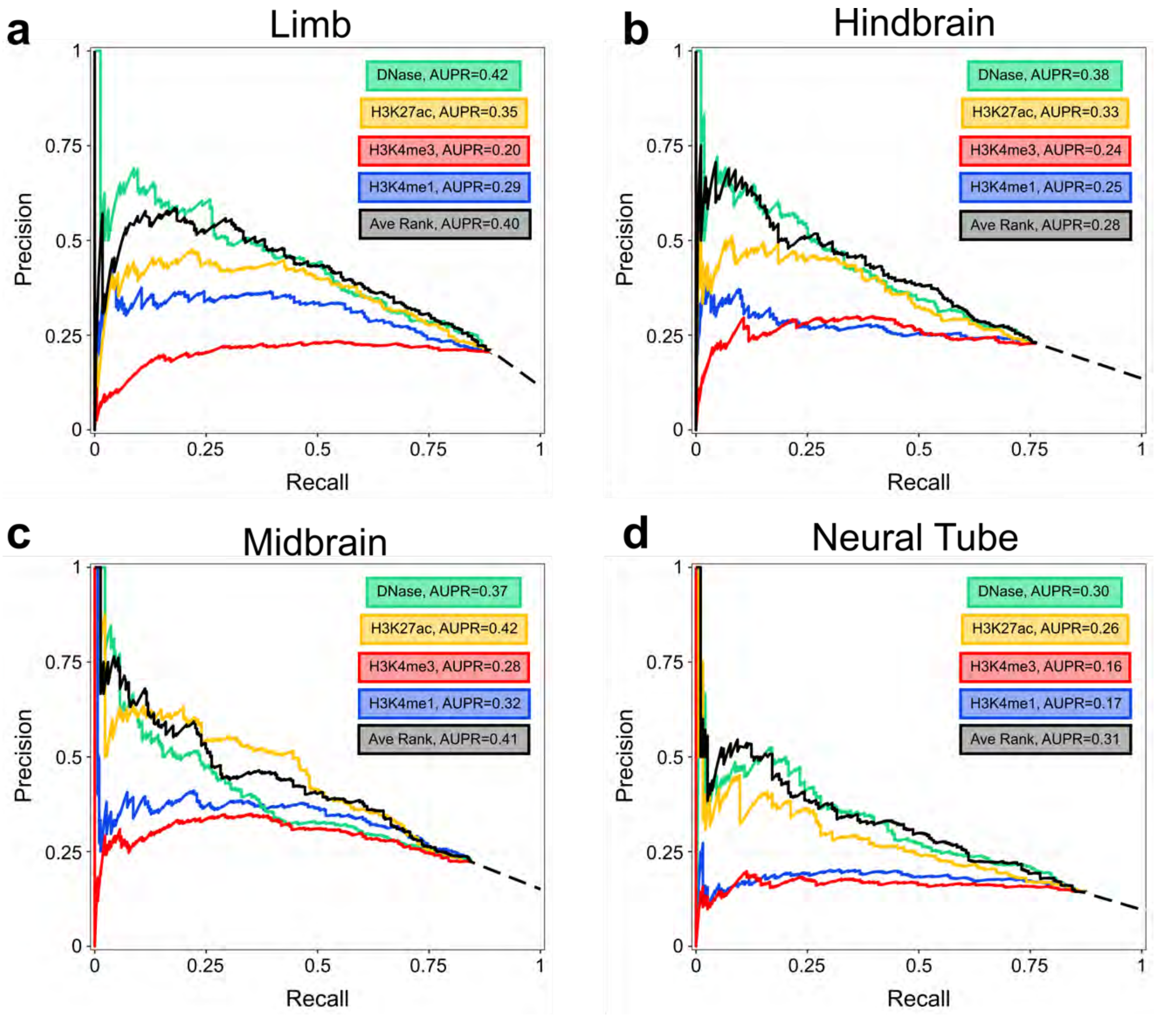


Figure 2.5 | PR curves for VISTA Enhancer prediction anchored on DHSs. PR curves for **a,** limb, **b,** hindbrain, **c,** midbrain, and **d,** neural tube enhancers at e11.5. All predictions were anchored on DHSs in the respective tissue. Colors indicate signals used for ranking predictions; black indicates the average of DNase and H3K27ac signals.

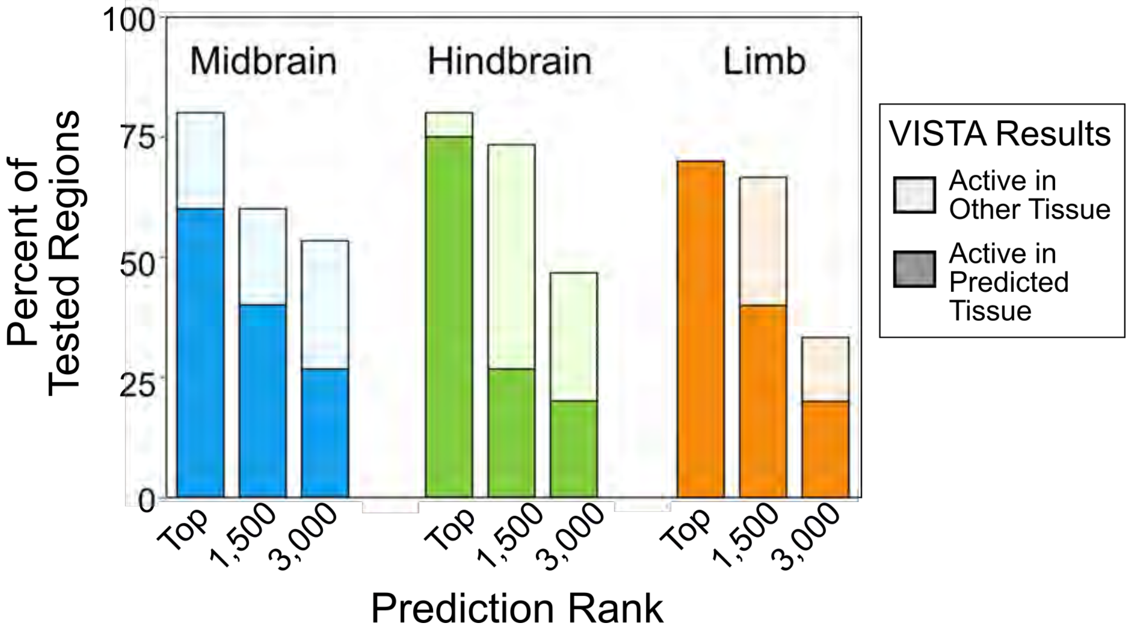


Figure 2.6 | Validation rates of predicted enhancer-like regions using transgenic mouse assays. Bars indicated the percent of tested regions that were positive in the transgenic mouse enhancer assay. Dark colors indicate the region is active in the predicted tissue (blue for midbrain, green for hindbrain, and orange for limb). The lighter color indicates a lack of activity in the predicted tissue with activity in other tissues.

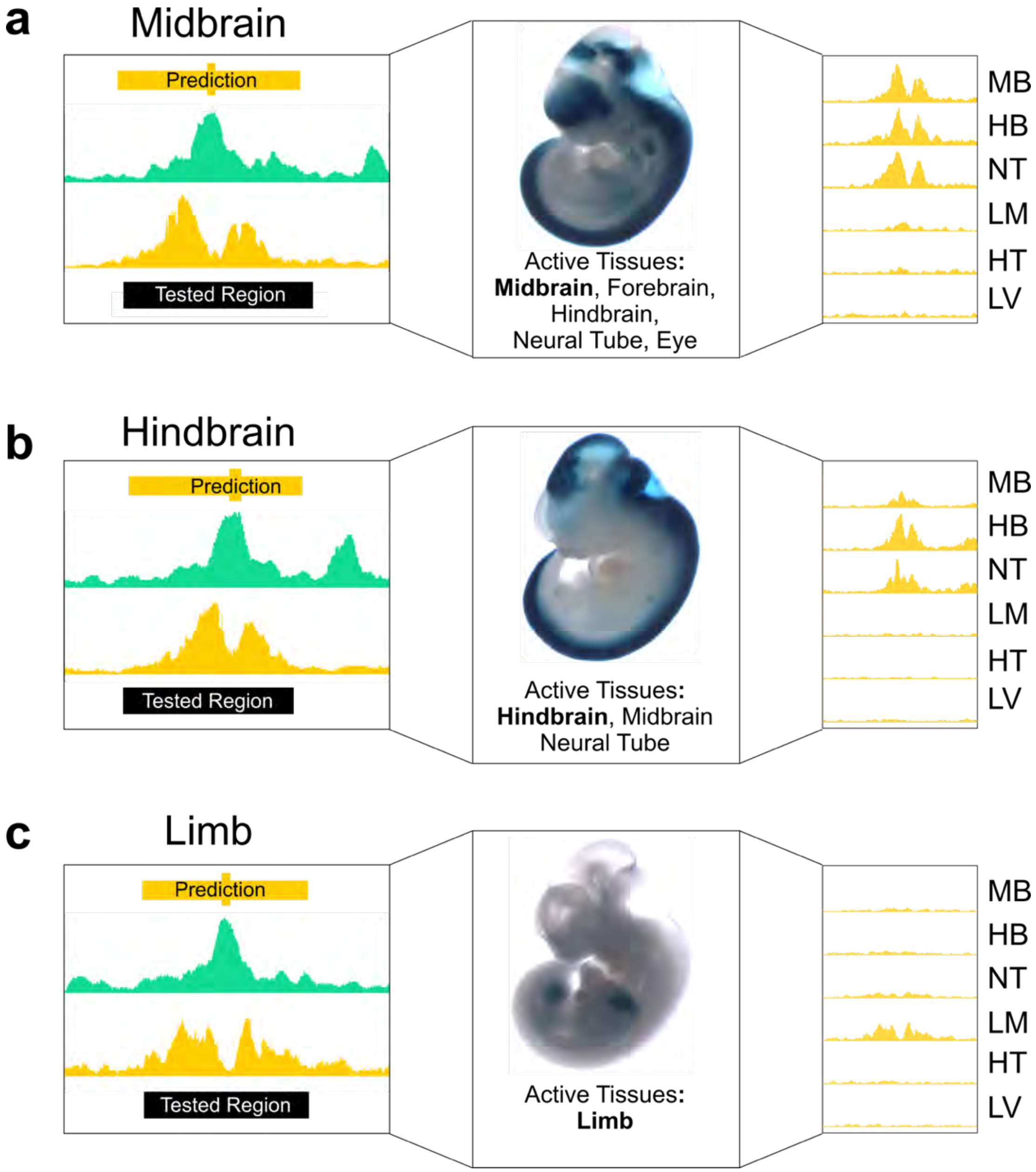


Figure 2.7 | Examples of predicted enhancer-like regions which tested positive in the transgenic mouse assays. Enhancer-like regions predicted using DNase signal (green) and H3K27ac signal (orange) in **a,** midbrain, **b,** hindbrain and **c**, limb. H3K27ac signal on accurately predicts additional observe activity in midbrain (MB), hindbrain (HB), neural tube (NT), limb (LM), heart (HT), and/or liver (LV)

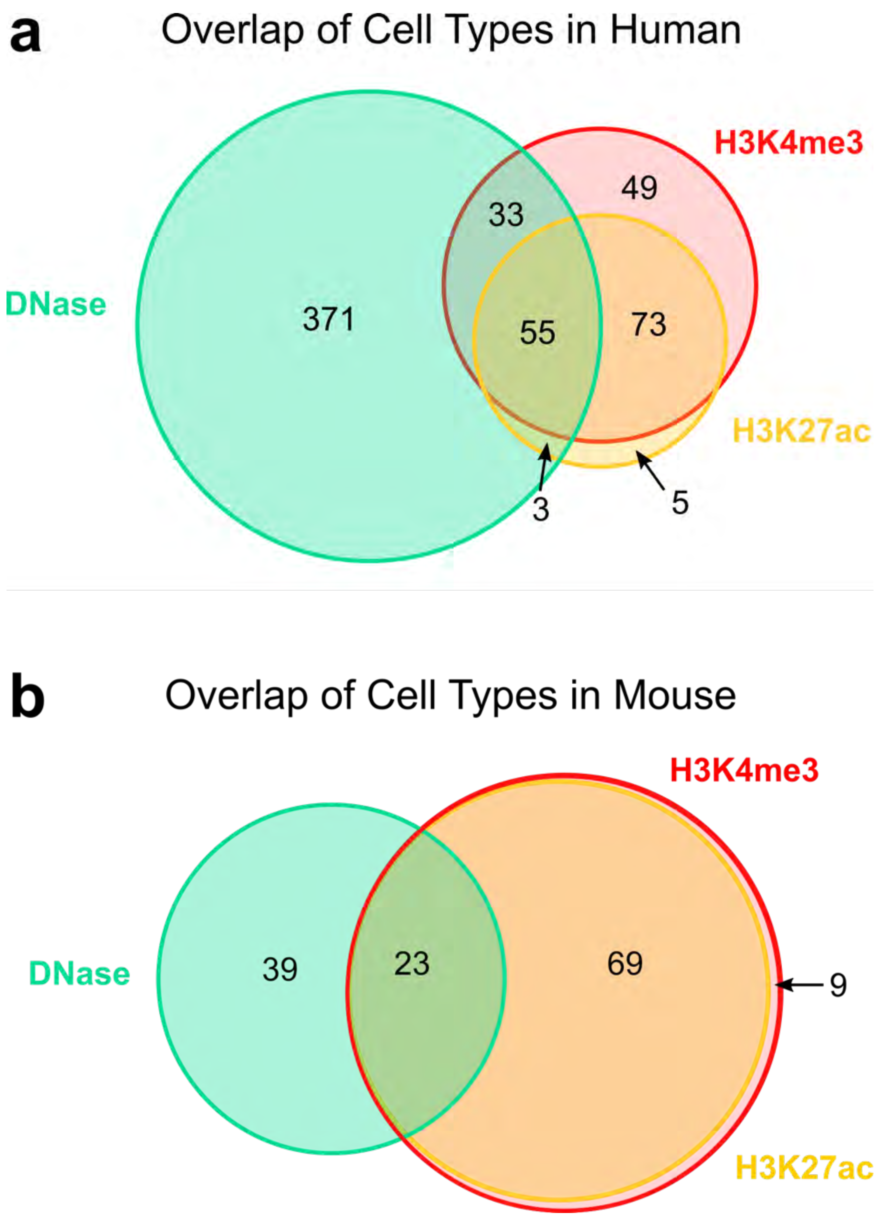


Figure 2.8 | Overlap of cell types with epigenomic datasets. Venn diagrams indicate the number of cell types that have either DNase-seq, H3K4me3 ChIP-seq, and/or H3K27ac ChIP-seq data in **a,** human and **b,** mouse.

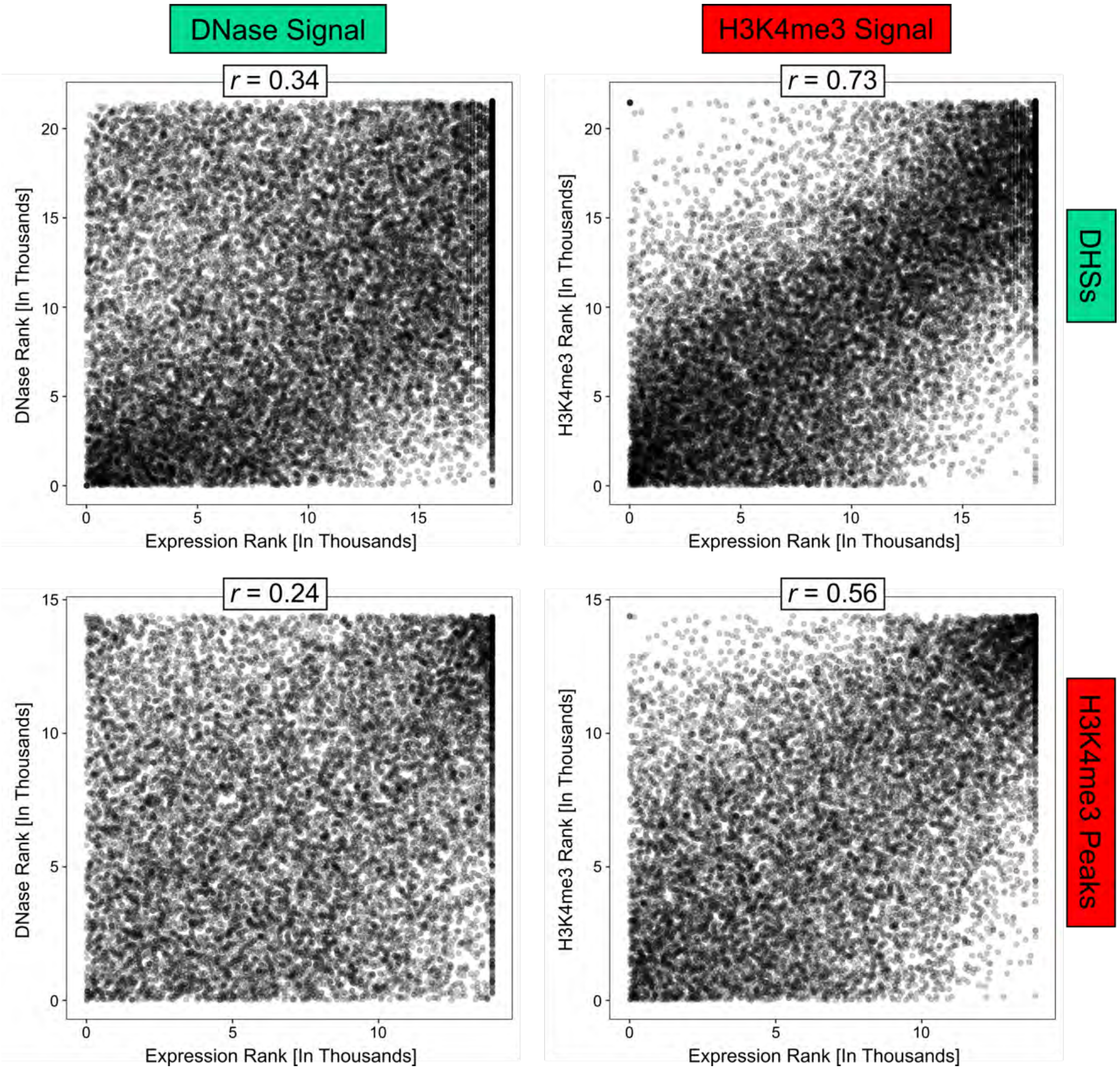


Figure 2.9 | Correlation of gene expression with epigenomic signals to predict promoter-like regions in mouse hindbrain e11.5. Scatterplots demonstrating correlation of expression with **a)** DHSs ranked by DNase signal (*r* = 0.34), **b)** DHSs ranked by H3K4me3 signal (*r* = 0.73), **c)** H3K4me3 peaks ranked by DNase signal (*r* = 0.24), and **d)** H3K4me3 peaks ranked by H3K4me3 signal (*r* = 0.56).

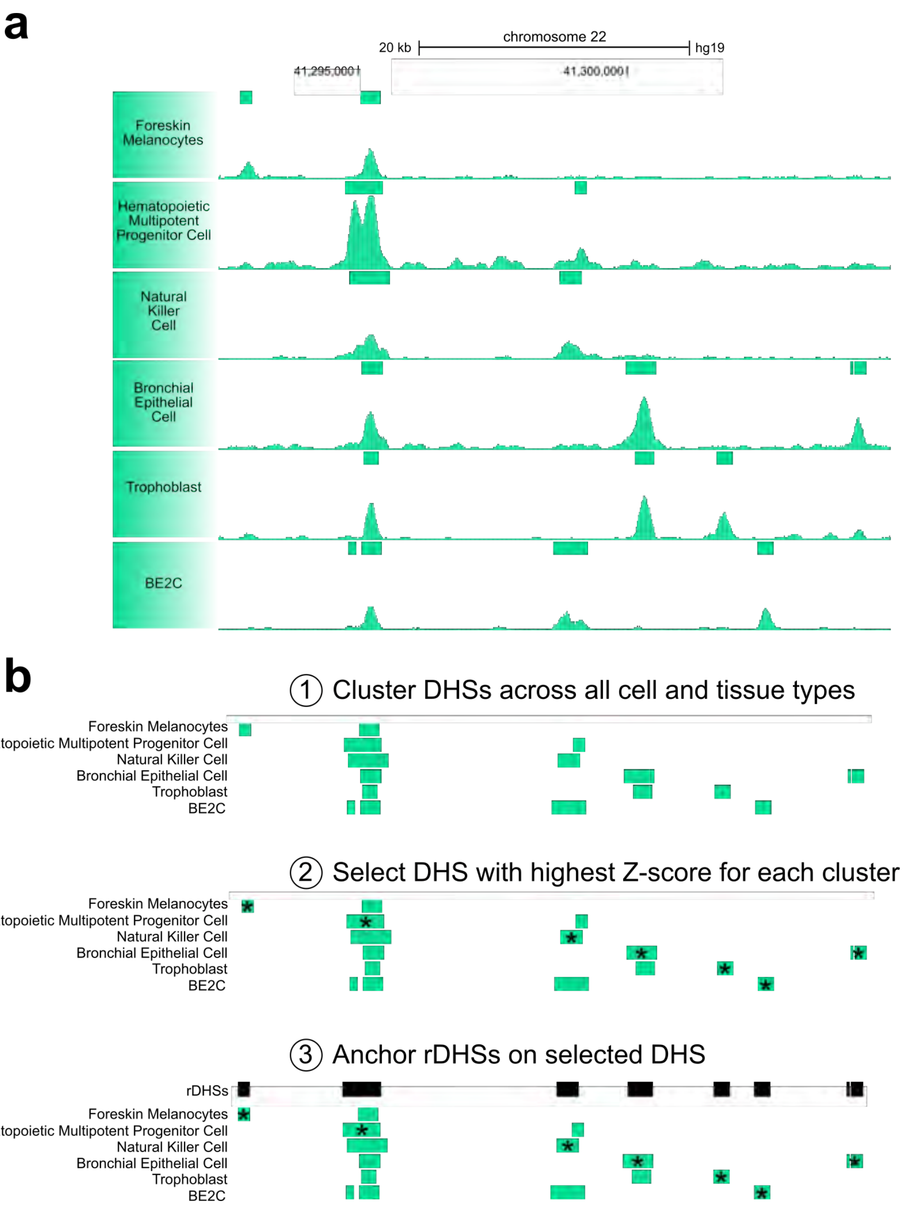


Figure 2.10 | Method for creating representative DNase hypersensitivity sites (rDHSs). a, DHSs across cell types tend to have similar boundaries. **b,** Method for generating rDHSs. We cluster DHSs if they overlap and then for each cluster select the DHS with the highest signal Z-score. This DHS serves as the representative DHS (rDHS) for the cluster. We iteratively repeat this process until all DHS overlap at least one rDHS.

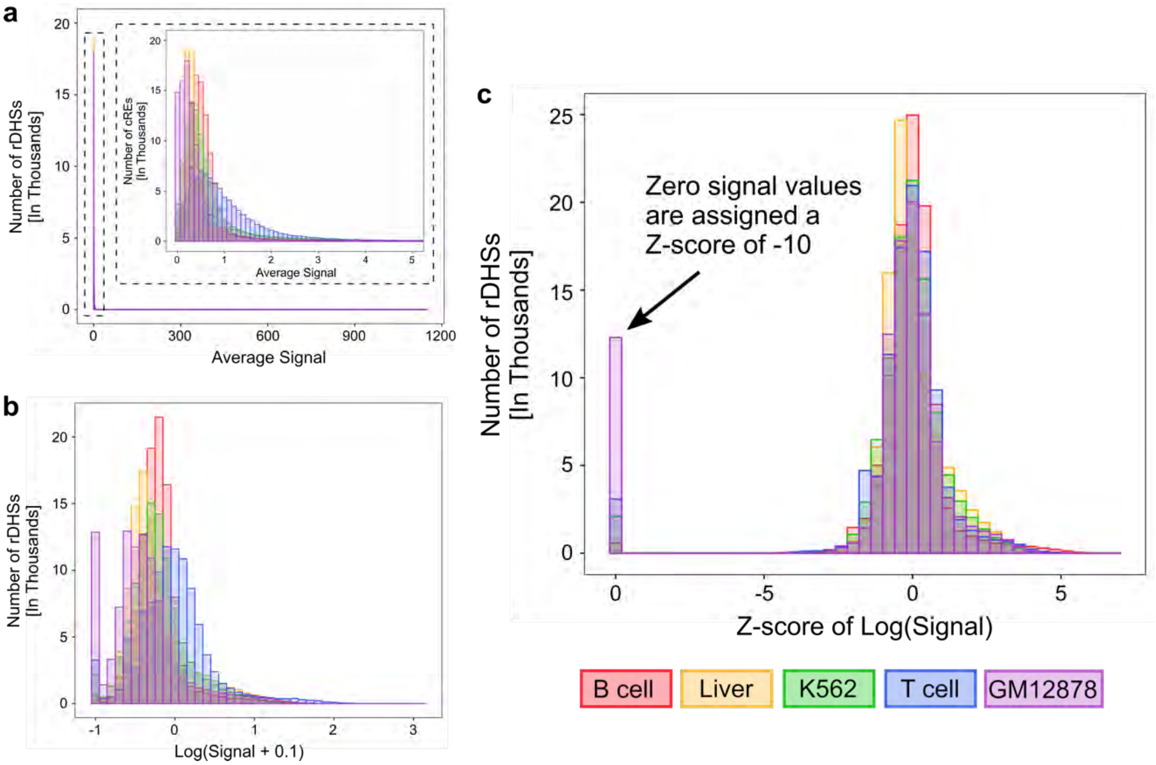


Figure 2.11 | Method for normalizing genomic signals. **a,** Distribution of the H3K27ac signals at rDHSs from five cell types (B cell, Liver, K562, T cell, and GM12878; shown in different colors). **b,** Distributions of the log of the H3K27ac signals in **a**. Individually, log(signal) values of the rDHSs in each cell type roughly follow a normal distribution. **c,** Distribution of the Z-scores corresponding to the log(signal) values in **b**. Signal values of zero are assigned a Z-score of –10.

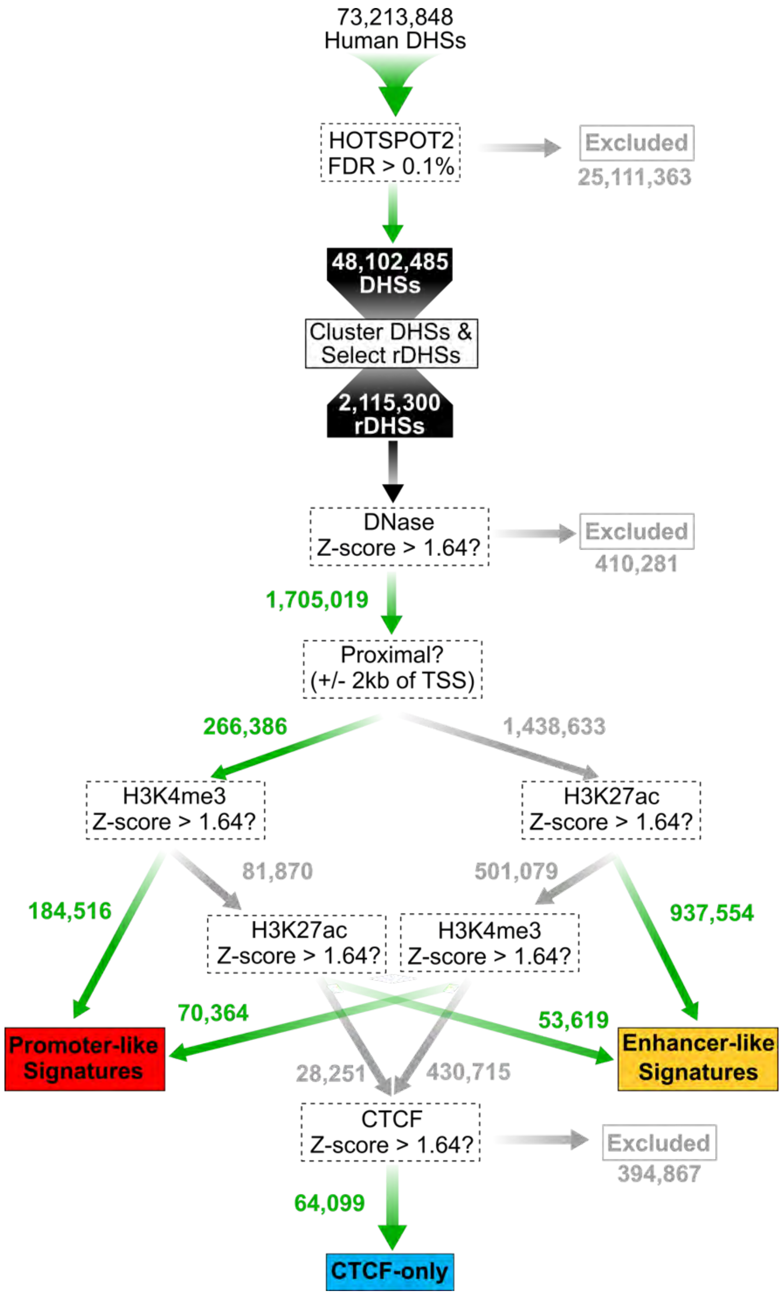


Figure 2.12 | Classification scheme for human cRES (hg19). We begin by clustering high-quality DHSs (FDR > 0.1%) to create representative DHSs (rDHSs). For each assay (DNase, H3K4me3, H3K27ac or CTCF), we calculate a Z-score for every rDHS in a particular cell or tissue type. We then obtain the maximum Z-score across all cell types which we denote the Max-Z. We use the decision tree to classify cREs into three cell-type-agnostic groups according to their Max-Z and proximity to the nearest TSS, including cREs with promoter-like signatures (cREs-PLS, n = 254,880), cREs with enhancer-like signatures (cREs-ELS, n = 991,173), and cREs bound by CTCF only (n = 64,099). The three groups comprise 1,310,152 cREs

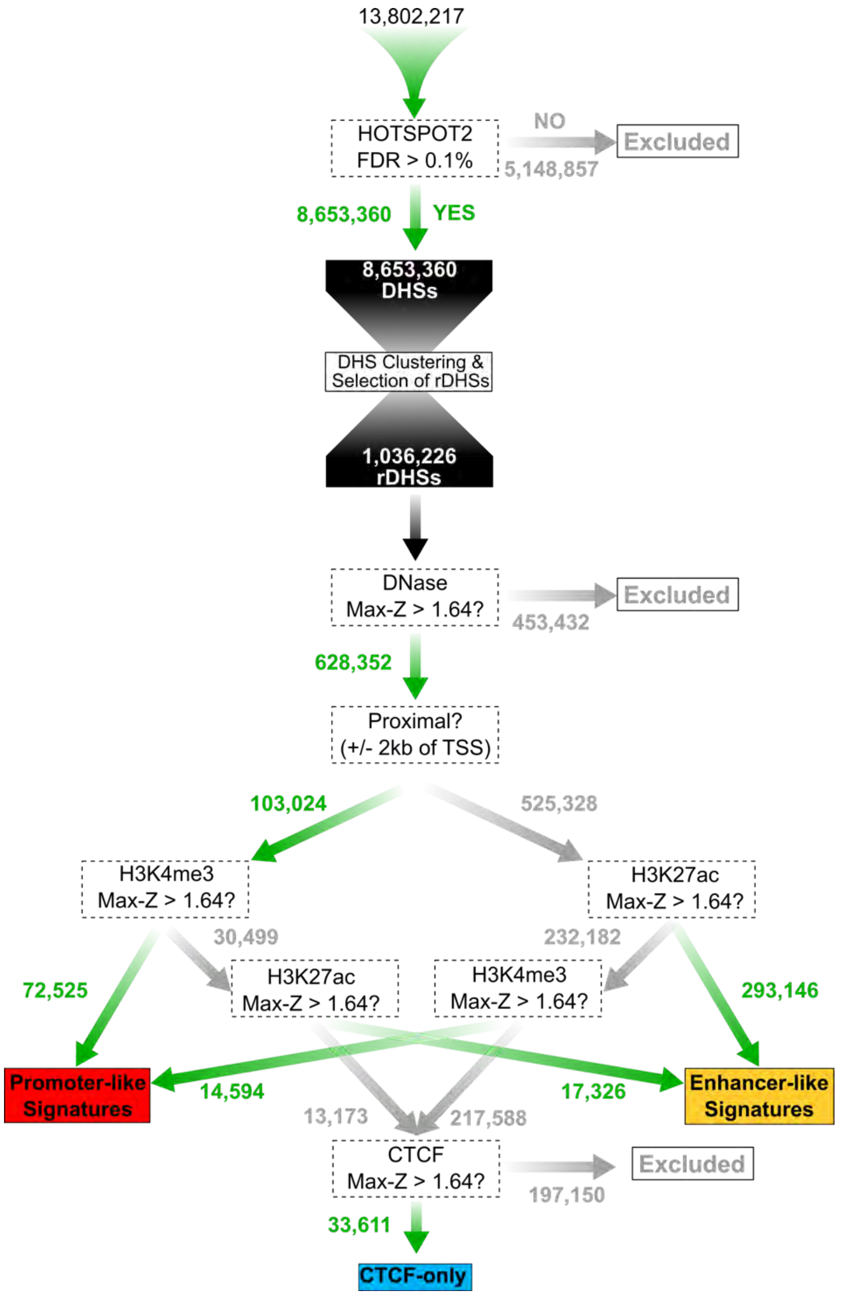


Figure 2.13 | Classification scheme for mouse cRES (mm10). We begin by clustering high quality DHSs (FDR > 0.1%) to create representative DHSs (rDHSs). For each assay (DNase, H3K4me3, H3K27ac or CTCF), we calculate a Z-score for every rDHS in a particular cell or tissue type. We then obtain the maximum 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 = 87,119), cREs with enhancer-like signatures (n = 310,472), and cREs bound by CTCF only (n = 33,611). The total number of cREs is the sum of the three groups: 431,202.

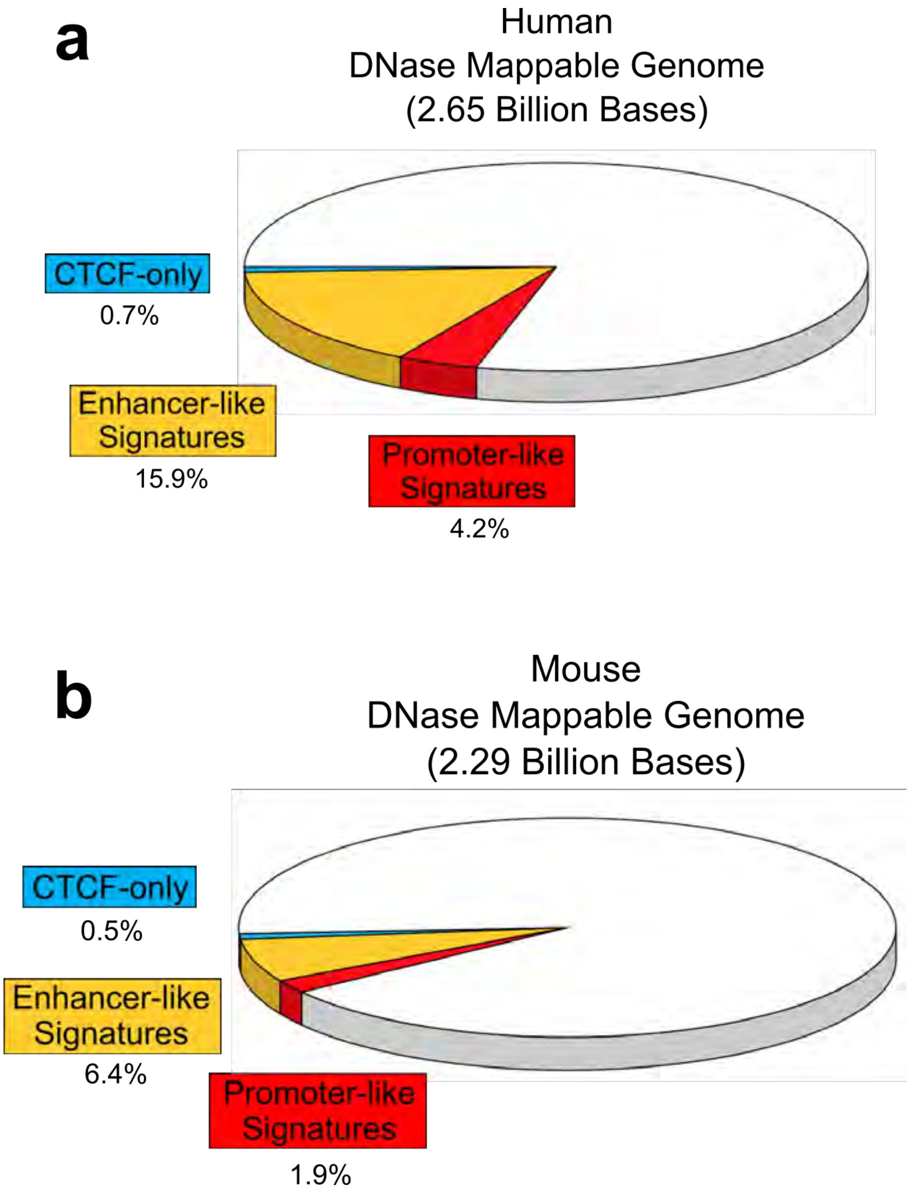


Figure 2.14 | Coverage of the registry of cREs. Percent of the DNase-mappable (36 nt, single-end reads) genome covered by each group of cREs in **a,** human and **b,** mouse.

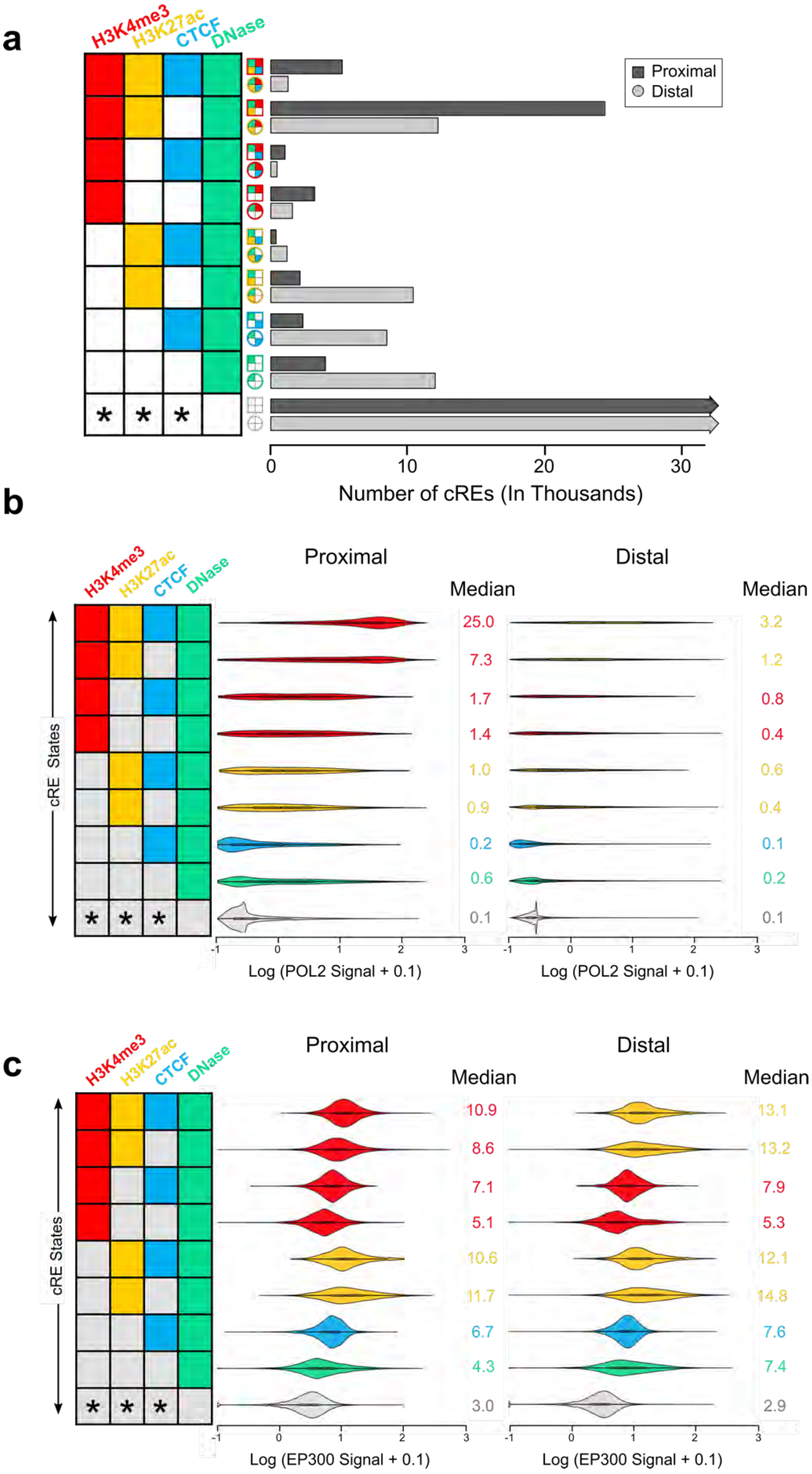

a
H3K4me3
H3K27ac
CTCF
DNase
Proximal
Distal
*
*
*
0
10
20
30
Number of cREs (In Thousands)
b
H3K4me3
H3K27ac
CTCF
DNase
Proximal
Distal
Median
Median
cRE States
25.0
7.3
1.7
1.4
1.0
0.9
0.2
0.6
0.1
3.2
1.2
0.8
0.4
0.6
0.4
0.1
0.2
0.1
*
*
*
-1
0
1
2
3
Log (POL2 Signal + 0.1)
Log (POL2 Signal + 0.1)
c
H3K4me3
H3K27ac
CTCF
DNase
Proximal
Distal
Median
Median
cRE States
10.9
8.6
7.1
5.1
10.6
11.7
6.7
4.3
3.0
13.1
13.2
7.9
5.3
12.1
14.8
7.6
7.4
2.9
*
*
*
-1
0
1
2
3
Log (EP300 Signal + 0.1)
Log (EP300 Signal + 0.1)

Figure 2.15 | Nine states of cell-type specific cREs in GM12878 **a**, Number of GM12878 cREs in each group **b,** Violin plots show the average POLII 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. **c,** 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.

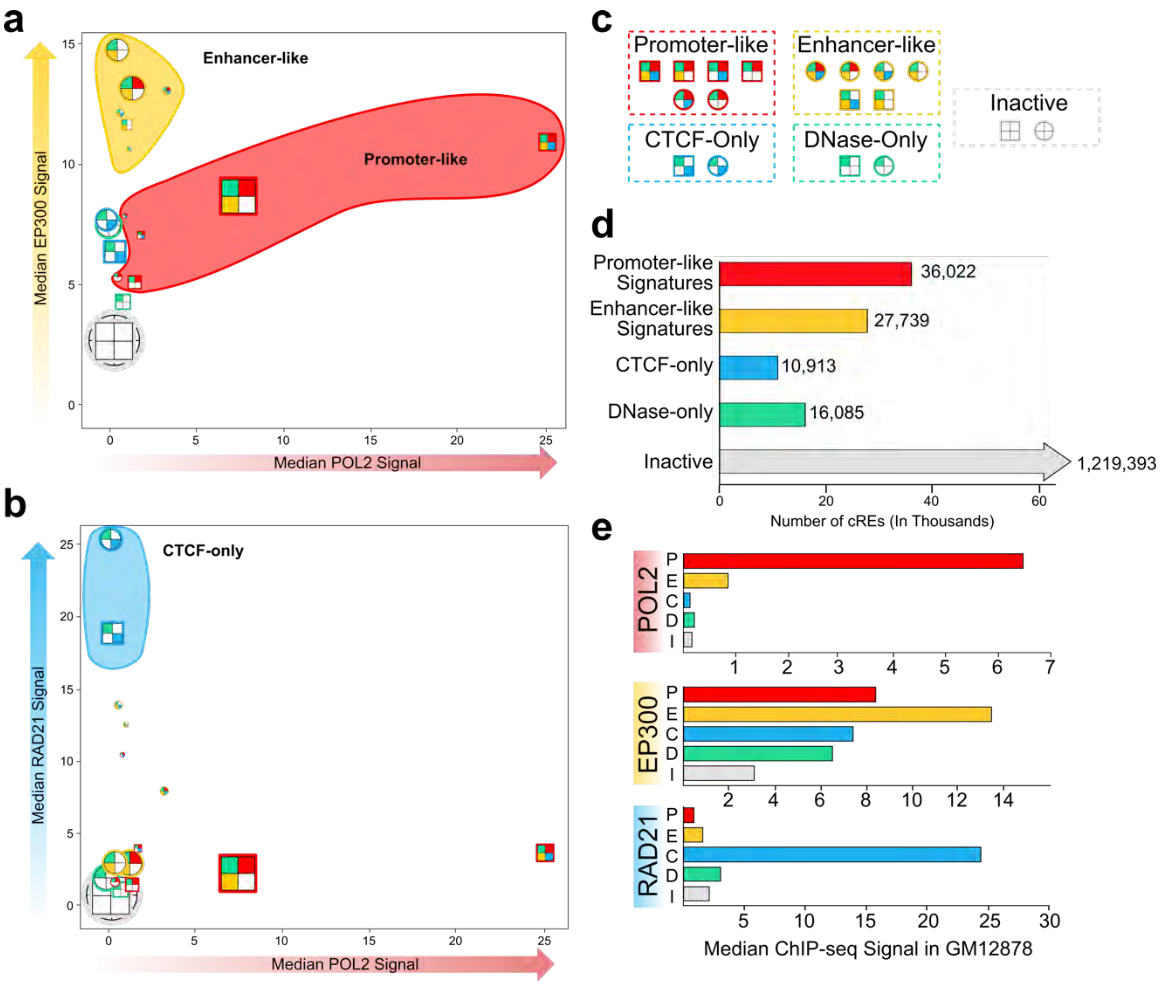


Figure 2.16 | Overview of 5 group classification method. Scatterplots of **a,** median EP300 signal or **b,** median RAD21 signal vs. median POLII signal for each cRE state in GM12878. The size of an icon is proportional to the number of cREs in that state except for the inactive state. Proximal cREs are represented by square icons. Distal cREs are represented by circular icons. **c**, Assignment of cRE states to the five following groups: with promoter-like signatures, with enhancer-like signatures, CTCF-only, DNase-only, and inactive. **d,** Number of GM12878 cREs in each group. **e**, Median ChIP-seq signal for POLII, EP300 and RAD21 in GM12878 for the cREs in each group.

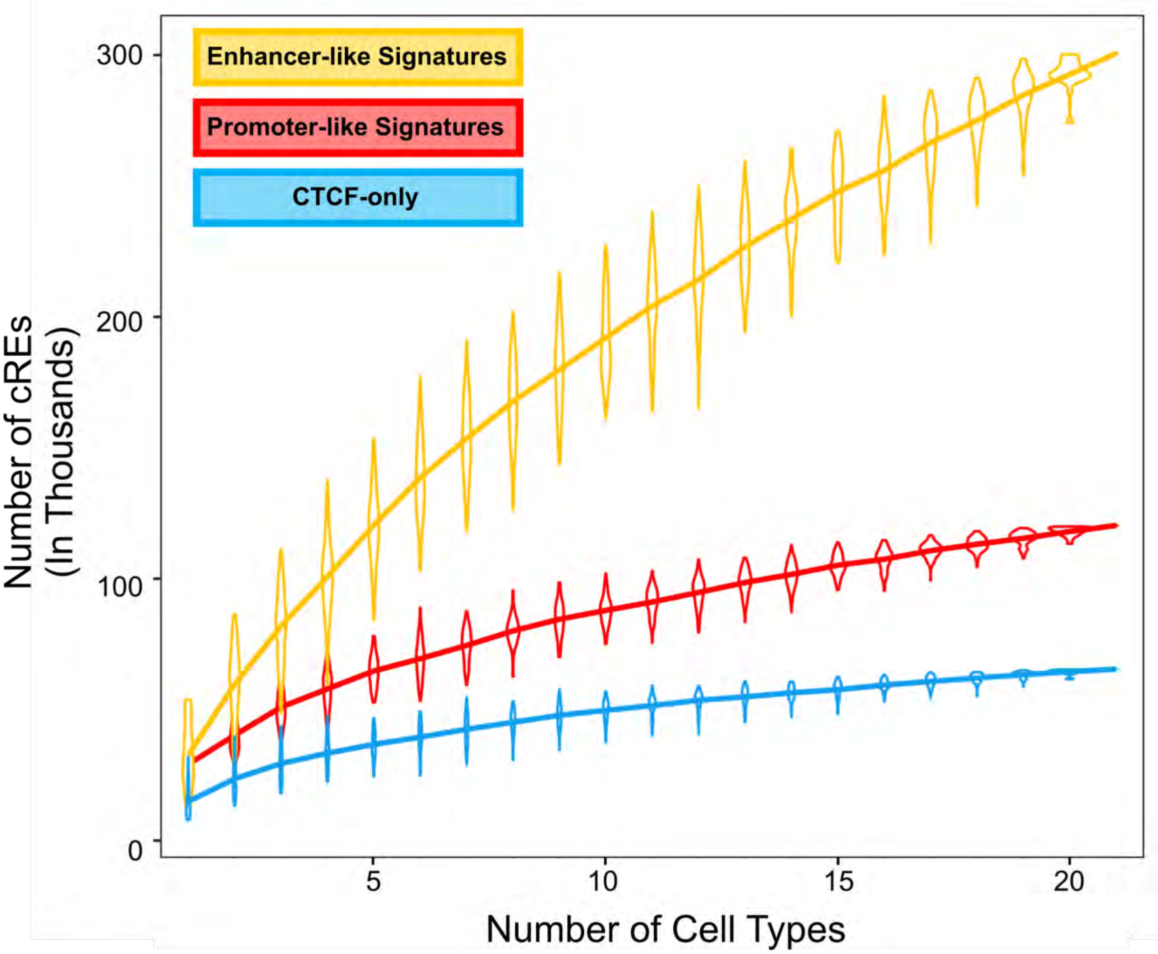


Figure 2.17 | Saturation of cREs across 21 cell types with all four datatypes. Total numbers of cREs with Promoter-like, Enhancer-like, or CTCF-only signatures grow when more cell types are considered. Enhancer-like cREs are more cell-type-restrictive than promotor-like cREs or CTCF-only cREs.

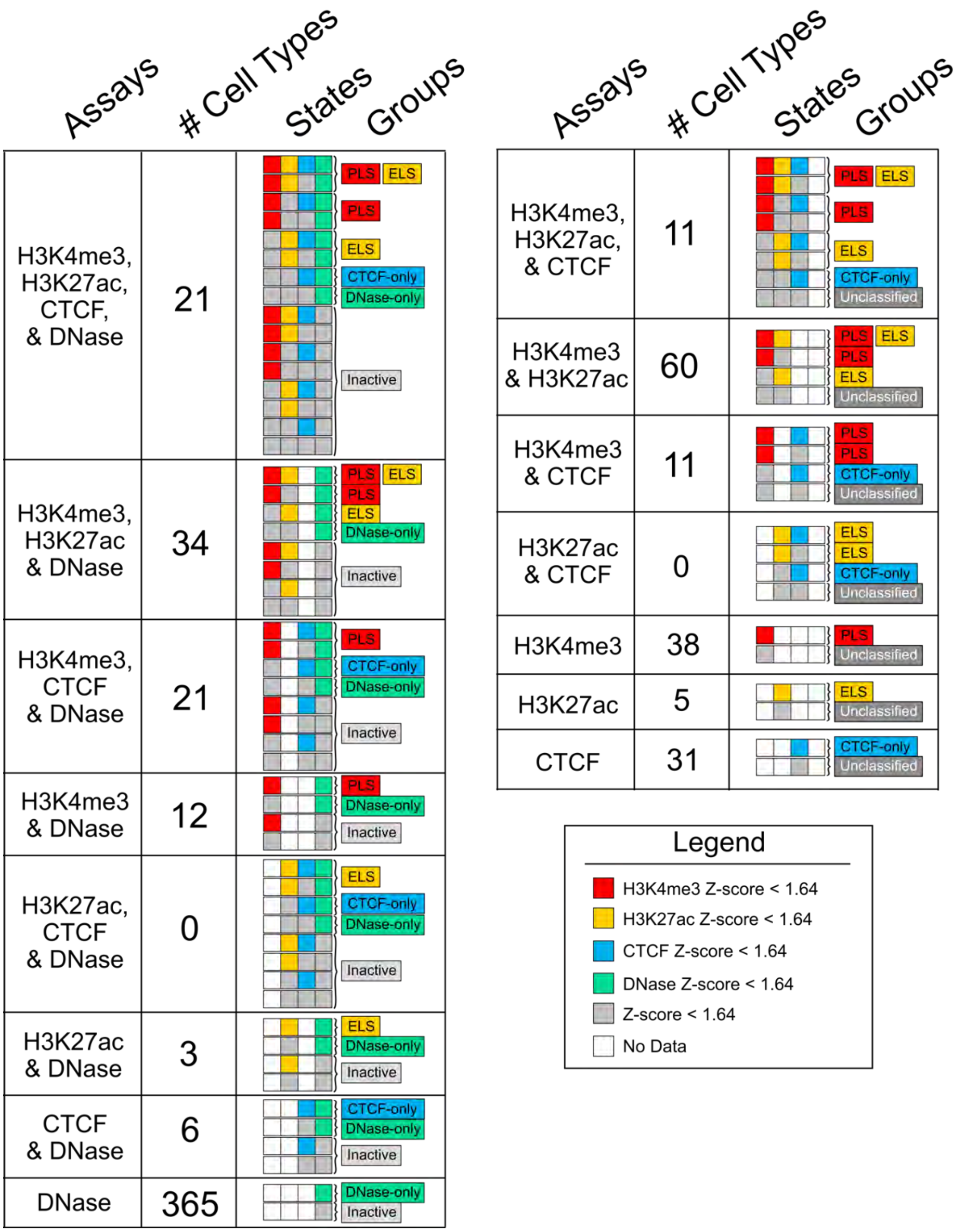


Figure 2.18 | Cell-type specific annotations of the Registry of cREs. Scheme for translating cell type specific state classifications into group classifications for cell types with different combinations of datasets.

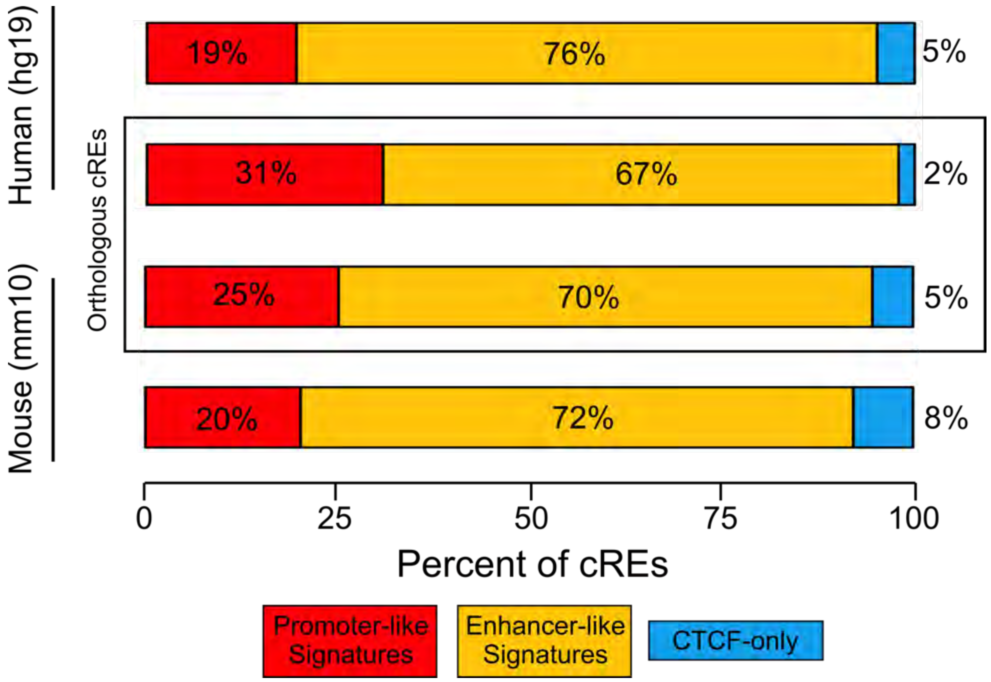


Figure 2.19 | Groups of orthologous cREs. Percentage of cREs with promoter-like (red), enhancer-like (yellow), or CTCF-only (blue) signatures for human and mouse cREs as well as orthologous human and mouse cREs.

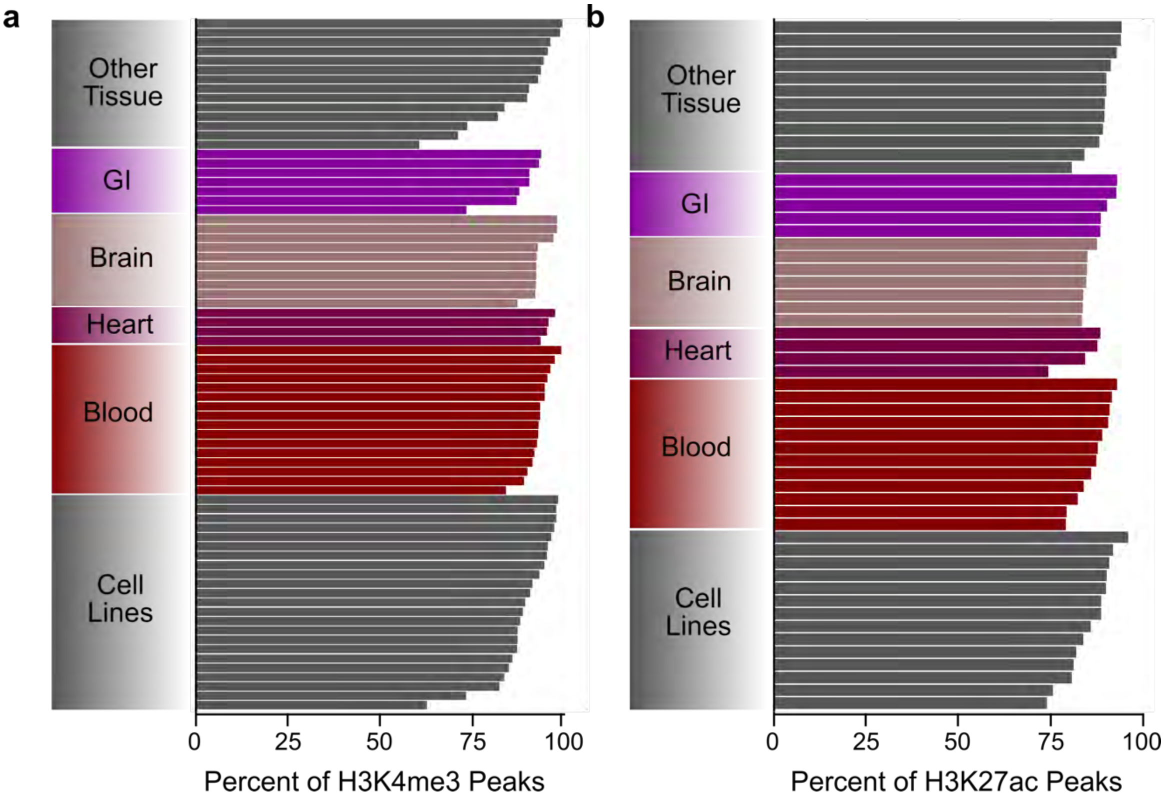


Figure 2.20 | Coverage of histone modification peaks by the current human Registry of cREs. Overlap of cREs with **a,** H3K4me3 peaks and **b,** H3K27ac peaks and **c,** CTCF peaks from cell types without DNase data. On average 89.7% and 86.8%, H3K4me3 and H3K27ac peaks overlap a cRE, respectively.

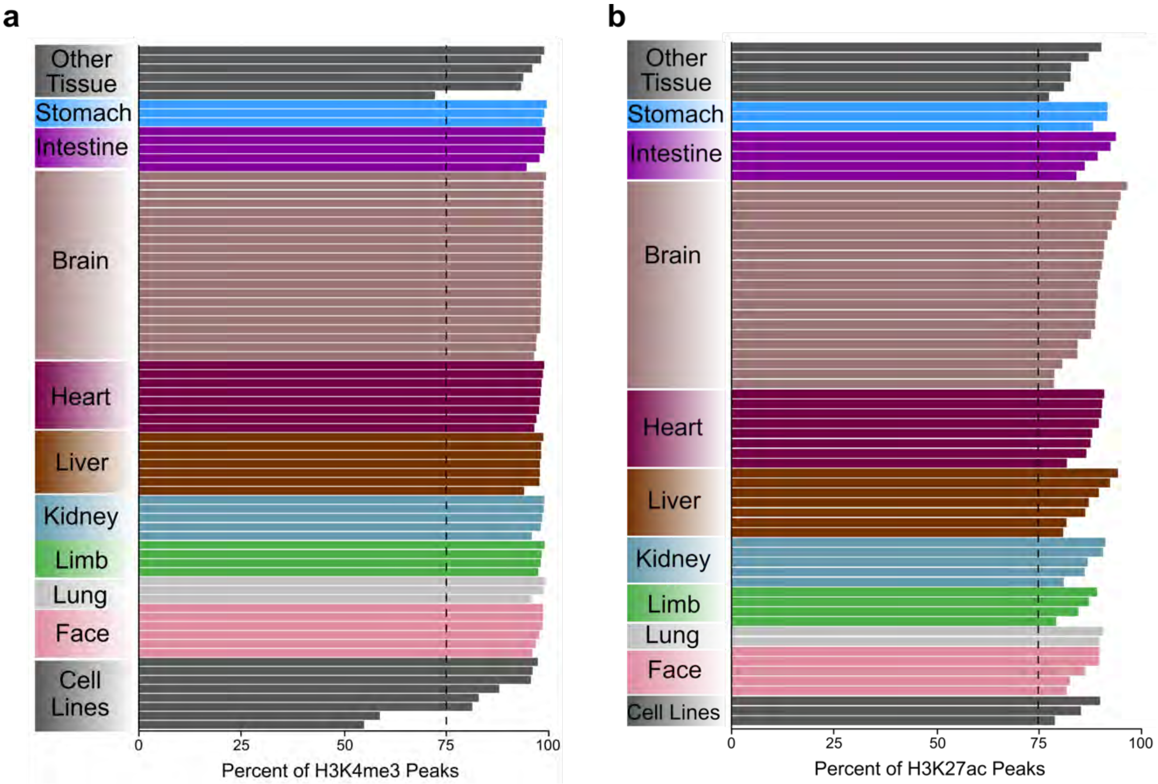


Figure 2.21 | Coverage of histone modification peaks by the current mouse Registry of cREs. Overlap of cREs with **a,** H3K4me3 peaks and **b,** H3K27ac peaks from cell types without DNase data. On average 95.8% and 87.6% of H3K4me3 and H3K27ac peaks overlap a cRE, respectively.

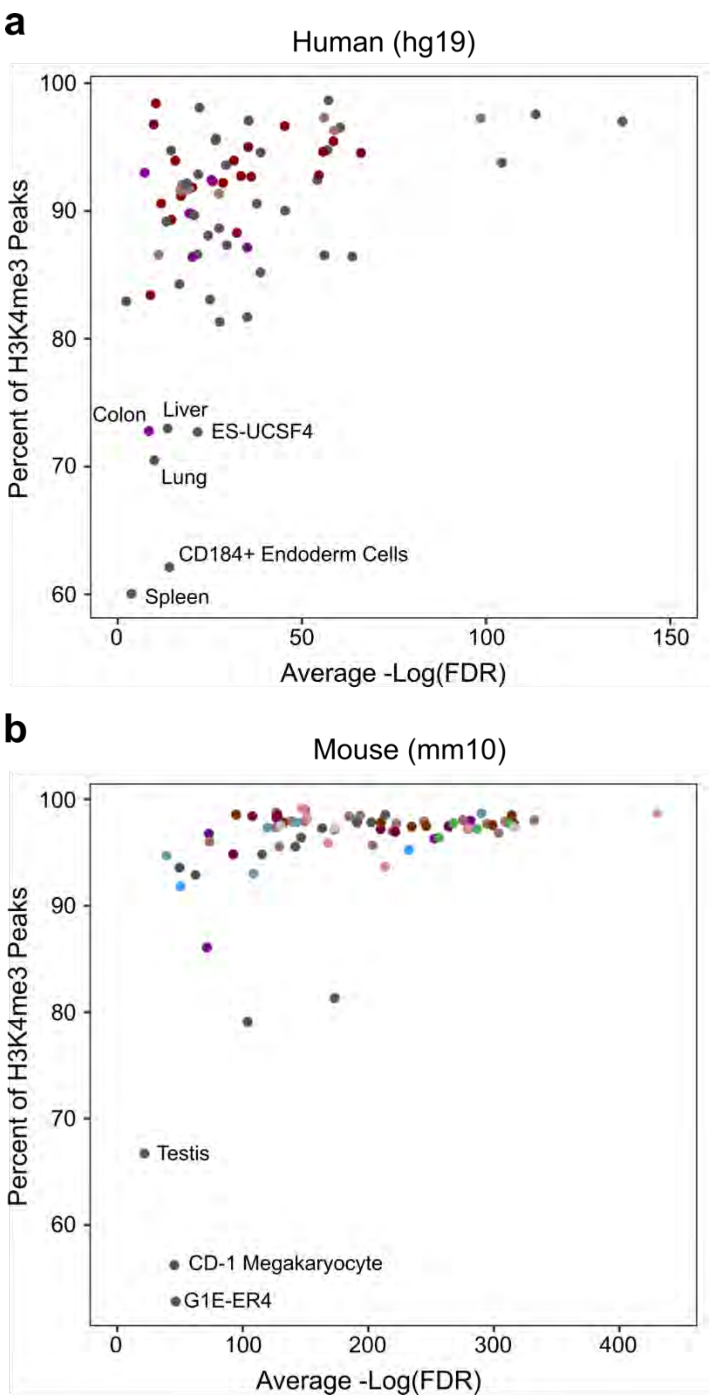


Figure 2.22 | Coverage of the H3K4me3 peaks by the current Registry of cREs is plotted against the average -log(FDR) of the H3K4me3 peaks. In **a,** human and **b,** mouse, cell-types with peaks that have a lower average -log(FDR) across all peaks tend to have a lower percentage of peaks covered. Manual inspection reveals that this lower coverage is due to lower-signal, false-positive peaks called by the algorithm for these datasets.

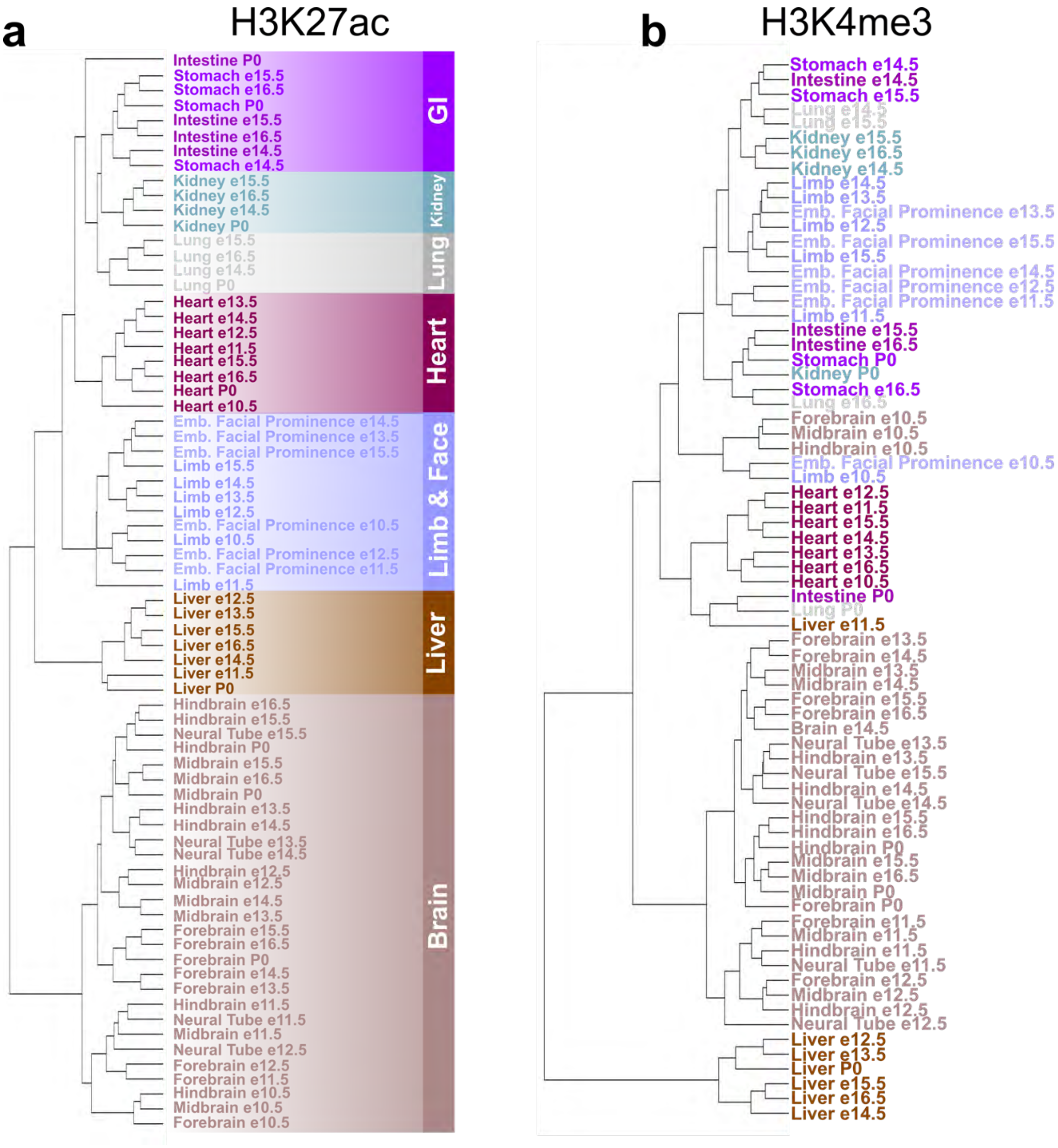


Figure 2.23 | Clustering of mouse cell types on the basis of cRE histone modification activity. Mouse embryonic tissues were hierarchically clustered according to the Jaccard similarity coefficient of cREs with high **a,** H3K27ac and **b,** H3K4me3 Z-scores. Colors indicate the organs of origin of the tissues. When clustered according to H3K27ac signals at cREs (panel **a**), the tissues segregate completely according to their organs of origin.

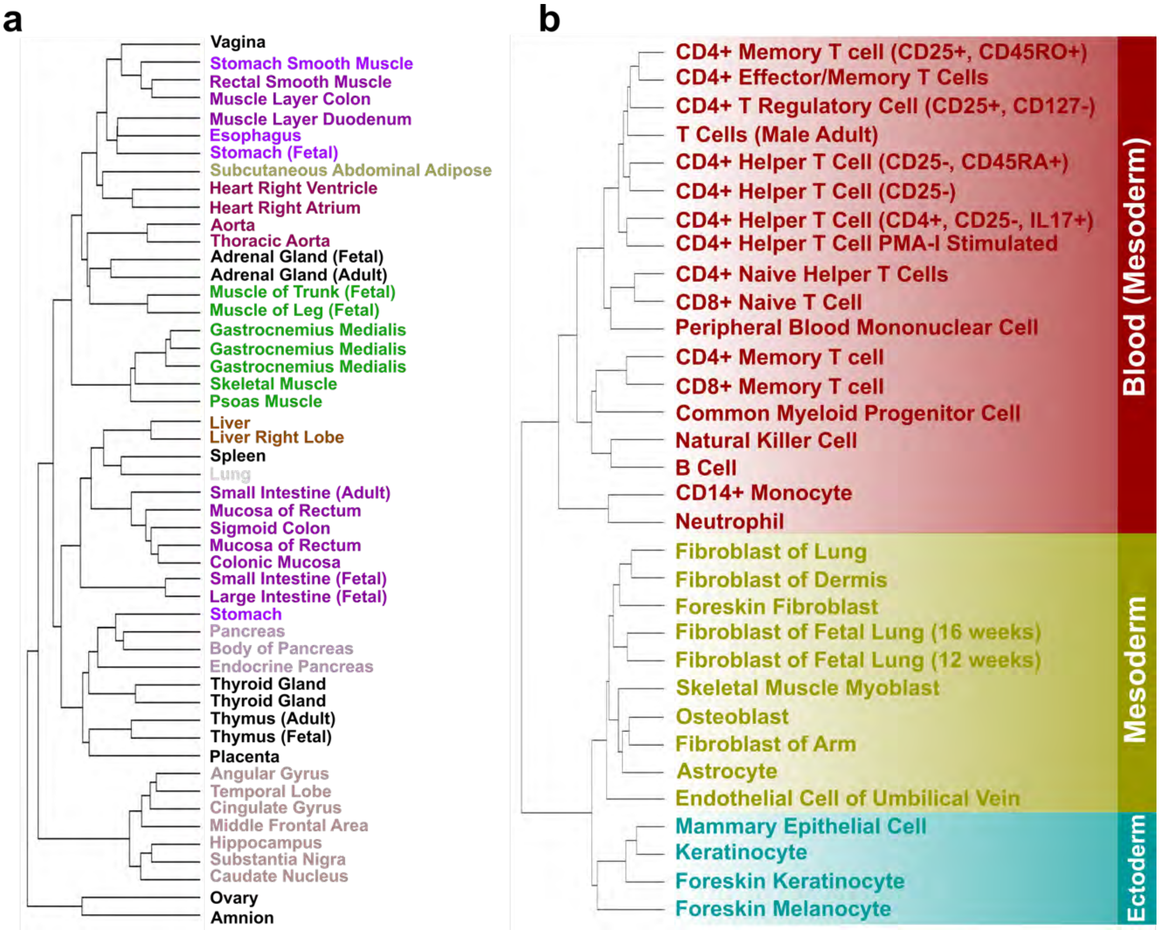


Figure 2.24 | Clustering of human cell and tissue types on the basis of cRE H3K27ac signal. Human **a,** primary cells and **b,** tissues were hierarchically clustered according to the Jaccard similarity coefficient of cREs with a high H3K27ac signal (Z-score > 1.64). The tissue samples in **a** are colored by their organ of origin and the primary cells in **b** are colored according to their lineages.

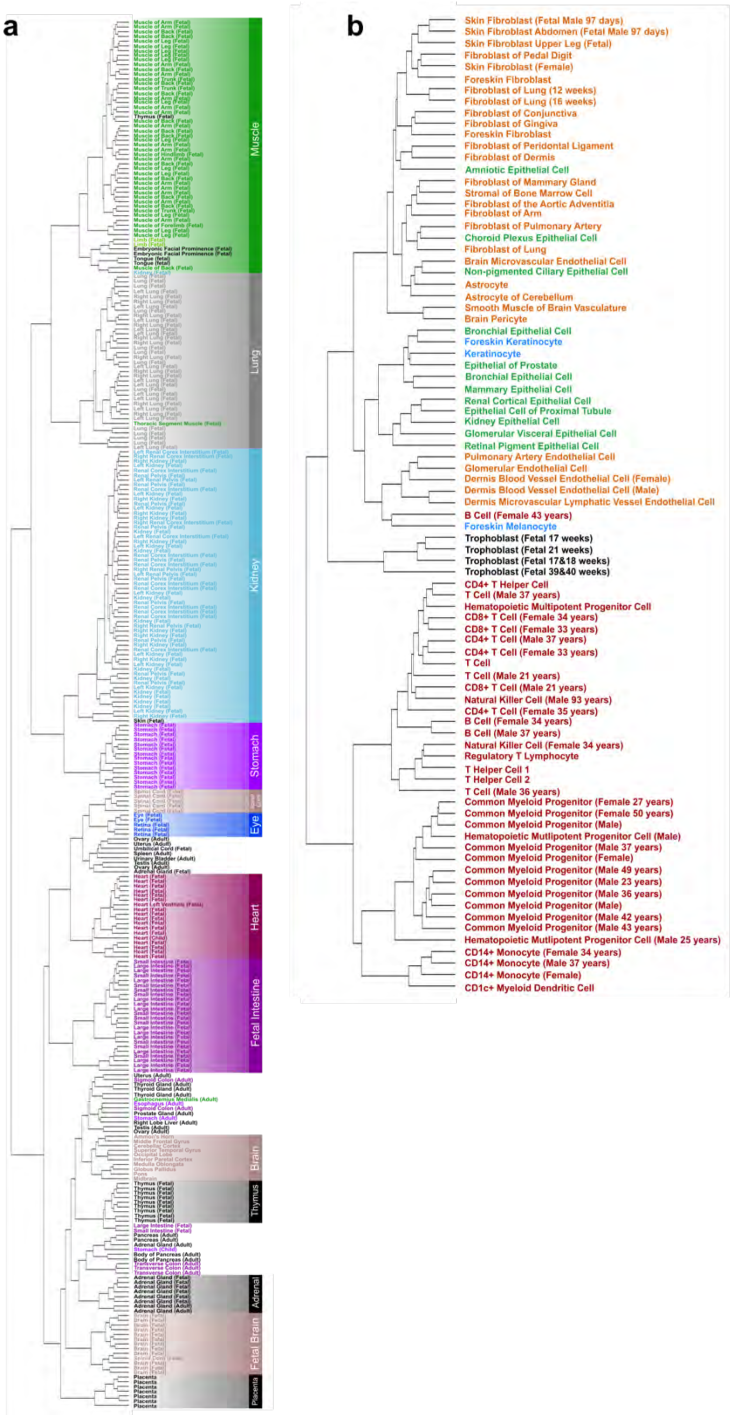


Figure 2.25 | Clustering of human cell and tissue types on the basis of cRE DNase signal. Human **a,** tissues and **b,** primary cells hierarchically clustered according to the Jaccard similarity coefficient of cREs with a high DNase signal (Z-score > 1.64). The tissue samples in **a** are colored by their organ of origin and the primary cells in **b** are colored according to their lineages.

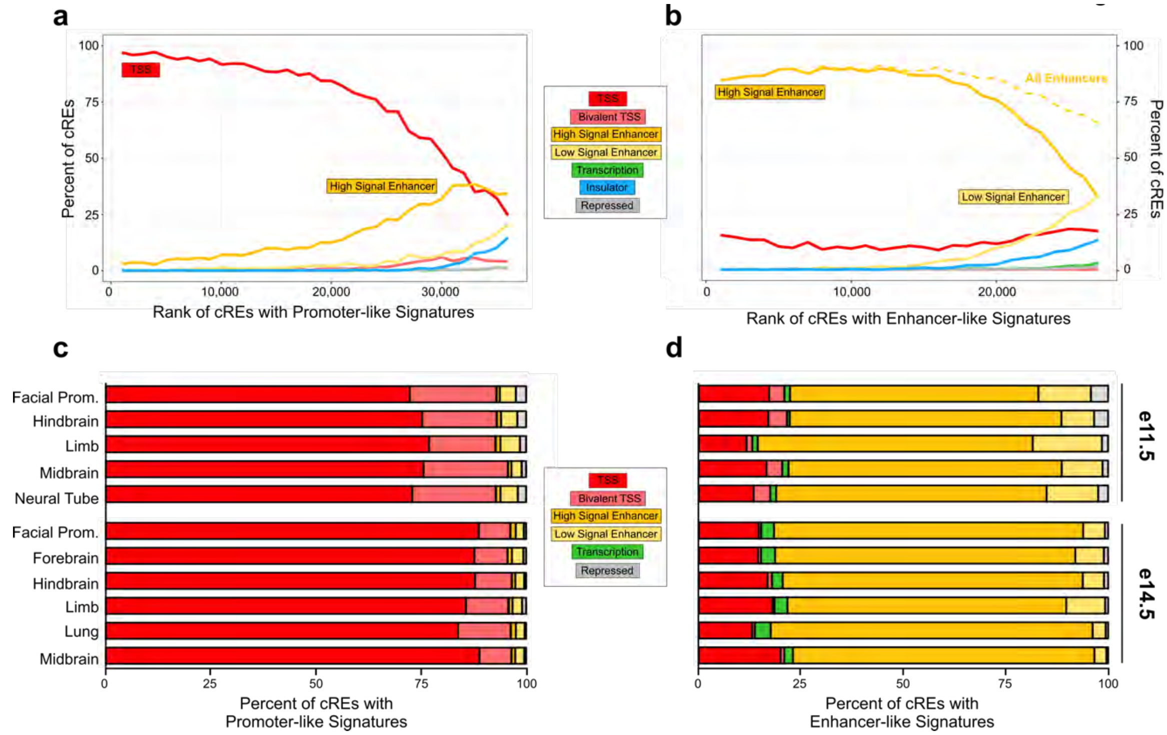


Figure 2.26 | Overlap of cREs with chromHMM states. In GM12878, we ranked cREs with **a,** promoter-like signatures and **b,** enhancer-signatures on the basis of H3K4me3 and H3K27ac Z-scores respectively. For each bin of 1 k cREs, we calculated the percentage of cREs overlapping each chromHMM state. In mouse, we selected all cREs with **c,** promoter-like and **d,** enhancer-like signatures from tissue–time-point combinations with both DNase and histone data. We then calculated the percent of cREs that overlapped each chromHMM state. In all panels, high- and low-signal enhancers denote chromHMM enhancer states with high or low H3K27ac signals.

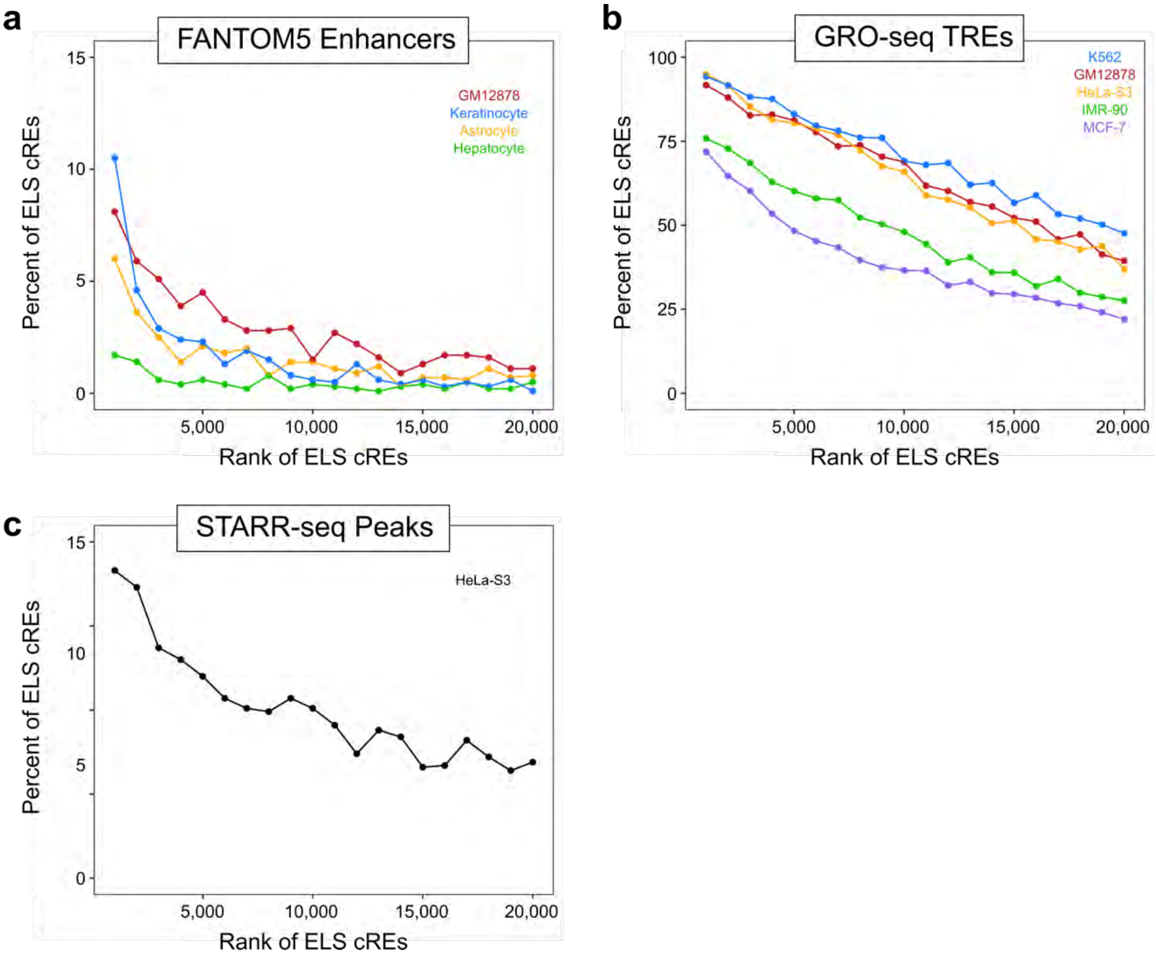


Figure 2.27 | Overlap of ELS cREs with previously predicted enhancers. The percentage of ELS cREs that overlap enhancers predicted by **a,** the FANTOM5 consortium, **b,** GRO-seq data, **c,** and STARR-seq. ELS cREs are ranked based on cell type specific H3K27ac Z-scores.

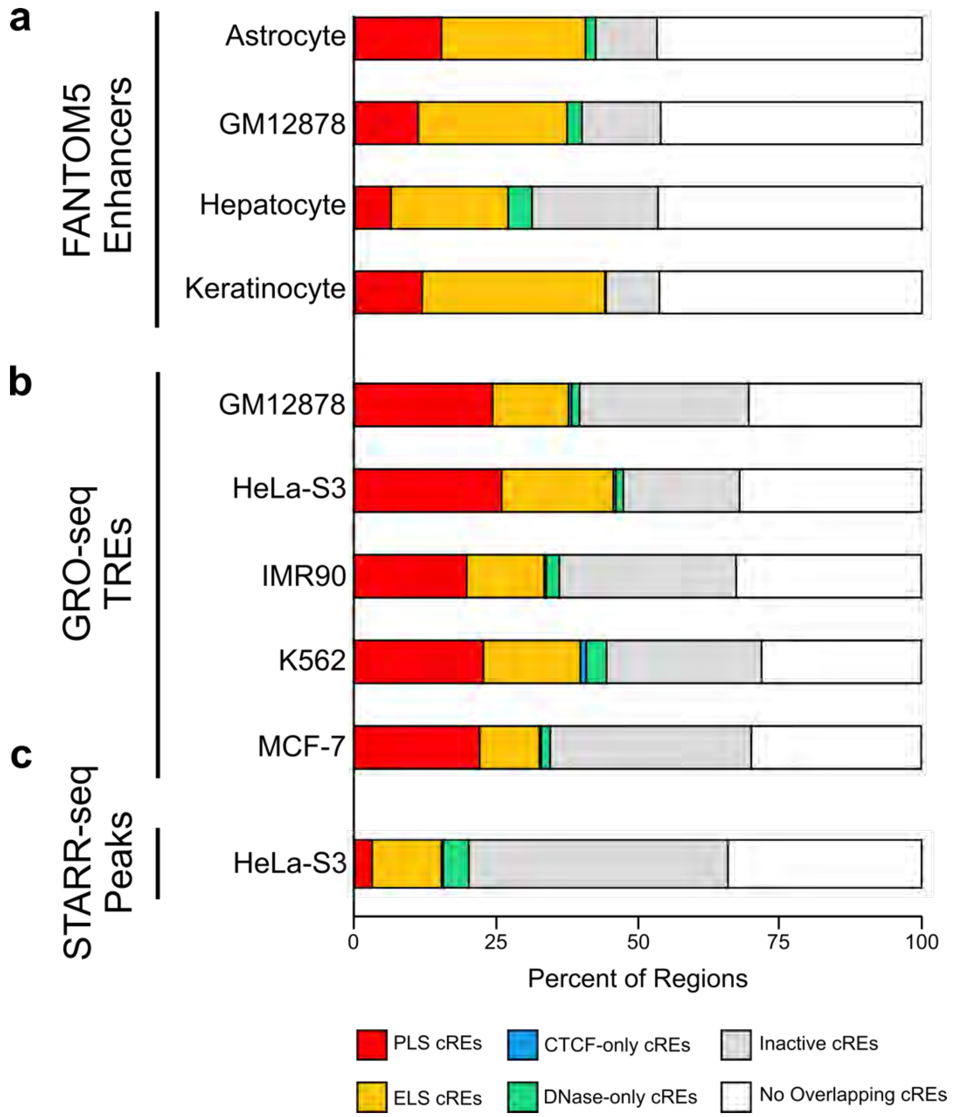


Figure 2.28 | Overlap of previously predicted enhancers with cREs. The percentage of enhancers predicted by **a,** the FANTOM5 consortium, **b,** GRO-seq data, **c,** and STARR-seq that overlap with cREs. Colors indicate groups of cREs active in each cell type: red for PLS, yellow of ELS, blue for CTCF-only, green for DNase-only and gray for inactive. Predicted enhancers that do not overlap any cREs are shown in white.

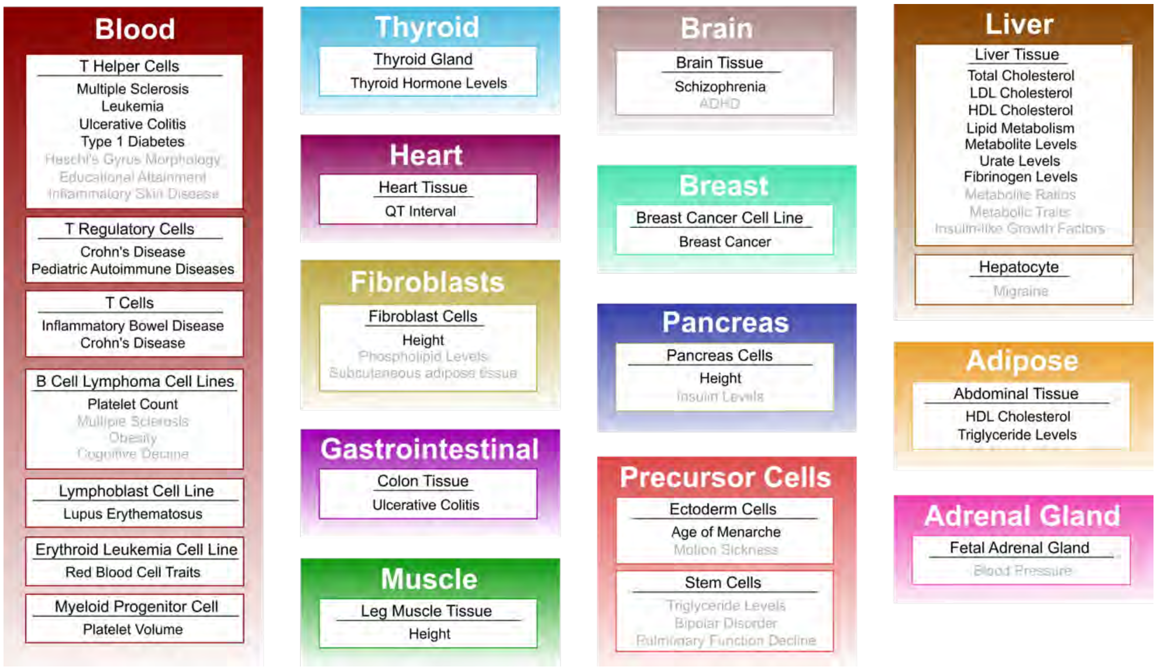
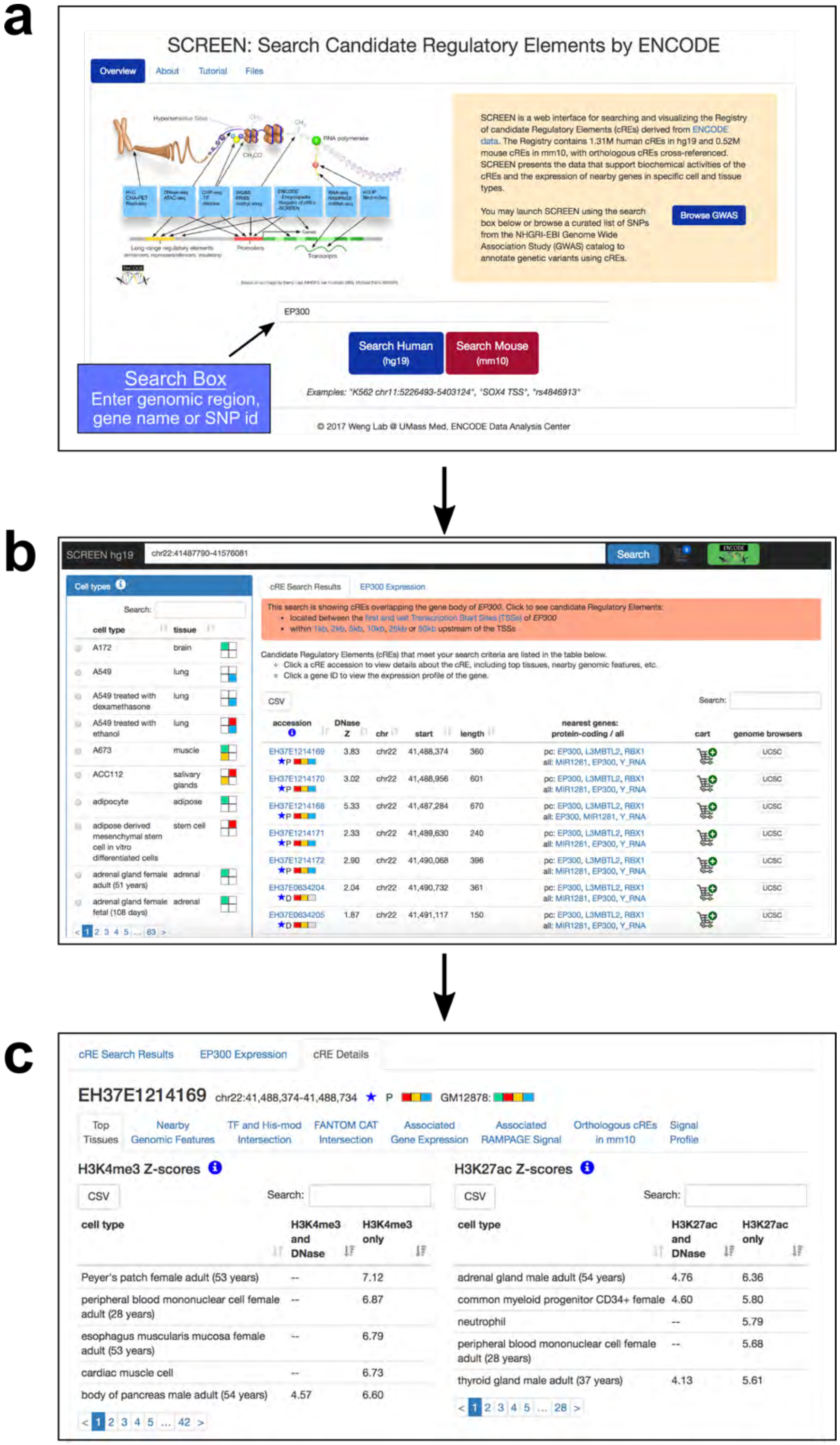


Figure 2.29 | Top cell type enrichments for variants reported by genome wide association studies (GWAS). For each GWAS we report the cell or tissue type of which active cREs are significantly enriched in the disease variants. Cell types that do not meet the FDR threshold of 0.05 are in grey. Most studies have multiple significantly enriched cell types but only the top hit is reported here. Traits listed multiple times are from different studies.



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.

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

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

cell type	tissue
A172	brain
A549	lung
A549 treated with dexamethasone	lung
A549 treated with ethanol	lung
A673	muscle
ACC112	salivary glands
adipocyte	adipose
adipose derived mesenchymal stem cell in vitro differentiated cells	stem cell
adrenal gland female adult (51 years)	adrenal
adrenal gland female fetal (108 days)	adrenal

< 1 2 3 4 5 ... 63 >

This search is showing cREs overlapping the gene body of *EP300*. Click to see candidate Regulatory Elements:

- located between the first and last Transcription Start Sites (TSSs) of *EP300*
- within 1kb, 2kb, 5kb, 10kb, 25kb or 50kb upstream of the TSSs

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.

accession	DNase Z	chr	start	length	nearest genes: protein-coding / all	cart	genome browsers
EH37E1214169 ★P	3.83	chr22	41,488,374	360	pc: EP300, L3MBTL2, RBX1 all: MIR1281, EP300, Y_RNA		UCSC
EH37E1214170 ★P	3.02	chr22	41,488,956	601	pc: EP300, L3MBTL2, RBX1 all: MIR1281, EP300, Y_RNA		UCSC
EH37E1214168 ★P	5.33	chr22	41,487,284	670	pc: EP300, L3MBTL2, RBX1 all: EP300, MIR1281, Y_RNA		UCSC
EH37E1214171 ★P	2.33	chr22	41,489,630	240	pc: EP300, L3MBTL2, RBX1 all: MIR1281, EP300, Y_RNA		UCSC
EH37E1214172 ★P	2.90	chr22	41,490,068	396	pc: EP300, L3MBTL2, RBX1 all: MIR1281, EP300, Y_RNA		UCSC
EH37E0634204 ★D	2.04	chr22	41,490,732	361	pc: EP300, L3MBTL2, RBX1 all: MIR1281, EP300, Y_RNA		UCSC
EH37E0634205 ★D	1.87	chr22	41,491,117	150	pc: EP300, L3MBTL2, RBX1 all: MIR1281, EP300, Y_RNA		UCSC

EH37E1214169 chr22:41,488,374-41,488,734 ★ P GM12878:

H3K4me3 Z-scores

cell type	H3K4me3 and DNase	H3K4me3 only
Peyer's patch female adult (53 years)	--	7.12
peripheral blood mononuclear cell female adult (28 years)	--	6.87
esophagus muscularis mucosa female adult (53 years)	--	6.79
cardiac muscle cell	--	6.73
body of pancreas male adult (54 years)	4.57	6.60

< 1 2 3 4 5 ... 42 >

H3K27ac Z-scores

cell type	H3K27ac and DNase	H3K27ac only
adrenal gland male adult (54 years)	4.76	6.36
common myeloid progenitor CD34+ female	4.60	5.80
neutrophil	--	5.79
peripheral blood mononuclear cell female adult (28 years)	--	5.68
thyroid gland male adult (37 years)	4.13	5.61

< 1 2 3 4 5 ... 28 >

Figure 2.30 | Overview of SCREEN. **a,** Landing page of SCREEN where user can enter a gene, SNP, or locus of interest to investigate cREs. Alternatively, users can select the GWAS app to analyze cREs overlapping disease associated genetic variants. **b,**

Example results page from searching in the main query box. SCREEN lists cREs with accession, location, nearby genes, and overview of activity. **c,** cRE details page displaying additional information about each cRE such as its activity across cell types and overlapping TF peaks

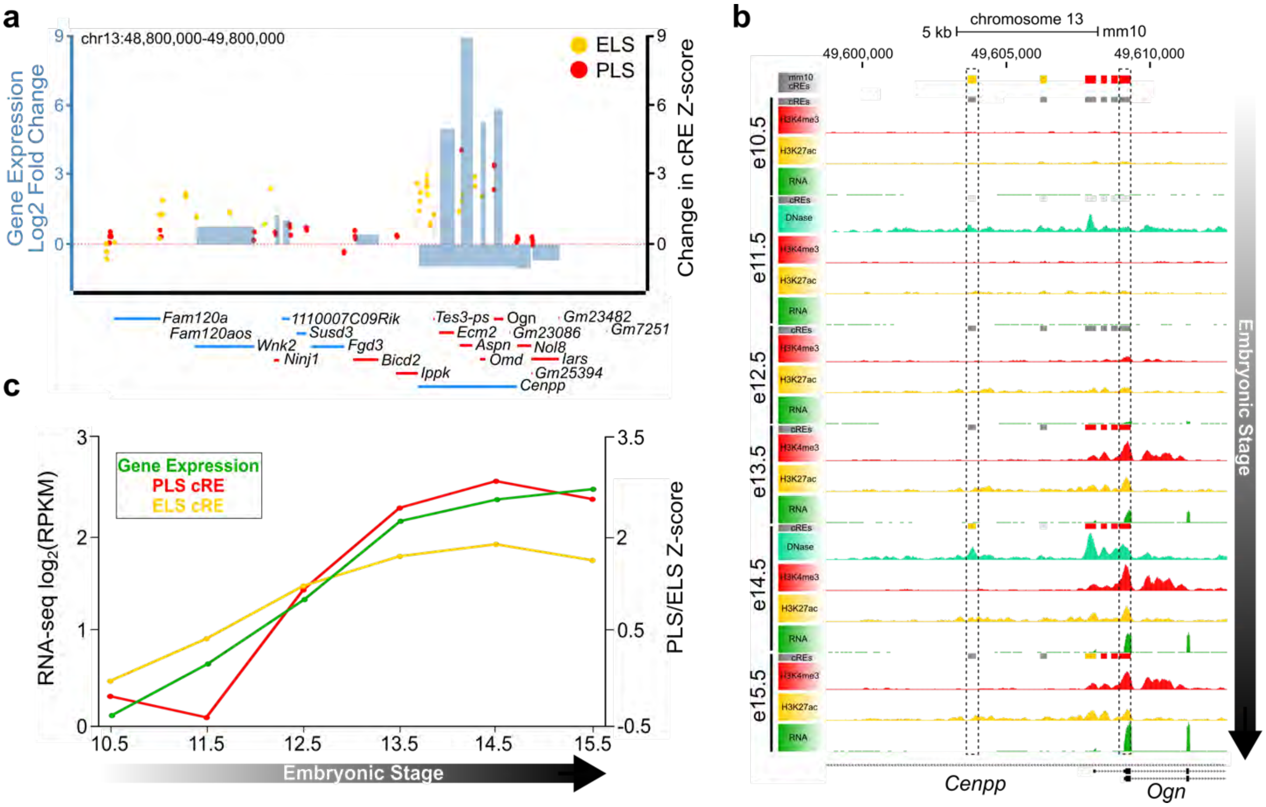


Figure 2.31 | Analyzing differential gene expression and cRE activity across developmental time-points. **a,** Comparison between limb e11.5 and e15.5 gene expression and cRE activity. Blue bars indicate differentially expressed genes, and red and yellow dots indicate cREs with promoter-like and enhancer-like signatures. The heights of bars or dots indicate changes (Log2 FC or difference in Z-score) between time-points. **b,** Genome browser view of the *Ogn* locus with H3K27ac, H3K4me4, DNase, and RNA-seq signals for the limb across all surveyed time-points. Promoter-like cREs are designated by red bars and enhancer-like cREs are designated by orange bars. **c,** *Ogn* gene expression and nearby cRE activity increase coordinately across time-points. The increase in gene expression lags behind the increases in cRE-PLS and cRE-ELS activities.

Table 2.1 | ENCODE mouse experiments used for enhancer prediction

	Midbrain	Hindbrain	Neural Tube	Limb
DNase	ENCSR292QBA	ENCSR358ESL	ENCSR312QVY	ENCSR661HDP
H3K27ac	ENCSR088UKA	ENCSR129LAP	ENCSR531RZS	ENCSR897WBY
H3K4me3	ENCSR283RFW	ENCSR928CYU	ENCSR427OZM	ENCSR654VMK
H3K4me1	ENCSR450ITF	ENCSR695FPP	ENCSR448TTC	ENCSR548BCO
H3K9ac	ENCSR502WUI	ENCSR734IEL	ENCSR547PLI	ENCSR286IGS
H3K36me3	ENCSR535NVF	ENCSR175QZX	ENCSR445UYH	ENCSR871YCT
H3K27me3	ENCSR545BRW	ENCSR375GSG	ENCSR240OUM	ENCSR085EYQ
WGBS	ENCSR091VFX	ENCSR398UCM	ENCSR613BMI	ENCSR916GKL

Table 2.2 | Area under PR curves for VISTA Enhancer prediction

Peak Space	Signal	Limb	Midbrain	Hindbrain	Neural Tube	Average
DNase	DNase	0.4108	0.3725	0.3862	0.2958	**0.3562**
H3K27ac	H3K27ac	0.3375	0.4320	0.3311	0.2712	0.3310
H3K4me3	H3K4me3	0.1228	0.2397	0.1994	0.1334	0.1749
H3K4me1	H3K4me1	0.2589	0.3227	0.2124	0.1601	0.2280

Table 2.3 | AUPR for VISTA Enhancer prediction anchored on DHSs

Peak Space	Signal	Hindbrain	Limb	Midbrain	Neural Tube	Average
DNase Peak	DNase Signal	0.3788	0.4159	0.3797	0.2951	0.3673
DNase Peak	H3K27ac Signal	0.3113	0.3265	0.3959	0.2526	0.3216
DNase Peak	Average Rank DNase-H3K27ac Signal	0.3764	0.3948	0.4148	0.3050	**0.3727**
DNase Peak	H3K4me3 Signal	0.2276	0.1828	0.2602	0.1615	0.2080
DNase Peak	Average Rank DNase-H3K4me3 Signal	0.2584	0.2392	0.2933	0.1751	0.2415
DNase Peak	H3K4me1 Signal	0.2442	0.2799	0.3122	0.1762	0.2531
DNase Peak	Average Rank DNase-H3K4me1 Signal	0.2527	0.2647	0.2901	0.1740	0.2454
DNase Peak	H3K9ac	0.2367	0.1977	0.2756	0.1721	0.2205
DNase Peaks	Average Rank DNase-H3K9ac Signal	0.2831	0.2574	0.3250	0.2147	0.2700
DNase Peak	H3K36me3 Signal	0.1910	0.1776	0.1911	0.1265	0.1715
DNase Peak	Average Rank DNase-H3K36me3 Signal	0.2280	0.2262	0.2212	0.1548	0.2075
DNase Peak	WGBS methylation	0.2470	0.2151	0.2663	0.1550	0.2208
DNase Peak	Average Rank DNase-WGBS Signal	0.3127	0.3031	0.3278	0.1981	0.2854
DNase Peak	H3K27me3 Signal	0.2187	0.1964	0.1853	0.1285	0.1822
DNase Peak	Average Rank DNase-H3K27me3 Signal	0.2700	0.2750	0.2325	0.1664	0.2360

Table 2.4 | Tested enhancer-like regions for midbrain e11.5

	VISTA ID	mm10 Coordinates	Result Summary: predicted tissue (Hb, Mb, Lb pos or neg)	Additional Tissue Activity Observed
Top Tier	mm1502	chr14:76253890-76257212	Mb positive (3/3)	
	mm1471	chr16:35584523-35589773	Mb positive (3/3)	Fb (3/3), Hb (3/3), neural tube (3/3)
	mm1461	chr2:154425426-154428462	Mb positive (3/3)	Fb (3/3), Hb (3/3), neural tube (3/3), nose (3/3), facial mesenchyme (3/3)
	mm1454	**chr2:25124845-25128090**	**Mb positive (3/4)**	**Fb (3/4), Hb (3/4), neural tube (4/4), eye (3/4)**
	mm1480	chr6:112808528-112813000	Mb positive (4/5)	Fb (5/5)
	mm1504	chr16:44528746-44534514	Mb positive (4/5)	Hb (4/5), Fb (4/5), neural tube (4/5)
	mm1503	chr14:40955282-40959325	Mb positive (4/8)	Hb (5/8), cranial nerve (5/8), trigeminal V (8/8), DRG (8/8), Lb (6/8)
	mm1469	chr12:80059965-80065110	Mb positive (5/5)	Fb (4/5), Hb (5/5), neural tube (5/5), Ht (5/5), branchial arch (3/5)
	mm1460	chr1:133185980-133189493	Mb positive (5/5)	Fb (5/5), Hb (5/5), neural tube (4/5), eye (5/5)
	mm1458	chr8:93916782-93919741	Mb positive (5/7)	Fb (5/7), Hb (5/7), neural tube (5/7)
	mm1456	chr12:109946504-109950294	Mb positive (6/6)	Fb (6/6), Hb (6/6), neural tube (5/6)
	mm1462	chr11:113783367-113787793	Mb positive (8/12)	Fb (7/12), Hb (9/12), neural tube (9/12), trigeminal V (7/12), DRG (7/12), Lb (7/12)
	mm1479	chr17:29354850-29357878	neg	
	mm1472	chr11:20654380-20657594	neg	
	mm1459	chr3:65973873-65976364	neg	
	mm1482	chr17:37028041-37029979	neg	
	mm1470	chr6:90780495-90784241	other positive	Hb (6/8), neural tube (5/8), trigeminal V (6/8), DRG (7/8)
	mm1481	chr11:120314873-120317426	other positive	DRG (3/4), other (3/4)
	mm1457	chr18:38378941-38381638	other positive	Fb (6/6), Lb (6/6), cranial nerve (4/6), DRG (4/6)

	mm1463	chr5:21731579-21734432	other positive	Fb (4/4)
Middle Tier	mm1553	chr9:86186309-86188376	Mb positive (3/5)	Hb (3/5) , Fb (3/5), neural tube (3/5)
	mm1557	chr6:65309702-65311689	Mb positive (3/6)	Fb (4/6), neural tube (4/6)
	mm1552	chr19:41715402-41718238	Mb positive (5/6)	Fb (6/6), Hb (6/6), DRG (4/6)
	mm1555	chr11:26772020-26774013	Mb positive (5/6)	Hb (3/6)
	mm1558	chr7:117685707-117687616	Mb positive (6/12)	Hb (8/12), Lb (9/12), Fb (9/12), neural tube (6/12)
	mm1546	chr1:9648223-9650965	Mb positive (6/8)	Hb (7/8), trigeminal V (6/8), neural tube (5/8), DRG (4/8), cranial nerve (5/8)
	mm1544	chr16:94723154-94725386	neg	
	mm1545	chr18:54759740-54761758	neg	
	mm1547	chr9:50451431-50453445	neg	
	mm1550	chr5:140767048-140769052	neg	
	mm1551	chr6:144165683-144167703	neg	
	mm1554	chr9:121359503-121361504	neg	
	mm1549	chr3:127705248-127707248	other positive	Fb (7/7), neural tube (6/7)
	mm1556	chr13:76809879-76811879	other positive	Fb (3/3), Hb (3/3)
	mm1548	chr6:145855046-145857046	other positive	Fb (4/4)
Bottom Tier	mm1524	chr6:52365164-52368224	Mb positive (4/4)	
	mm1583	chr11:85857229-85860254	Mb positive (4/7)	
	mm1526	chr11:113307639-113310548	Mb positive (4/7)	Hb (4/7), neural tube (4/7)
	mm1522	chr5:114298723-114301210	Mb positive (6/7)	
	mm1520	chr8:124313443-124316007	neg	
	mm1580	chr13:113913477-113916032	neg	

	mm1521	chr12:111161954-111163777	neg	
	mm1582	chr9:107644011-107645376	neg	
	mm1523	chr14:100374296-100376562	neg	
	mm1585	chr6:83434774-83436371	neg	
	mm1587	chr15:84259559-84261268	neg	
	mm1581	chr18:55784854-55787217	other positive	Hb (4/4)
	mm1584	chr15:74155467-74158454	other positive	Lb (4/4), other (4/4), eye (4/4), neural tube (4/4), branchial arch (4/4)
	mm1525	chr17:28105717-28108643	other positive	Fb (3/4), Hb (3/4)
	mm1586	chr7:48749001-48750980	other positive	trigeminal V (9/12), Hb (6/12)

Table 2.5 | Tested enhancer-like regions for hindbrain e11.5

	VISTA ID	mm10 Coordinates	Result Summary: predicted tissue (Hb, Mb, Lb pos or neg)	Additional Tissue Activity Observed
Top Tier	mm1444	chr12:86822930-86827112	Hb positive (7/8)	Mb (3/8)
	mm1496	chr11:94158655-94162177	Hb positive (3/4)	Fb (3/4)
	mm1494	chr4:136754322-136758801	other positive	Mb (3/3), nose (3/3)
	mm1445	chr1:134060554-134066172	Hb positive (4/4)	Fb (4/4), Mb (4/4), DRG (4/4), cranial nerve (4/4), neural tube (4/4), other (4/4), trigeminal V (4/4)
	mm1446	chr14:25107155-25110736	Hb positive (3/3)	Fb (3/3), Mb (3/3), neural tube (3/3), eye (3/3)
	mm1488	chr17:10333563-10337174	Hb positive (4/4)	Mb (4/4) neural tube (4/4)
	mm1447	chr2:21008806-21012242	Hb positive (6/6)	Mb (5/6)
	mm1448	chr7:111121648-111123424	neg	
	mm1449	chr15:98967479-98972541	Hb positive (5/7)	Fb (6/7), Mb (6/7), neural tube (4/7)
	mm1450	chr19:45057481-45059842	neg	
	mm1497	chr7:145204877-145207696	Hb positive (5/6)	Mb (6/6), Fb (6/6), neural tube (5/6), Ht (4/6), Lb (4/6)
	mm1498	chr14:19984462-19988065	Hb positive (3/4)	Mb (4/4)
	mm1499	chr17:66478396-66482145	Hb positive (6/6)	Fb (4/6), Mb (6/6), neural tube (6/6)
	mm1451	chr5:112236129-112240520	Hb positive (10/10)	Fb (9/10), Mb (10/10), neural tube (10/10), cranial nerve (5/10)
	mm1489	**chr7:140056813-140059080**	**Hb positive (5/5)**	**Mb (5/5), neural tube (5/5)**
	mm1452	chr3:51787852-51791869	Hb positive (4/6)	Mb (3/6), neural tube (5/6)
	mm1453	chr5:125140066-125142757	neg	
	mm1500	chr14:66492741-66497119	Hb positive (3/4)	Fb (3/4), Mb (3/4), neural tube (3/4)
	mm1501	chr13:84344551-84350110	neg	
	mm1478	chr15:91016522-91020048	Hb positive (3/4)	Fb (3/4), Mb (3/4), neural tube (3/4)

Middle Tier	mm1540	chr8:108243823-108246831	Hb positive (3/5)	
	mm1534	chr7:93060869-93063626	Hb positive (4/5)	Fb (5/5), neural tube (5/5), Lb (4/5)
	mm1532	chr4:148859471-148860956	Hb positive (5/5)	cranial nerve (3/5)
	mm1542	chr14:63585329-63587537	Hb positive (8/9)	neural tube (9/9), Mb (8/9), Fb (7/9), cranial nerve (6/9)
	mm1535	chr18:54610385-54613247	neg	
	mm1536	chr12:86798248-86799852	neg	
	mm1560	chr4:154593394-154596395	neg	
	mm1539	chr7:70329472-70332465	neg	
	mm1543	chr15:77335353-77337280	neg	
	mm1537	chr3:104540534-104542671	other positive	neural tube (5/6)
	mm1538	chr17:62851927-62854939	other positive	Mb (5/6)
	mm1603	chr2:93277888-93280895	other positive	Lb (3/4)
	mm1561	chr2:163188663-163191536	other positive	Fb (3/6), eye (3/6), Lb (4/6)
	mm1541	chr3:5386841-5389926	other positive	Fb (6/7), neural tube (6/7)
	mm1562	chr19:21164417-21167450	other positive	tail (5/6)
	mm1533	chr5:125214963-125216753	other positive	Fb (3/4)
Bottom Tier	mm1515	chr2:166215189-166217391	Hb positive (4/5)	neural tube (6/5)
	mm1604	chr12:105972686-105975121	Hb positive (7/9)	other/abdomen (6/9)
	mm1577	chr3:5348379-5351461	Hb positive (8/10)	branchial arch (4/10)
	mm1510	chr16:33593467-33596065	neg	
	mm1511	chr2:70559085-70560255	neg	
	mm1512	chr12:16854925-16857237	neg	
	mm1578	chr16:72690693-72694735	neg	

	mm1513	chr6:95173859-95175593	neg	
	mm1579	chr18:15173064-15176000	neg	
	mm1517	chr1:182981864-182984314	neg	
	mm1519	chr7:46452307-46455505	neg	
	mm1509	chr9:63058010-63060489	other positive	Mb (4/4), Fb (4/4)
	mm1516	chr6:89241345-89243919	other positive	Mb (5/11), facial mesenchyme (8/11)
	mm1514	chr7:82713868-82717092	other positive	Lb (6/6)
	mm1518	chr9:63958315-63961551	other positive	Lb (6/8)

Table 2.6 | Tested enhancer-like regions for limb e11.5

	VISTA ID	mm10 Coordinates	Result Summary: predicted tissue (Hb, Mb, Lb pos or neg)	Additional Tissue Activity Observed
Top Tier	mm1473	chr9:72639094-72641466	Lb positive (11/11)	eye (5/11)
	mm1505	chr3:101394801-101399299	Lb positive (12/12)	somite (10/12), branchial arch (10/12), facial mesenchyme (9/12)
	mm1464	chr8:126825533-126828569	Lb positive (3/3)	
	mm1476	chr2:128712175-128715819	Lb positive (3/4)	facial mesenchyme (3/4), branchial arch (3/4), DRG (3/4)
	mm1486	chr9:41227902-41230545	Lb positive (3/5)	
	mm1474	chr6:72710179-72712952	Lb positive (4/4)	branchial arch (3/4)
	mm1485	**chr14:24261692-24264509**	**Lb positive (4/4)**	
	mm1493	chr9:41949140-41954525	Lb positive (4/7)	eye (5/7)
	mm1475	chr9:106354104-106358319	Lb positive (4/7)	facial mesenchyme (6/7)
	mm1492	chr4:154707415-154711162	Lb positive (5/5)	facial mesenchyme (5/5)
	mm1506	chr4:139597441-139601336	Lb positive (5/8)	blood vessels (5/8)
	mm1484	chr7:123096694-123099867	Lb positive (6/6)	Ht (6/6), eye (6/6), facial mesenchyme (5/6)
	mm1490	chr4:108361018-108364528	Lb positive (7/7)	Mb (7/7), neural tube (7/7), facial mesenchyme (7/7), Hb (7/7), nose (7/7), branchial arch (6/7), somite (5/7), genital tubercle (7/7)
	mm1483	chr8:90860971-90863490	Lb positive (8/8)	Fb (6/8), Hb (6/8), Mb (6/8), cranial nerve (6/8), trigeminal V (7/8), DRG (7/8)
	mm1507	chr8:11584424-11587803	neg	
	mm1491	chr9:41934743-41936889	neg	
	mm1477	chr13:51312386-51316734	neg	
	mm1508	chr10:91179718-91183751	neg	
	mm1495	chr14:65342605-65345502	neg	

	mm1487	chr10:59919229-59922102	neg	
Middle Tier	mm1574	chr7:136389300-136391216	Lb positive (3/4)	
	mm1567	chr8:87659116-87661534	Lb positive (4/5)	tail (4/5)
	mm1564	chr2:60785660-60787563	Lb positive (5/5)	branchial arch (4/5)
	mm1576	chr5:16778223-16779671	Lb positive (5/5)	
	mm1571	chr15:95638744-95641449	Lb positive (5/6)	Hb (6/6), neural tube (6/6)
	mm1570	chr14:49121140-49122633	Lb positive (7/8)	
	mm1563	chr9:32823262-32824682	neg	
	mm1559	chr6:5801484-5804011	neg	
	mm1568	chr11:104288616-104290675	neg	
	mm1572	chr3:81782584-81784020	neg	
	mm1573	chr10:38972079-38974105	neg	
	mm1565	chr13:51919913-51922132	other positive	Fb (3/3), Mb (3/3), Hb (3/3), neural tube (3/3)
	mm1566	chr4:148984734-148987251	other positive	nose (4/4)
	mm1575	chr1:59333923-59335433	other positive	tail (4/6), somites (4/6)
	mm1569	chr5:65414975-65416572	other positive	Unidentifiable structures in chest and abdomen (9/10)
Bottom Tier	mm1597	chr6:89613039-89615855	Lb positive (10/11)	branchial arch (8/11), somites (7/11), Mb (6/11), facial mesenchyme (5/11)
	mm1598	chr8:26832486-26834858	Lb positive (3/4)	Mb (3/4), Fb (3/4), Hb (3/4), somites (3/4), nose (3/4), other (3/4)
	mm1599	chr3:37934422-37937092	Lb positive (8/9)	Hb (7/9), Fb (7/9), eye (5/9)
	mm1588	chr11:60365089-60367681	neg	
	mm1589	chr12:107951151-107953994	neg	
	mm1590	chr7:135888196-135890453	neg	

	mm1591	chr8:108018634-108020810	neg	
	mm1592	chr12:108913631-108917155	neg	
	mm1593	chr11:118083180-118085181	neg	
	mm1594	chr6:85022961-85025939	neg	
	mm1595	chr12:73510499-73513049	neg	
	mm1596	chr7:112867931-112870363	neg	
	mm1602	chr7:132203278-132205205	neg	
	mm1600	chr14:21673440-21675700	other positive	tail (3/3)
	mm1601	chr1:61202202-61204582	other positive	facial mesenchyme (3/5)

Table 2.7 | Correlation of ranked peaks with ranked gene expression in mouse tissues

Peak Space	Signal	Hindbrain	Limb	Midbrain	Neural Tube	Average
DNase	DNase	0.3454	0.3643	0.3973	0.4714	0.3946
DNase	H3K4me3	0.7332	0.7472	0.7507	0.7488	**0.7450**
H3K4me3	DNase	0.2364	0.2603	0.2239	0.1055	0.2065
H3K4me3	H3K4me3	0.5551	0.6112	0.5691	0.5555	0.5727

Table 2.8 | Correlation of ranked peaks with ranked gene expression in human cells

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

Table 2.9 | Combined ChromHMM States

Combined State	State 1	State 2	State 3	State 4
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 2.10 | Overlap of previous enhancer predictions with cREs

Assay	Cell Type	PLS	ELS	CTCF	DNase	Inactive	No Overlap
FANTOM5	GM12878	274	646	5	64	342	1,135
FANTOM5	hepatocyte	36	118	2	24	127	266
FANTOM5	keratinocyte	139	380	1	2	111	546
FANTOM5	astrocyte	256	427	0	30	182	786
GRO-seq	GM12878	27,771	15,312	612	1,601	34,030	34,703
GRO-seq	K562	26,815	20,206	1,193	4,230	32,206	33,233
GRO-seq	IMR-90	15,993	11,020	293	1,852	25,177	26,343
GRO-seq	MCF-7	21,283	10,124	320	1,493	34,150	28,876
GRO-seq	HeLa	21,738	16,463	353	1,127	17,143	26,807
STARR-seq	HeLa	687	2,800	67	1,035	10,447	7,809

Table 2.11 | GWAS studies included in analysis

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

METHODS

Mouse Transgenic Assays

To test for enhancer activity, each region was cloned into a construct containing the Hsp68 promoter, which lacks activity in embryonic mice, and the LacZ reporter gene. Individually, these constructs were injected into fertilized mouse eggs which were then implanted into pseudopregnant female mice. At e11.5, the embryos were harvested and stained for LacZ reporter gene activity. If the tested genomic region has tissue specific enhancer activity, the tissue will stain blue. For a tissue to have enhancer activity, at least three embryos were required to test positive.

Cell type specific enhancer prediction

In November 2015, we downloaded regions from the VISTA database. We lifted these regions from the mm9 to the mm10 genome and merged overlapping regions, generating 1,994 unique regions. To test anchoring schemes on different datatypes, we selected the top 20,000 peaks ranked by p-value and shrank the size of the peak to 300 bp. For histone modifications, we used peaks call by MACS2 which were in both biological replicates. To modify the width of these peaks, we used +/- 150 bp around the peak summit. For DHS, we used peaks call by HOTSPOT2 for the first biological replicate. To modify the width of these peaks, we used +/- 150 bp around the peak center. To rank peaks, we used signal from each experiment. For histone modifications, we used "fold-change over control" signal from the combined biological replicates and calculated signal over a +/- 1 kb window around the peak summit. For DHSs, we used raw signal from replicate one and calculated signal over the 300 bp DHS.

To evaluate performance, we intersected peaks with all tested VISTA regions. If more than one peak overlapped a region we selected the peak with the highest signal. VISTA regions were then ranked by the signal of their overlapping peak. VISTA regions that did not overlap peaks were assigned a signal of 0. We plotted the PR curves using the R package ROCR and calculated the area under the curves using custom R scripts.

To compare ranking schemes, we ran the same pipeline except we anchored all predictions on the 300 bp DHSs. For histone modification and methylation (WGBS) signals we used a +/- 1 kb window centered at each DHS. For H3K27me3 and methylation signals we reversed the order of ranking when making our PR curves as high H3K27me3 and methylation correlate with repressed regions.

Cell type specific promoter prediction

To evaluate the performance of promoter prediction methods, we downloaded transcript expression quantifications from the ENCODE DCC produced from the ENCODE RNA-seq uniform processing pipeline. Using TSS-proximal (± 2 kb) DHSs or H3K4me3 peaks, we computed the Pearson correlation between the ranks of these peaks and the ranks of the expression (TPM) of transcripts within 2 kb. We tested 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. Overall, the method with the high correlation was anchoring predictions on DHSs and ranking by H3K4me3 signal.

Creating representative DHSs (rDHSs)

As of February 1, 2017 there were 449 hg19 DNase experiments and 62 mm10 DNase experiments on the ENCODE data portal with HOTSPOT2 calls. As a preprocessing step, we normalized the signal at these DHSs so that we can compare relative signals across datasets. For each experiment, 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 for all DHSs passing an FDR threshold of <0.1%. Using a script adapted from the Stamatoyannopoulos lab, we clustered these high quality DHSs 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 iteratively repeated this process until we obtained a list of non-overlapping rDHSs representing all DHSs.

Normalizing epigenomic signals

To normalize DNase, H3K4me3, H3K27ac, and CTCF signal for each rDHS, we transformed average signals into Z-scores.For each experiment, we used the UCSC tool *bigWigAverageOverBed* to compute the average signal across each cRE. For H3K4me3 and H3K27ac we added +/- 500 to both ends of the rDHS when computing the signal. Using a custom Python script, we calculated the log of these signals and computed a Z-score for each rDHS compared with all other rDHS signals within the cell type. rDHSs with a raw signal of 0 were assigned a Z-score of -10 due to inconsistencies with pseudocounts.

Classification of cREs

To classify rDHSs as cREs we used the classifications trees in Figures 2.12 and 2.13. Based on maximum DNase, H3K4me3, H3K27ac, and CTCF Z-scores across all cell types as well as distance from GENCODEV19 annotated TSS, rDHSs can be classified as PLS, ELS, or CTCF-only cREs. Because both promoters and enhancers can have high levels of H3K4me3 and H3K27ac, the classification tree first splits based on whether a rDHS is proximal (+/- 2 kb) to a TSS. rDHSs that are not classified as cREs are discarded. To classify cRE activity in a particular cell type we used the classification scheme in Figure 2.18 relying on the Z-scores of genomic signals in the cell type of interest.

Saturation of cREs within each 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 genomic 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 unique 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[25], we calculated the cREs in each group to be at 95% saturation for each curve using a Weibull distribution.

Overlap of cREs with ChromHMM states

To compared PLS and ELS cREs with chromatin states called by chromHMM, we first combined similar chromHMM states to generate seven broad states (Table 2.9). For human, we analyzed chromHMM regions for GM12878 cells from the ENCODE 2012 paper (ENCODE experiment accession ENCFF001TDH). We selected all PLS or ELS cREs 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. Each cRE was assigned to only one chromHMM state—the state that overlapped the largest number of basepairs. 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 modifications in the same tissue–time-point.

Clustering cell types on the basis of their cRE activities

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 and H3K4me3. 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.

Enrichment of GWAS variants in cREs

We curated studies from the NHGRI-EBI Catalog (Table 2.11) 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 MAF quartile and the same distance to TSS quartile. 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[101] 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.

SCREEN

SCREEN was engineered by Michael Purcaro and Henry Pratt of Zhiping Weng's lab. Their code is available at https://github.com/weng-lab/SCREEN.

Scripts

Scripts for this analysis can be found on GitHub: https://github.com/Jill-Moore/Dissertation/tree/master/Chapter-II/

CHAPTER III: Systematic evaluation of enhancer target gene prediction methods

PREFACE

Results from this chapter were adapted from

Moore*, Purcaro*, Pratt*, Epstein*, Shoresh*, Adrian*, Kawli*, Davis*, Dobin*, Kaul*, Halow*, Van Nostrand*, Freese*, Gorkin*, He*, Mackiewicz*, The ENCODE Consortium. Cherry, Myers, Bing Ren, Graveley, Stamatoyannopoulos, Gerstein, Pennacchio, Gingeras, Snyder, Bernstein, Wold, Hardison, and Weng. "ENCODE Phase III: Building an Encyclopaedia of candidate Regulatory Elements for Human and Mouse,"

which is currently under review at *Nature* and from

Moore, Garnick, and Weng. "A Systematic Evaluation of Enhancer Target Gene Prediction Methods using the ENCODE Encyclopedia."

which is currently in preparation. I performed all analysis and generated all the figures that were used in the chapter.

ABSTRACT

To interpret the biological function of an enhancer, we need to determine the genes it regulates. While many enhancers target nearby genes, there are examples of enhancers regulating genes up to 1 Mb away. Many groups have developed computational methods for linking enhancers with target genes, yet these methods are trained and tested on different enhancer-gene links, making comparisons between methods difficult. To systematically and accurately evaluate enhancer-gene linking methods, we developed a benchmark of chromatin and

genetic interaction datasets. In this benchmark, we used Hi-C, ChIA-PET and eQTL links to generate positive and negative ELS-gene pairs that can be used for training, validating and testing computational methods. Using this benchmark, we evaluated correlation based methods and found that they had low overall performance and did not outperform ranking genes by distance. We then developed a Random Forest model which outperformed unsupervised methods and was applicable across cell types. We used our Random Forest model to predict genes linked with a variant associated with multiple sclerosis identifying a novel GWAS risk gene. Our results establish a pipeline for generating a benchmark of ELS-gene pairs, which can be used to evaluate published target gene methods.

INTRODUCTION

Chapter II of this thesis detailed the Registry of candidate Regulatory elements (cREs), a collection of putative regulatory regions in human and mouse. The majority, 75%, of these cREs have enhancer-like signatures (i.e. high DNase and H3K27ac signals) and are distal from TSSs. To interpret the biological function of these cREs, we need to determine the genes they regulate. While many enhancers target nearby genes, there are examples of enhancers regulating genes up to 1 Mb away[2] so simply assigning an enhancer to its nearest gene may not be an ideal method. Therefore, labs have developed experimental and computational methods for investigating enhancer-gene interactions.

Experimental assays such as Hi-C and ChIA-PET survey physical interactions between genomic regions[62,66]. By overlapping the anchors of these interactions with annotated enhancers and promoters, we can infer regulatory connections. However, these assays are expensive to perform and have only been conducted with high resolution in a small number of cell types. Therefore, we need to rely on computational methods to more broadly predict enhancer-gene interactions.

Previous work by members of the ENCODE consortium demonstrated that they could identify the target genes of enhancers by correlating enhancer activity with transcriptional activity. Correlation based methods rely on the hypothesis that enhancers are active in the same cell types in which their target gene is expressed. These labs used DNase signal[6] or H3K4me1[70] signal to estimate enhancer activity and DNase signal[6,68], POLII[70] signal or gene expression[69] to estimate transcriptional activity. While these methods identified biologically relevant enhancer-gene links, they have yet to be systematically analyzed to evaluate their overall precision and recall.

Other target gene prediction methods such as, IM-PET[72], PETmodule[73] and TargetFinder[74], use supervised machine learning algorithms to predict enhancer-gene links utilizing features such as epigenomic signal[72-74], gene ontology terms[73], and conservation[72]. While these methods have overall high performance they require known enhancer-gene pairs for training. These methods use Hi-C and ChIA-PET data as their gold standards but since each method uses data from different studies, as well as different collections of enhancers, it is difficult to compare the performance of each model.

In order to determine the best method for linking ELS cREs with potential target genes, we developed a collection of benchmark datasets using the Registry of cREs and experimentally derived enhancer-gene interactions. We then tested common methods of linking enhancers with genes such as using distance and signal correlation. Ultimately, we developed a high performing Random Forest approach that can be applied across cell types to predict target genes. Using this Random Forest model, we predicted targets genes for a SNP associated with multiple sclerosis and identified a novel GWAS risk gene. Our analysis lays the groundwork for future comparisons of gene-enhancer linking methods and a push towards standardizing the comparison of computational models.

RESULTS

Curating Benchmark Datasets

In order to compare methods of predicting target genes, we curated a collection of potential enhancer-gene interactions. We focused on two types of data: three-dimensional chromatin interactions (e.g. ChIA-PET and Hi-C data) and genetic interactions (e.g. eQTLs). We chose to use both types of data because they complement each other's limitations. For example, ChIA-PET reports proximal physical interactions between two genomic regions, but does not imply regulation of one region by another. eQTLs, on the other hand, suggest a functional relationship between a SNP and gene, but do not imply direct regulation by the SNP. Therefore, by testing methods on both types of data we can investigate what features are more indicative of physical interactions versus genetic interactions.

We decided to create our initial benchmark using interaction data surveyed in GM12878 and lymphoblastoid cell lines due to the large amount of genomic data that has been generated by the ENCODE project and the biological community. For chromatin interaction data, we selected ChIA-PET and Hi-C datasets from the ENCODE DCC and the gene expression omnibus (GEO). During the second phase of the ENCODE project, the Snyder lab generated ChIA-PET data in GM12878 targeting RAD21, a component of the cohesion complex[129]. We supplemented this data with ChIA-PET datasets generated by the Ruan lab targeting POLII and CTCF[67]. For Hi-C, we included links from promoter capture Hi-C (CHi-C) data generated by the Osborne lab[65] and high resolution Hi-C loops generated by the Aiden lab[63]. For genetic interaction data, we included eQTLs reported by the Dermitzakis lab[114] and the GTEx consortium[108] in lymphoblastoid cell lines.

To generate candidate ELS-gene pairs from these datasets, we required one end of a link to overlap an ELS cRE and the other end to fall within 2 kb of a GENCODE annotated TSS (Figure 3.1). To correctly identify the target gene, we excluded ambiguous links that connected to multiple gene TSSs. For eQTLs, we linked a cRE to a gene if it directly overlapped the eQTL SNP. To create our negative set, for cREs with positive links, we selected all genes with TSSs within a window that were not a part of a positive or ambiguous pair. For each dataset, we determined the size of this window by calculating the 95th percentile of distance between positive ELS-gene pairs (Figure 3.2). This window ranged from +/- 170 kb for POLII ChIA-PET to +/- 983 kb for Aiden Hi-C links (Table 3.1). Using this method, we generated thousands of GM12878 specific ELS-gene pairs. For each dataset, we split pairs into three groups with 50% of pairs forming the training set, 25% forming the validation set, and the remaining 25% forming the test set. Therefore, this benchmark can be used to train, validate, and test any target gene prediction method.

Comparing Benchmark Datasets

To determine the similarity of the benchmark datasets, we calculated the overlap coefficient for the number of positive ELS-gene pairs between each dataset. When we clustered the datasets, we observed two large groups, one for genetic interaction data and the second for chromatin interaction data (Figure 3.3). Within the second cluster, we observed two sub-clusters with ChIA-pet and Hi-C datasets aggregating together. We also observed that POLII ChIA-PET and Osborne CHi-C data had a higher overlap with eQTL datasets compared to Aiden Hi-C, CTCF ChIA-PET or RAD21 ChIA-PET (max overlap coefficient = 0.14 vs. max overlap coefficient = 0.03) though the overall overlap was still very low.

For each benchmark dataset, we also analyzed the activity and expression levels of the ELS-gene pairs. POLII ChIA-PET cREs tended to have higher levels of H3K27ac and H3K4me3 signals (p<4.2E-8, p<8.1E-8) compared to the other benchmarks, while RAD21 ChIA-PET cREs had lower levels (p<3.6E-9, p<2.9E-3) (Figure 3.4a,b). We did not observe notable differences for these signals between the other datasets. For CTCF, however, there were differences in the distributions of signal for ELS cREs. For RAD21 ChIA-PET, CTCF ChIA-PET and Aiden Hi-C pairs, there were populations of ELS cREs with very high levels of CTCF signal (Figure 3.4c). For RAD21 ChIA-PET this was a large group of about 79% of total cREs; for CTCF ChIA-PET and Aiden Hi-C, these groups were smaller, containing 35% and 20% of cREs respectively. This enrichment for CTCF signal is biologically consistent with these chromatin interaction experiments. CTCF was the target for the CTCF ChIA-PET experiment, RAD21 is known to co-localize with CTCF and the Aiden lab reported that their links are anchored at CTCF binding sites. When we analyzed gene expression, we found that genes in the POLII ChIA-PET pairs had higher expression levels than other datasets (median 15.3 TPM) while genes in the Osborne CHi-C pairs had low expression levels (median of 0.2 TPM) (Figure 3.4d). This suggests that some of these CHi-C pairs may be false positives or are links for ELS cREs that have yet to regulate gene expression.

Overall, these results suggest that these benchmark datasets capture different types of genomic interactions and that to accurately determine the performance of a target gene method, we should test it using all of these benchmark datasets.

Ranking by Distance Outperforms Correlation Based Methods

We began by evaluating the simplest method of enhancer target gene prediction: selecting the closest gene by linear distance. We tested this method using TSSs for all annotated genes and only TSSs from protein coding genes by calculating the precision and recall for each benchmark dataset. For all datasets except RAD21 ChIA-PET, we observed higher performance using TSSs for protein coding genes, rather than all genes. POLII ChIA-PET links had the highest performance with a precision of 0.66 and recall of 0.46 for protein coding gene TSSs (Figure 3.5). eQTLs and other ChIA-PET links had moderate performance and the Hi-C datasets had the lowest performance. We then tested whether we could predict links simply by ranking genes by their linear distance from the ELS cRE. Since this method was independent of cell type and did not require any genomic data, we considered it our baseline method. For each benchmark dataset, we evaluated performance by calculating the areas under the receiver operating characteristic (ROC) and precision recall (PR) curves focusing primarily on the area under PR (AUPR) curves due to class imbalance in the benchmark datasets (Table 3.2). As expected POLII ChIA-pet pairs had the highest AUPR (0.41) and Aiden Hi-C pairs had the lowest (0.06) (Figure 3.6 and 3.7, Table 3.2b). We compared all subsequent methods to these baseline results.

Our next step was to evaluate correlation based approaches. Adapting methods from Thurman *et al*.[68], we calculated the correlation coefficient for average signal across the ELS cRE and the TSS across hundreds of cell types (462 for DNase and 136 for H3K27ac). We tested this method using different epigenomic signals (e.g. DNase vs H3K27ac), signal normalization techniques (e.g. Raw vs Z-score) and correlation methods (e.g. Pearson vs Spearman). For all benchmark datasets, the best performing method was calculating the Spearman correlation of Z-score normalized DNase signals (Table 3.2). However, for all of the benchmark datasets except RAD21 ChIA-PET, this method did not outperform our baseline model (Figure 3.6 and 3.7). For RAD21 ChIA-PET links, though we achieved a 57% improvement using correlation over the baseline method (AUPR 0.18 vs 0.11), the overall performance was still poor. To understand why the correlation methods had such low AUPR values, we analyzed specific ELS-gene pairs. We found that there were some pairs that did have high correlation coefficients. For example, the highest ranked POLII ChIA-PET ELS-gene pair was EH37E0572541-*WNT10A*, which had a Spearman correlation coefficient of 0.82 (Figure 3.8a,b). Both *WNT10A's* promoter and EH37E0572541 have high DNase signal in over a hundred of DNase experiments suggesting that EH37E0572541 regulates *WNT10A* in many types of cells. This example, however, is in the minority. There are many other cases where the DNase correlation for the ELS-gene pair is low. For example, ELS cRE EH37E0853090 is paired with *AKIRIN2* by both a POLII ChIA-PET link and a GTEx eQTL. However, EH37E0438944 and

AKIRIN2 have low DNase correlation (ρ=0.06) (Figure 3.8c,d). *AKIRIN2* is expressed across many cell types and has high DNase activity (PLS cRE DNase Z-score > 1.64) in all of the 462 surveyed cell types. EH37E0365491, on the other hand, only has high DNase signal (Z-score > 1.64) in six cell types, four of which are lymphoblastoid or lymphoma cell lines. Therefore, we predict that *AKIRIN2* is only regulated by EH37E0853090 in B cell related cell types and that because of this cell type specific regulation, we are unable to identify this ELS-gene pair using correlation. In general, we hypothesized that the low overall performance for correlation based methods is because enhancers are much more likely to be cell type specific than promoters (Chapter II, Figure 2.17). Therefore, while correlation methods are simple to implement because they do not require cell-type specific data, they cannot identify cell-type specific ELS-gene pairs, such as EH37E0853090-*AKIRIN2,* which results in worse performance compared to our baseline method.

When we compared the shape of the PR curves for correlation with our baseline method, we observed that correlation methods tended to have higher AUPR for the top ranked ELS-gene pairs whereas distance had better performance with the lower ranked pairs. We decided to combine the methods by taking the averaging rank of the two different schemes. For all benchmark datasets except eQTLs, this resulted in an average increased AUPR of 38% (Figures 3.6 & 3.7, Table 3.3). Our baseline method remained the best performing method for both GTEx and Dermitzakis eQTL pairs. In conclusion, while correlation based methods did not outperform selecting genes using distance, we observed an increase in performance when we combined the two features.

Random Forest Models Outperform Unsupervised Methods

Because simply combining distance and DNase correlation resulted in higher performance than our baseline model, we sought to develop a more sophisticated computational method that could combine multiple features to predict ELS-gene links. We decided to utilize the Random Forest algorithm[130], a supervised machine learning approach that is able to handle class imbalanced datasets (i.e. different number of positive and negatives), and used two approaches for developing models (Figure 3.9). First, we focused on developing models that can be applied across many cell and tissue types. These models would only require cell type specific DNase and/or H3K27ac data along with cell type independent features such as signal correlation, sequence features (k-mers), and distance (see methods for full feature list). Second, we focused on developing the best performing model in GM12878 using all available data. For this model, we included gene expression, TF and histone modification ChIP-seq, and RAMPAGE data. This model may only be able to be applied across a few cell types (e.g. GM12878, K562, H1-hESC) but may indicate which types of experiments we should prioritize in the future.

Starting with the DNase and H3K27ac only models, we found these relatively simple models had higher AUROC and AUPR compared to our average rank method for all benchmarks (Figure 3.10 and 3.11, Table 3.4). Including both signals as features resulted in the best performance, but models using only one of the signals (for cell types with just one datatype) still had comparable performance. We analyzed feature importance for each model and found that distance was consistently ranked the most important feature (Table 3.4). The next most important features were average DNase and H3K27ac signal around the gene's TSS suggesting that activity at the gene's promoter is indicative of ELS-gene links.

We then decided to expand on this basic model by adding in other features such as expression, TF ChIP-seq signal and other histone modifications (Table 3.3). We found that expression universally improved performance, with AUPR for the eQTL benchmarks having the highest increases (13% and 9%) (Figures 3.10 and 3.11). This is consistent with the methodology of defining eQTLs as the gene must be expressed in LCLs to be detected (Table 3.5). When we added CTCF signal to this expression model, considering both signal at the ELS cRE and TSS, AUPR increase 13% for CTCF ChIA-PET and 18% for RAD21 ChIA-PET. When we analyzed feature importance for this model, TSS CTCF signal was the top ranked feature for RAD21 ChIA-PET and the second ranked feature for CTCF ChIA-PET after distance (Table 3.6). For Aiden Hi-C, we only observed an increase of 2% for including CTCF and TSS CTCF was the fifth most important feature after distance and TSS DNase, H3K27ac, and conservation signals. For all three benchmarks, CTCF-signal at the enhancers was not even one of the top ten most important features. When we added POLII, EP300, RAMPAGE, or additional histone modification signals to our expression model, we did not observe dramatic increases in performance (Table 3.7). Our comprehensive model, which includes all of these features only had an average improvement of 0.5% over the expression + CTCF model.

Our Random Forest Model Can be Applied Across Cell Types

As we do not have chromatin or genetic interaction data for the majority of cell types covered by the Registry of cREs, if we were to apply our method to predict cell type specific ELS-gene pairs, we would need to train our model in a cell type with benchmark data. In order to evaluate the versatility of our models across different cell types, we tested our basic RF model using POLII and CTCF ChIA-PET data from HeLa cells generated by the Ruan Lab[67]. We compared performance of models trained and validated with data from the cell type versus models trained and validated in different cell types. We also compared performance to our best performing unsupervised method: taking the average rank of distance and DNase correlation. We found that while the cross-cell type models had lower AUPR than the same cell type models, they outperformed the average rank method with an average increase in AUPR of 25% (Figure 3.12, Table 3.10). The POLII ChIA-PET datasets retained higher performance across cell types compared to the CTCF ChIA-PET datasets. We hypothesize this is because distance is a more important feature in the POLII ChIA-PET models and is truly cell type independent. DNase and H3K27ac data quality can vary between cell types and though we normalized the signals, these biases can lower performance. Therefore, while we aim to improve the cross cell type application of our model, our random forest model still outperformed our best unsupervised approach.

Our Random Forest Model Identifies a New GWAS Gene for Multiple Sclerosis

Since our expression random forest model had high performance across all benchmark datasets without requiring multiple features, we used this model to predict ELS-gene pairs in GM12878 cells. Our previous GWAS enrichment analysis (Chapter II) demonstrated that variants associated with multiple sclerosis are enriched in cREs active in GM12878, which is in agreement with previously findings by the Stamatoyannopoulos lab[102]. Multiple sclerosis is an disease in which the body's immune system attacks the myelin sheaths of axons, resulting in in neurological deficits and deterioration[131]. The role for B cells in the pathology of multiple sclerosis has recently been recognized[103] with clinical trials targeting B cell antigens reporting success for slowing disease progression[104]. Therefore, identifying potential target genes in GM12878 may present new therapeutic targets.

One SNP of particular interest was rs1250568, which is in LD with two SNPs associated with multiple sclerosis[132,133]. Rs1250568 overlaps ELS cRE EH37E0182314 which has high H3K27ac and DNase signal in blood cells like GM12878. Rs1250568 also overlaps both a ChIP-seq peak and motif site for ELF1

(Figure 3.13a). ELF1 is primarily expressed in lymphoid cells and is involved in the IL-2 and IL-23 immune response pathways, both of which have previously been implicated in multiple sclerosis[134,135]. Additionally deltaSVM, a computational method that predicts the functional impact of variants using DNA sequence k-mers, predicted that rs1250568 is likely a casual SNP[136]. Because rs1250568 may disrupt ELF1 binding, thus affecting gene regulation, we aimed to identify genes that interact with EH37E0182314.

Using our expression Random Forest model trained on POLII ChIA-PET pairs, we predicted genes links using a probability cutoff of 0.5 (precision=0.80, recall=0.59). Our model reported two linked genes: *ZMIZ1* and *PPIF* (Figure 3.13b). These predictions are also supported by POLII ChIA-PET links that were not a part of our training set (Figure 3.13c). *ZMIZ1* is involved with androgen receptor signaling pathway and is expressed at lower levels in patients with multiple sclerosis[137]. Due to its proximity to the GWAS lead SNPs, *ZMIZ1* was reported as the risk gene in both GWAS[132,133]. The other linked gene, *PPIF* (Cyclophilin D), is located downstream of EH37E0182314 and encodes a mitochondrial permeability transition pore protein. While *PPIF* was not previously reported as a MS susceptibility gene by GWAS, evidence demonstrates that its dysregulation likely plays a role in the onset of multiple sclerosis. For example Forte *et al.* demonstrated that knocking out Ppif in mice with experimental autoimmune encephalomyelitis (EAE, a mouse disease model for multiple sclerosis) protected spinal cord axons, enabling the knock out mice to partially recover from EAE[138]. These findings are supported by recent work from Warne *et al*. who demonstrated that treating EAE mice with a Ppif inhibitor protects axons and improves motor ability[139]. Therefore, our comprehensive Random Forest model is able to predict biologically significant enhancer-gene links which can be used to better understand disease etiology and identify potential therapeutic targets.

DISCUSSION

In this chapter, we evaluated methods for linking enhancers with putative target genes. We curated a benchmark of ELS-gene pairs using chromatin and genetic interaction datasets and then used this benchmark to test common methods of target gene prediction. We found that overall, correlation of epigenomic signal across cell types is not an ideal method for predicting ELS-gene pairs. Though we observed that some ELS-gene pairs such as EH37E0572541-*WNT10A* have high correlation, the overall performance of these methods are low. We demonstrated that one reason for low performance is that with correlation methods we are unable to detect cell type specific regulation, which is the case for EH37E0853090-*AKIRIN2*. Another possible reason for poor performance is we survey signal across a biased set of cell and tissues types. For example, we have many more types of blood cells than lung or heart tissue samples. This imbalance could dramatically alter the results depending on the gene's expression patterns. In the future, we could curate a set of recommended cell types to use for correlation based methods (e.g., equal representation from different tissues of origin) but this still does address the problem of cell type specificity. Our results, however, do not discredit all correlation based methods. For example, correlation may be a much more powerful tool when analyzing activity and expression across differentiation or development. For example, in Chapter IV we link cREs with target genes using correlation of signal activity and gene expression within the same tissues across embryonic development. Our results simply suggest that using correlation across a wide range of different cell and tissue types is less than ideal.

We also demonstrated that even basic Random Forest Models, with few cell type specific features had a considerable improvement in performance over unsupervised methods. Adding other features such as expression and CTCF signal improved performance for specific benchmarks, but a comprehensive model including all features did not result in a much higher AUPR. Since simply adding more types of epigenomic signal did not drastically improve performance, moving forward we will integrate different types of features. For example, the Aiden lab reported enrichment for CTCF motifs at the anchors of their Hi-C loops[63] so we plan to include distance to nearest CTCF motif site in our model. While our Random Forest model had high performance when trained and validated on the same cell type, we were surprised of its lower performance when validated across cell types. With models dependent on cell type specific signals, we need to be aware of differences in data quality and biases in signal. For our analysis used Z-score normalized signals to try and alleviate this problem but we may need to investigate alternative methods of normalization. It is also possible fundamental differences between cell types may causing this lower performance. For example, GM12878 and HeLa cells are histologically very different from one another and may have different numbers of ubiquitous or cell type specific cREs. We plan on continuing to modify our models so that they are applicable across cell types perhaps implementing new normalization methods or using a ranking metric. Our next step will be to compare our Random Forest model with other target gene prediction methods such as PET-Module, IM-PET and Target Finder. These models have all been trained and evaluated using different enhancer-gene links so our benchmark will allow for an unbiased comparison between the methods.

Finally, our analysis also revealed key differences between the benchmark interaction datasets. In general, there is little overlap between these datasets and the ELS-genes pairs have different features. For example, CTCF ChIA-PET, RAD21-ChIA-PET and Aiden Hi-C ELS cREs have higher CTCF signal and their links are better predicted when CTCF is included in the Random Forest model. Additionally, the Osborne CHi-C data link ELS cREs with genes with low expression. This suggests that either many of these interactions are random noise or CHi-C captures new interactions that have yet to result in gene expression. Both cases warrant further investigation. In the future, we can also investigate features that differentiate between types of links. For example, the Ruan lab reported differences between the gene expression patterns of genes at the anchors of their CTCF and POLII ChIA-PET with the CTCF genes having ubiquitous expression and POLII genes having cell type specific expression. We can incorporate these features into our model to predict what types of interactions link the ELS-gene pair.

During the next phase of the ENCODE project, the Ruan labs and Aiden labs plan on generating new ChIA-PET and Hi-C datasets in cell types relevant to the Registry of cREs. Therefore, establishing these methods for creating benchmarks will aid in the future evaluation of target gene methods.

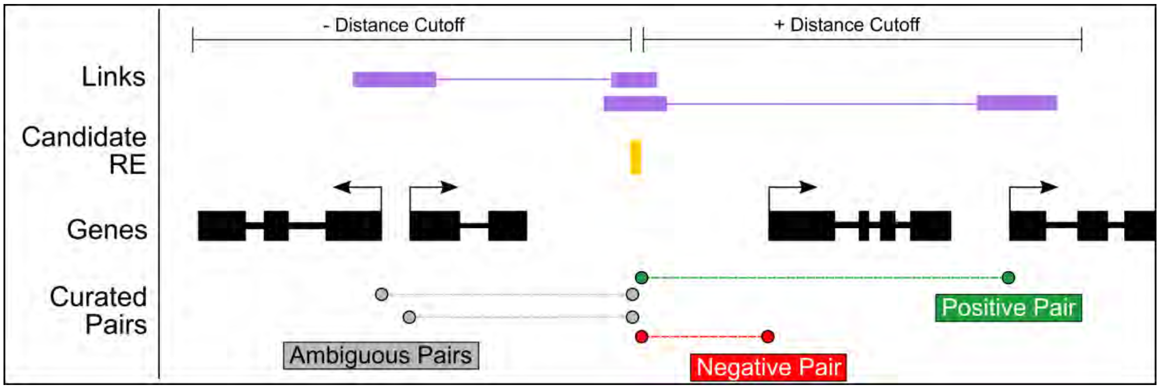


Figure 3.1 | Method for curating ELS-Gene pairs. To include a link in one of our benchmark datasets, it must overlap a ELS cRE active GM12878 (yellow) and the proximal region surrounding a TSS (+/- 2 kb). Links that overlap multiple TSSs are not included in either the positive or negative set of links (black listed). For the negative set we included all genes with a TSS within a +/- distance based on the 95th percentile.

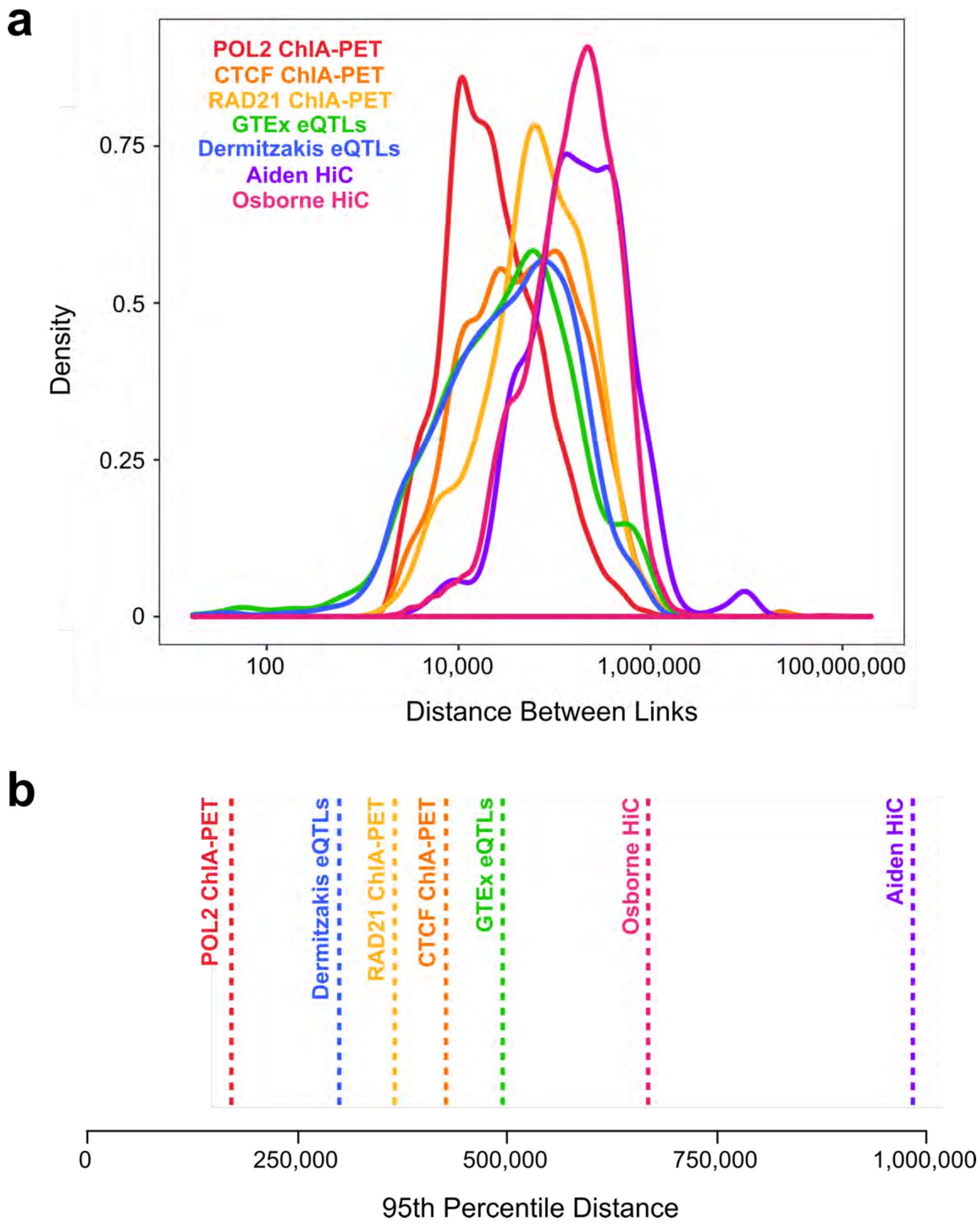


Figure 3.2 | Distance distributions for benchmark datasets. **a,** Distribution of distances between ELS cREs and gene TSSs in benchmark links. **b,** Lines indicating 95th percentile of distance for each benchmark dataset.

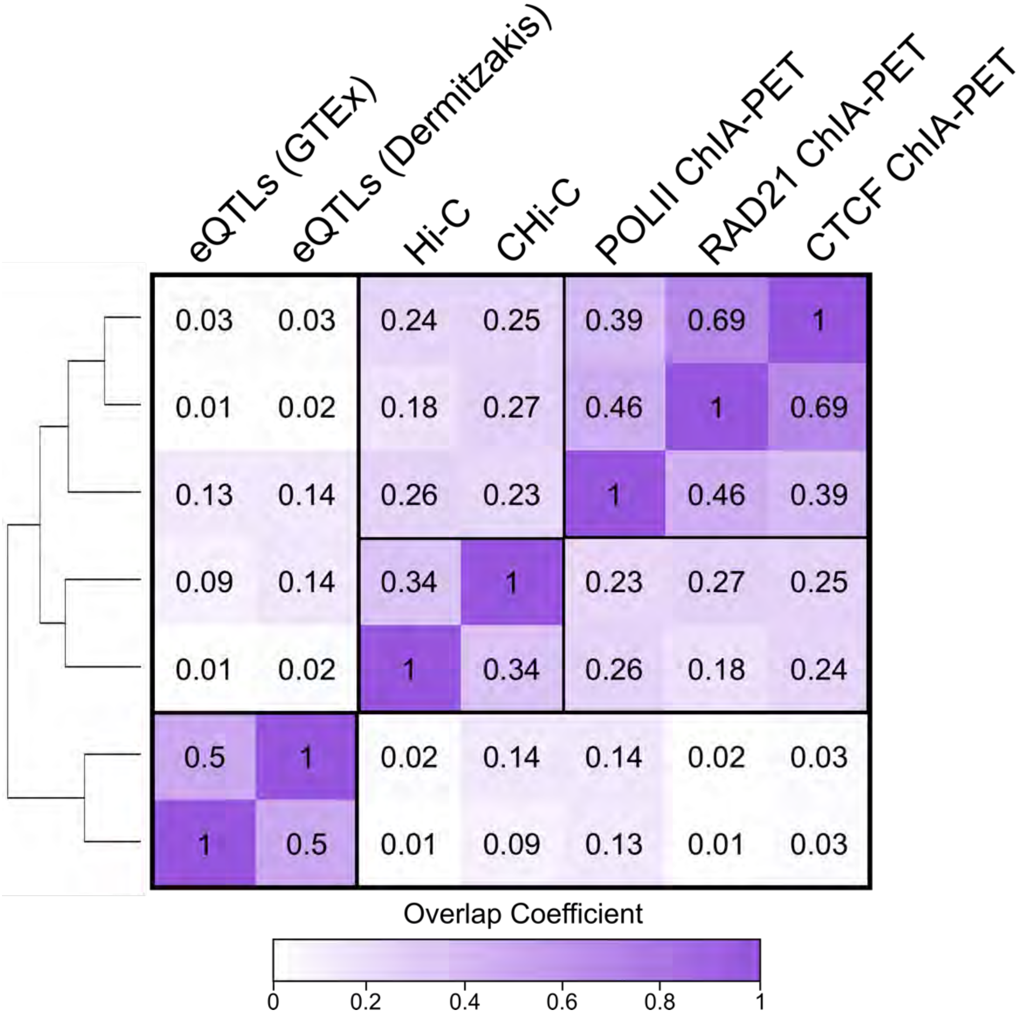


Figure 3.3 | Overlap of benchmark datasets. Heatmap displaying overlap coefficients for each pairwise comparison of benchmark datasets. Benchmark datasets cluster by type of dataset (i.e. ChIA-pET, Hi-C, and eQTLs).

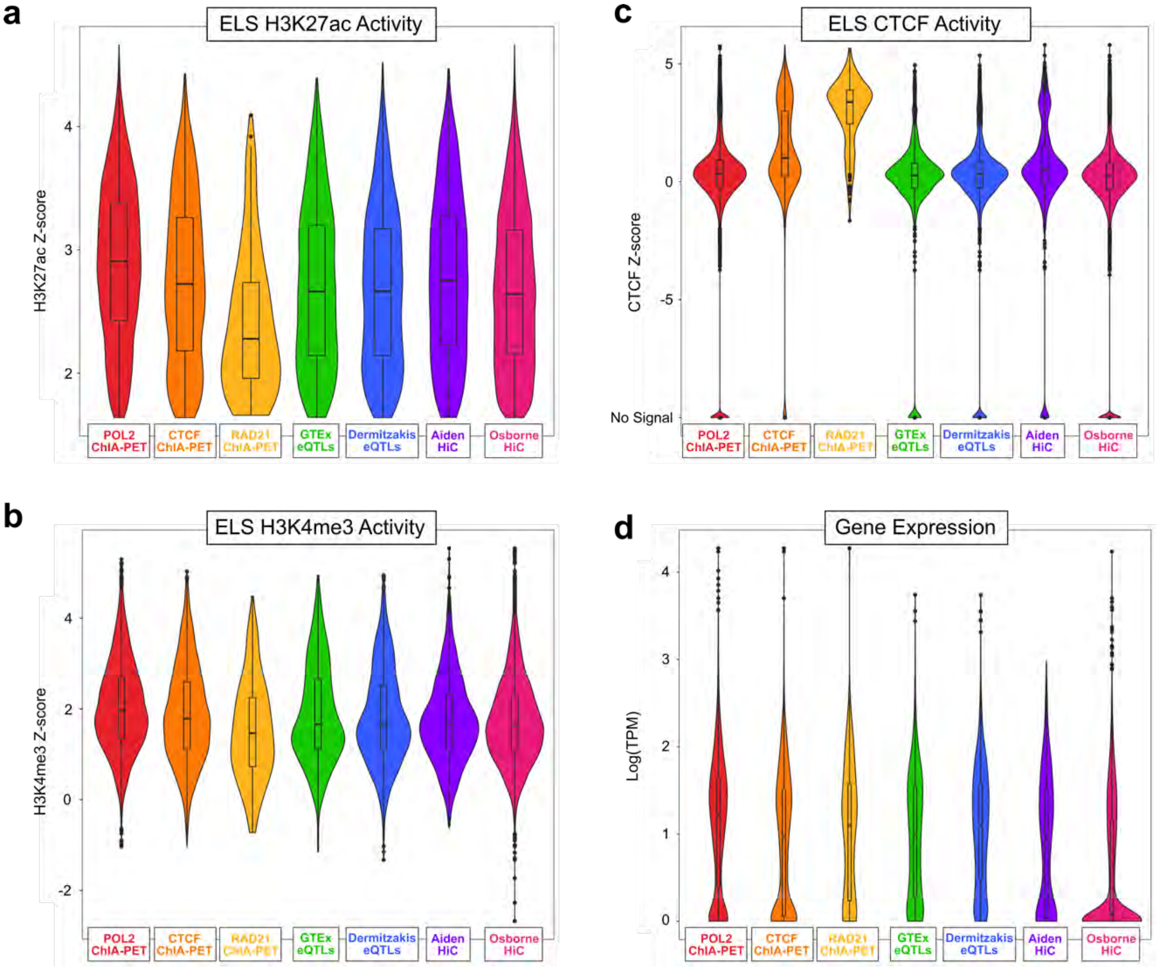


Figure 3.4 | Activity of ELS cREs and expression of genes in benchmark datasets. **a,** H3K27ac **b,** H3K4me3 and **c,** CTCF Z-scores in GM12878 for ELS cREs in benchmark datasets. **d,** Gene expression in TPM for genes in benchmark datasets

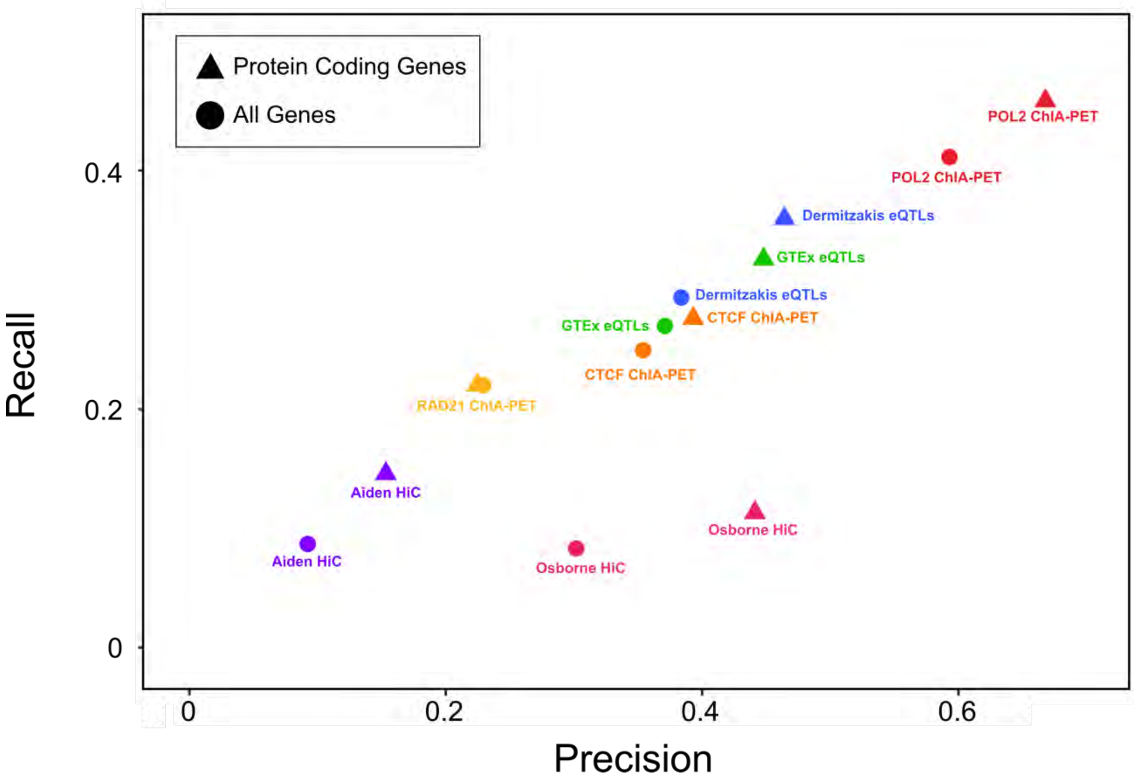


Figure 3.5 | Performance for closest gene method. Precision (X-axis) and recall (Y-axis) for each benchmark dataset using the closest gen method. Results from using all genes are indicated by circles. Results from using protein coding genes are indicated by triangles.

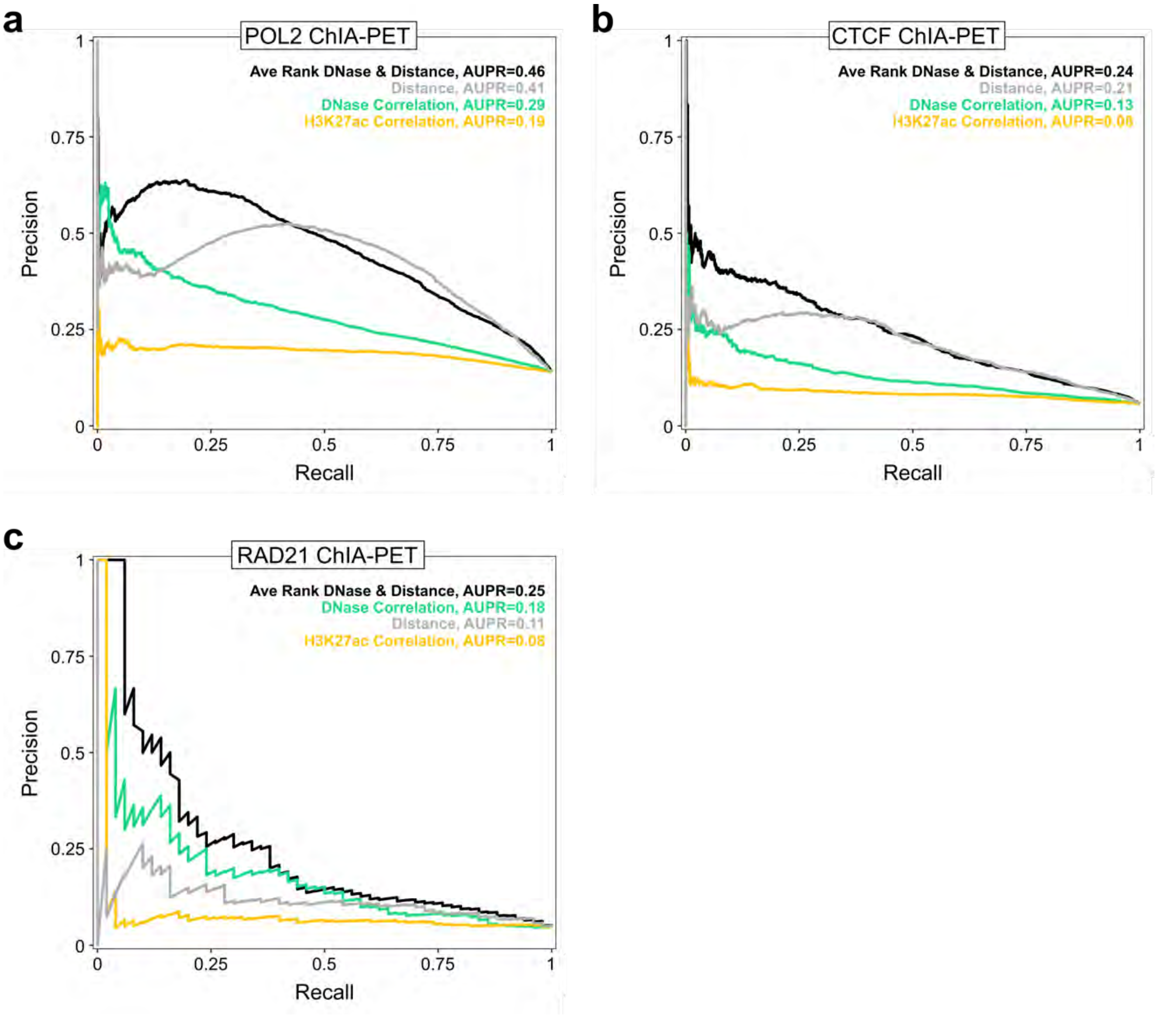


Figure 3.6 | PR curves for unsupervised target gene prediction methods: ChIA-PET datasets. Precision recall curves for enhancer gene pairs ranked by distance, DNase and H3K27ac Z-score Spearman correlation, and the average rank of DNase correlation and distance for **a,** POLII, **b,** CTCF, and **c,** RAD21 ChIA-PET datasets

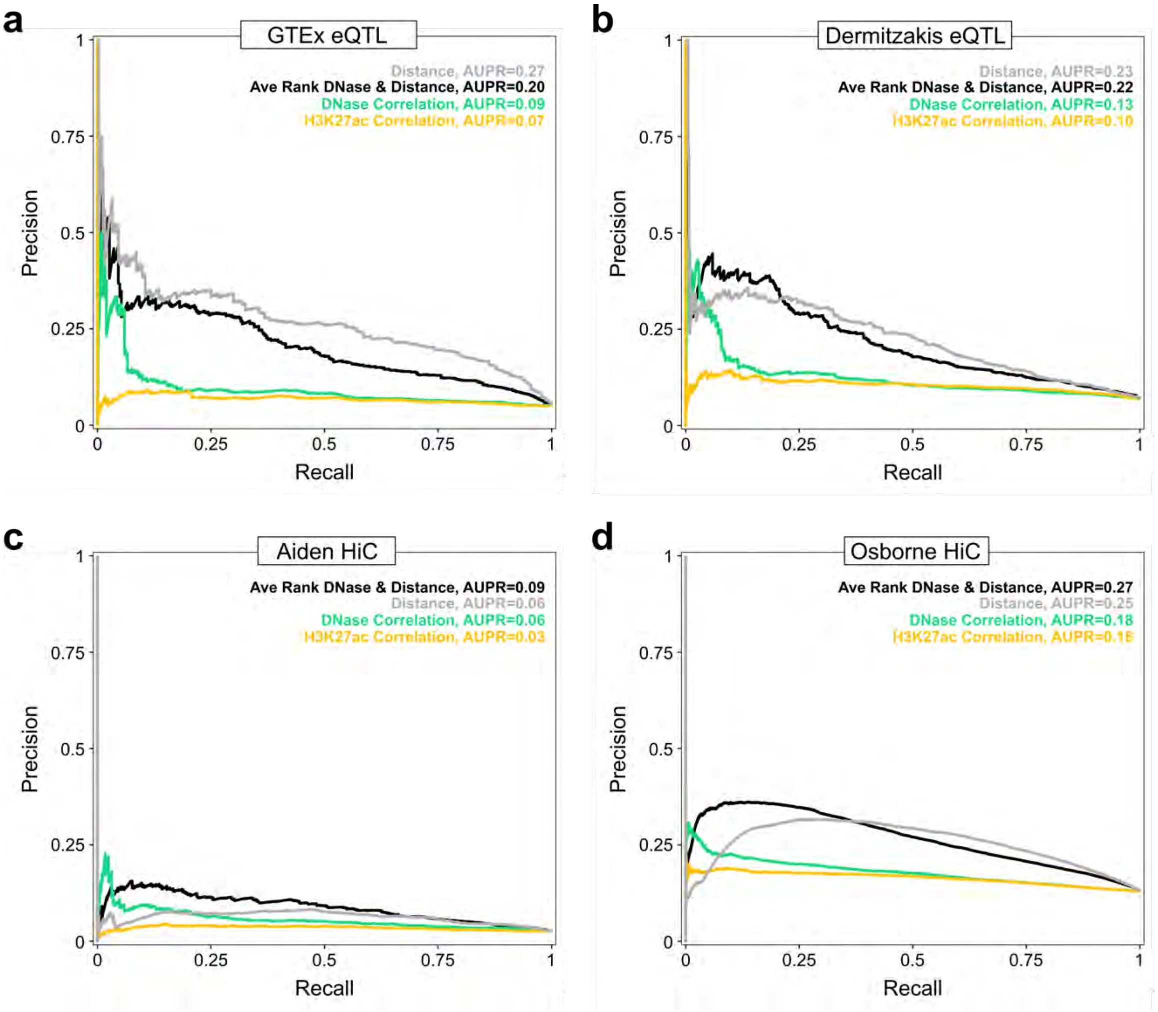


Figure 3.7 | PR curves for unsupervised target gene prediction methods: eQTL and Hi-C datasets. Precision recall curves for enhancer gene pairs ranked by distance, DNase and H3K27ac Z-score Spearman correlation, and the average rank of DNase correlation and distance for **a**) GTEx eQTLs, **b**) Dermitazkis lab eQTL, **c**) Aiden lab Hi-C and **d,** Osborne lab CHi-C datasets

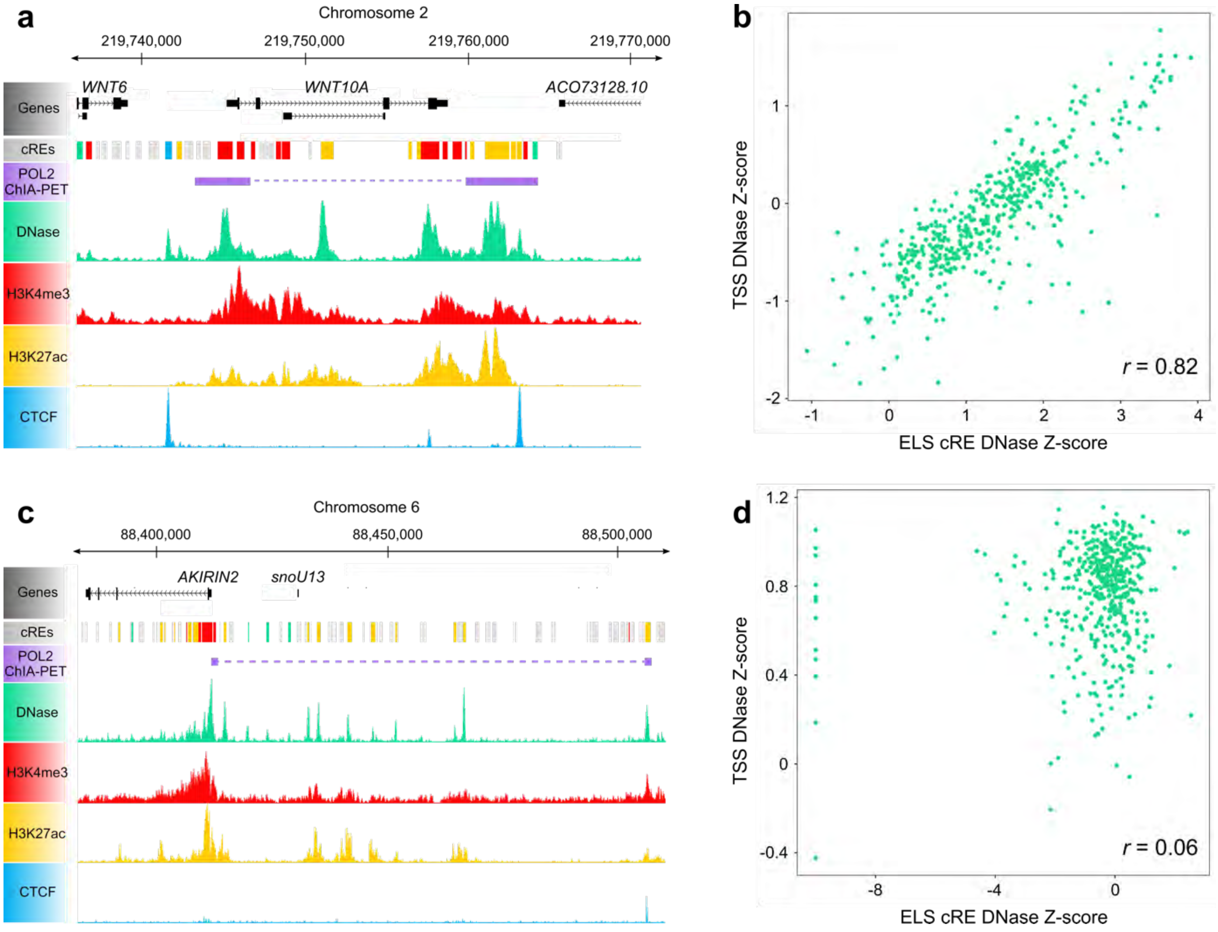


Figure 3.8 | Correlation of DNase signal between of ELS-gene pairs. **a,** ELS cRE EH37E0572541 *WNT10A* are paired by a POLII ChIA-PET link. EH37E0572541and the promoter of *WNT10A* DNase signal correlation coefficient of 0.82. **b,** ELS cRE EH37E0853090 and *AKIRIN2* are paired by a POLII ChIA-PET link and GTEx eQTL. EH37E0853090 and the promoter of *AKIRIN2* DNase signal correlation coefficient of 0.06.

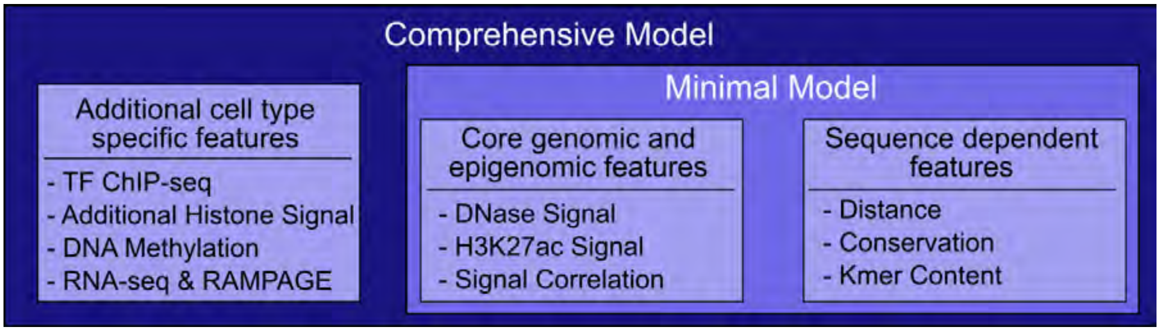


Figure 3.9 | Proposed random forest models for predicting ELS-gene pairs. We propose developing two models. The minimal model will only cell type specific DNase and H3K27ac as well as cell type agnostic features such as conservation and distance. With the comprehensive model, we will integrate as many data types as possible to generate the best performing model.

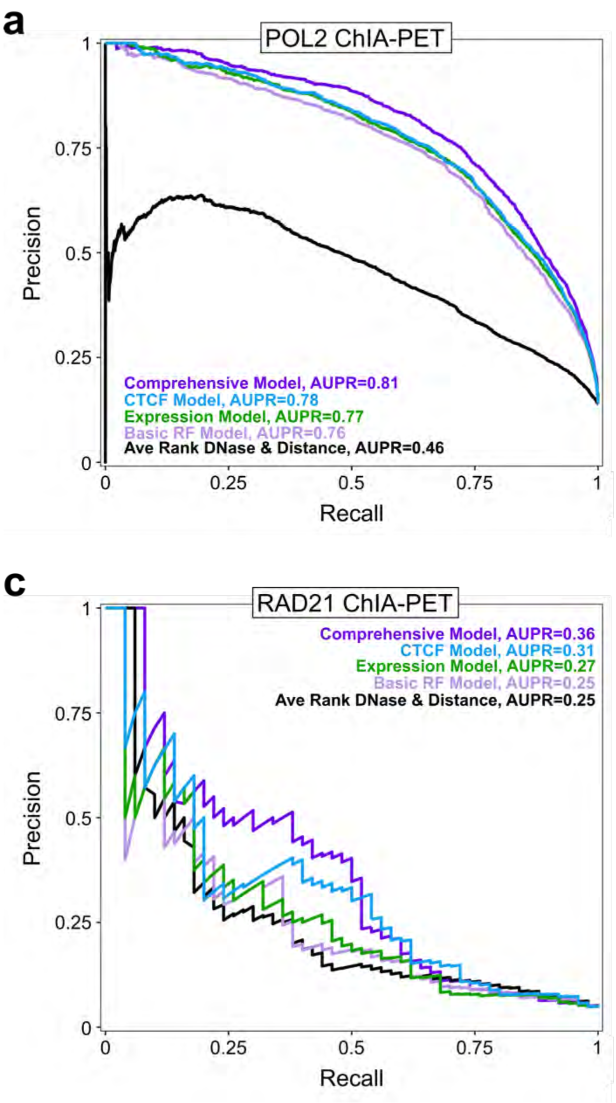


Figure 3.10 | PR curves for Random Forest model predicting ELS-Gene links: ChIA-PET datasets. Precision recall curves for the average rank of DNase correlation and distance, and basic, expression, CTCF, and comprehensive Random Forest models for **a,** POLII, **b,** CTCF, and **c,** RAD21 ChIA-PET datasets

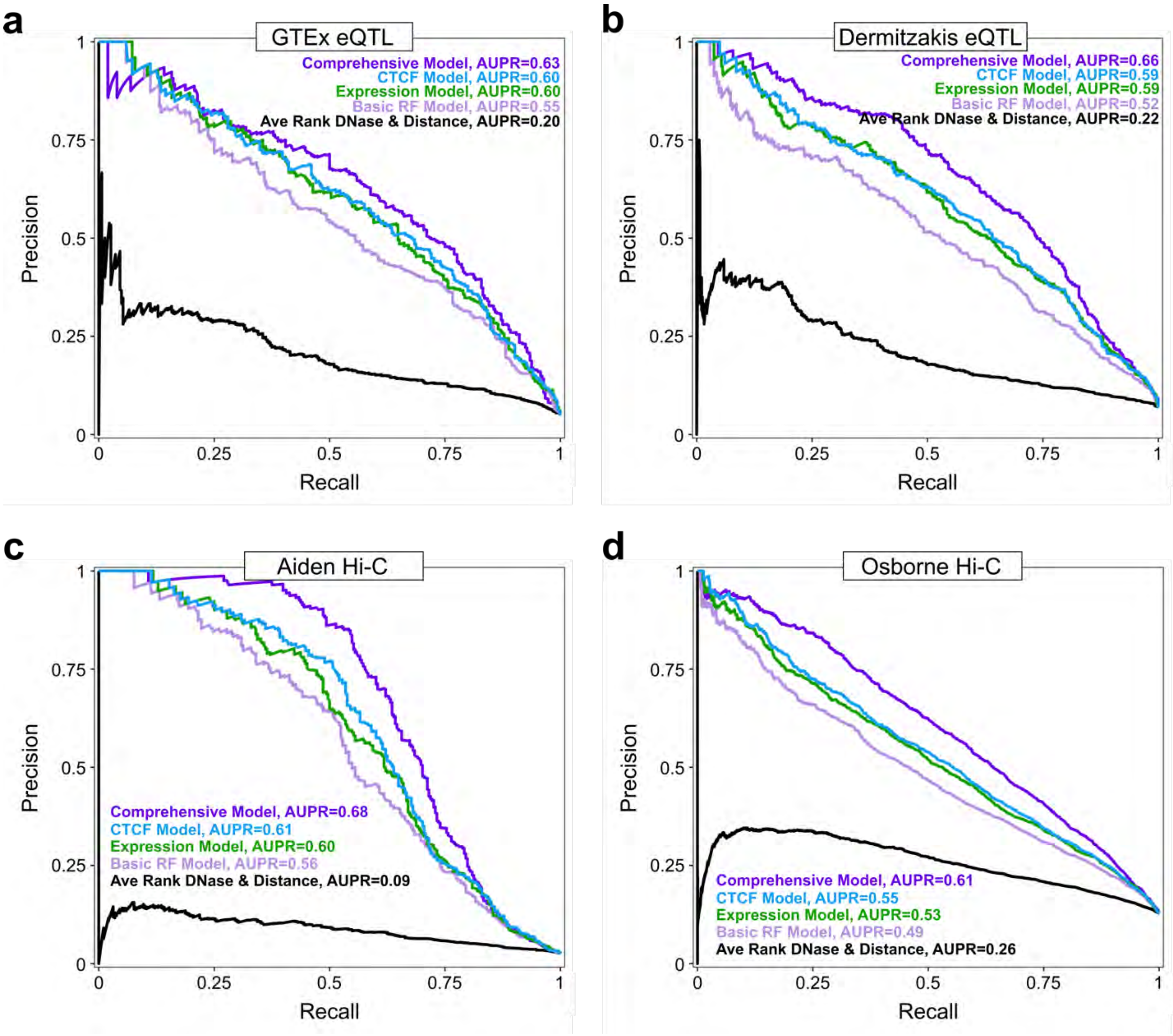


Figure 3.11 | PR curves for Random Forest model predicting ELS-Gene links: eQTL and Hi-C datasets. Precision recall curves for the average rank of DNase correlation and distance, and basic, expression, CTCF, and comprehensive Random Forest models for **a**) GTEx eQTLs, **b**) Dermitazkis lab eQTL, **c**) Aiden lab Hi-C and **d,** Osborne lab CHi-C datasets

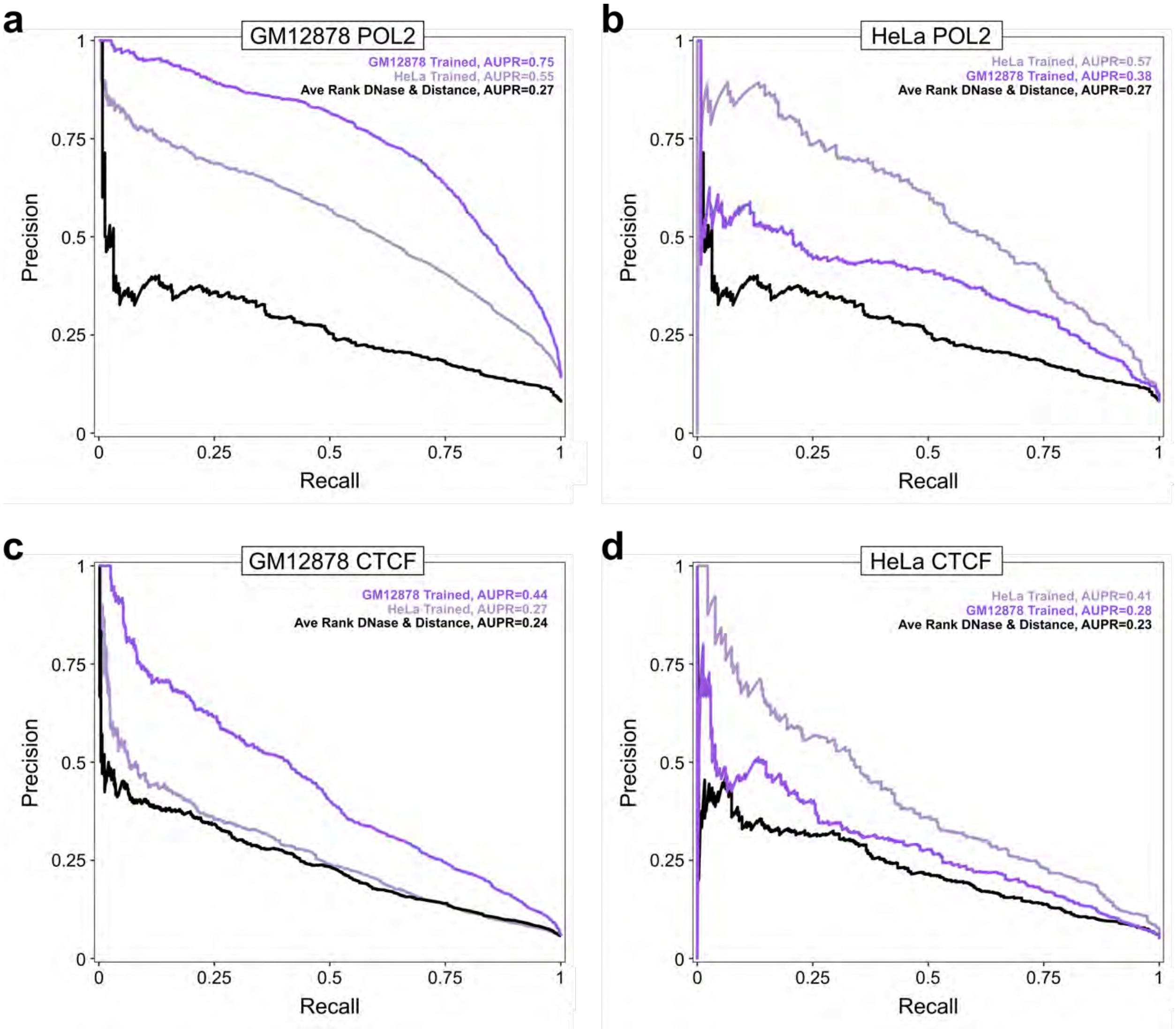


Figure 3.12 | PR curves for Random Forest models trained and validated in different cell types. Precision recall curves for the average rank of DNase correlation and distance, and Random Forest models trained using GM12878 data or HeLa data for ELS-gene pairs from **a,** POLII ChIA-PET from GM12878, **b,** POLII ChIA-PET from HeLa, **c,** CTCF ChIA-PET from GM12878, and **d,** CTCF ChIA-PET from HeLa

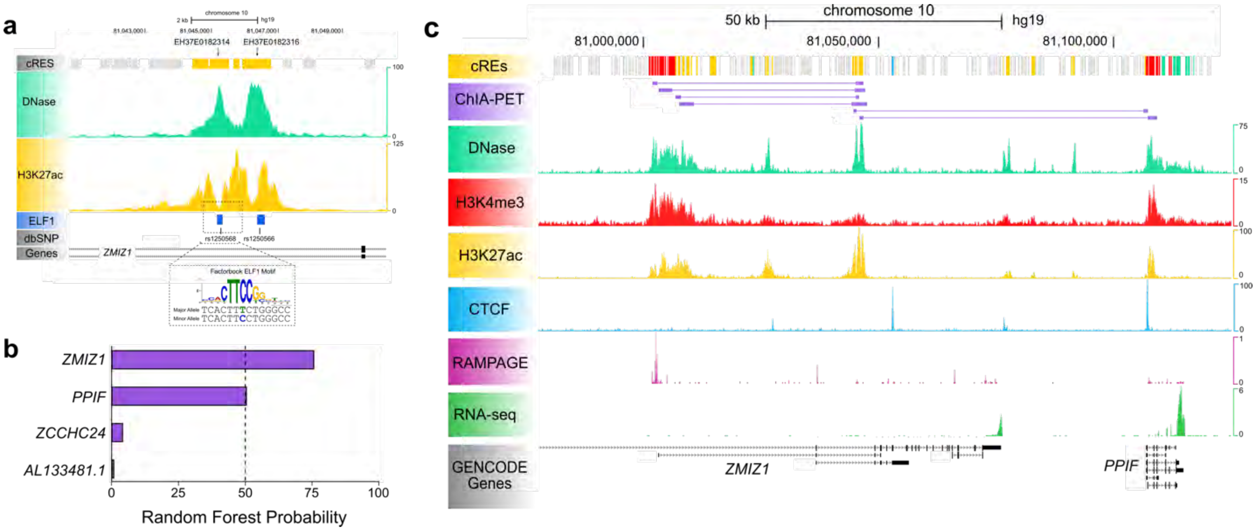


Figure 3.13 | Predicting genes linked with MS variant rs1250568. **a**, MS variant rs1250568 overlaps ELS cRE EH37E0182314 which has high DNase (green) and H3K27ac (yellow) signal in GM12878. Rs1250568 overlaps a ELF1 ChIP-seq peak (blue) and ELF motif site. **b**, Results from comprehensive Random Forest model where bars indicated the probability of each gene being linked with EH37E0182314. **c**, Genome browser view of the locus showing POLII ChIA-PET links validating the predicting links between rs1250568 and *ZMIZ1* and *PPIF*

Table 3.1 | Benchmark Datasets

	Number of Enhancer-Gene Pairs			
Dataset	**Positive Set**	**Negative Set**	**% Positive**	**95th Percentile Distance**
POLII ChIA-PET[67]	12,118	75,380	13.85%	170,163
CTCF ChIA-PET[67]	4,354	72,070	5.70%	426,155
RAD21 ChIA-PET	198	3,979	4.74%	365,097
Promoter Capture Hi-C[65]	48,638	328,724	12.89%	667,267
High Resolution Hi-C[63]	1,132	43,472	2.54%	983,020
GTEX eQTLs	1,162	24,082	4.60%	493,764
Dermitazakis (2013) eQTLs[114]	2,145	30,223	6.63%	298,777

Table 3.2a | AUROC for unsupervised methods

	H3K27ac				DNase					
	Pearson		Spearman		Pearson		Spearman			
	Raw	Z-score	Raw	Z-score	Raw	Z-score	Raw	Z-score	Distance	Average Rank
POLII ChIA-PET	0.5837	0.6234	0.6159	0.6220	0.6225	0.6969	0.6424	0.7070	0.8339	0.8327
CTCF ChIA-PET	0.6082	0.6149	0.6202	0.6197	0.6520	0.6859	0.6835	0.6904	0.8044	0.8065
RAD21 ChIA-PET	0.5897	0.6328	0.6358	0.6044	0.6486	0.7220	0.7143	0.7253	0.7466	0.7954
GTEx eQTLs	0.5511	0.5919	0.6159	0.6010	0.5945	0.6223	0.6053	0.6346	0.8764	0.8170
Dermitzakis eQTL	0.5920	0.6346	0.6363	0.6487	0.5829	0.6388	0.5960	0.6442	0.7436	0.7594
Osborne CHi-C	0.5799	0.5838	0.5832	0.5898	0.5704	0.5868	0.5732	0.6003	0.7633	0.7257
Aiden HiC	0.5706	0.5994	0.5956	0.6008	0.6265	0.6795	0.6188	0.6855	0.7767	0.7823
Average	0.5849	0.6195	0.6248	0.6192	0.6294	0.6818	0.6614	0.6893	0.8153	0.8129

Table 3.2b | AUPR for unsupervised methods

	H3K27ac				DNase					
	Pearson		Spearman		Pearson		Spearman			
	Raw	Z-score	Raw	Z-score	Raw	Z-score	Raw	Z-score	Distance	Average Rank
POLII ChIA-PET	0.1761	0.1891	0.1894	0.1898	0.2152	0.2700	0.2285	0.2911	0.4119	0.4553
CTCF ChIA-PET	0.0833	0.0806	0.0847	0.0846	0.1012	0.1234	0.1097	0.1300	0.2066	0.2399
RAD21 ChIA-PET	0.0641	0.0657	0.0683	0.0819	0.0853	0.1544	0.0995	0.1800	0.1145	0.2507
GTEx eQTLs	0.0597	0.0630	0.0710	0.0674	0.0802	0.0822	0.0810	0.0939	0.2667	0.2049
Dermitzakis eQTL	0.0969	0.1053	0.0999	0.1044	0.0928	0.1142	0.0944	0.1256	0.2252	0.2158
Osborne CHi-C	0.1579	0.1590	0.1588	0.1614	0.1643	0.1662	0.1635	0.1740	0.2481	0.2595
Aiden HiC	0.0321	0.0342	0.0345	0.0345	0.0393	0.0517	0.0396	0.0564	0.0624	0.0875
Average	0.0960	0.1007	0.1027	0.1056	0.1149	0.1488	0.1226	0.1641	0.2450	0.2877

Table 3.3a | AUROC for Random Forest models

				Gene Expression						
	DNase	H3K27ac	Basic	Basic	CTCF	POLII	p300	RAMPAGE	His Mods	All
POLII ChIA-PET	0.9285	0.9327	0.9348	0.9396	0.9409	0.9415	0.9417	0.9411	0.9464	0.9487
CTCF ChIA-PET	0.8930	0.8934	0.8967	0.9022	0.9158	0.9064	0.9046	0.9039	0.9111	0.9227
RAD21 ChIA-PET	0.7798	0.7847	0.7841	0.7762	0.8184	0.7777	0.7748	0.7797	0.7676	0.8037
GTEx eQTLs	0.9289	0.9292	0.9297	0.9401	0.9416	0.9419	0.9401	0.9405	0.9432	0.9447
Dermitzakis eQTL	0.8970	0.8984	0.9043	0.9221	0.9232	0.9241	0.9251	0.9238	0.9312	0.9337
Osborne CHiC		0.8263	0.8330	0.8488	0.8521	0.8536	0.8522	0.8507	0.8731	0.8736
Aiden HiC	0.9216	0.9218	0.9220	0.9270	0.9285	0.9263	0.9247	0.9256	0.9310	0.9302
Average	0.8915	0.8838	0.8864	0.8937	0.9029	0.8959	0.8948	0.8950	0.9005	0.9082

Table 3.3b | AUPR for Random Forest models

				Gene Expression						
	DNase	H3K27ac	Basic	Basic	CTCF	POLII	p300	RAMPAGE	His Mods	All
POLII ChIA-PET	0.7353	0.7466	0.7557	0.7718	0.7761	0.7778	0.7803	0.7773	0.7985	0.8065
CTCF ChIA-PET	0.4342	0.4387	0.4488	0.4712	0.5342	0.4861	0.4819	0.4826	0.5306	0.5857
RAD21 ChIA-PET	0.2594	0.2546	0.2532	0.2651	0.3137	0.3026	0.2812	0.2874	0.3041	0.3569
GTEx eQTLs	0.5301	0.5392	0.5484	0.5982	0.6040	0.6051	0.6030	0.6041	0.6260	0.6320
Dermitzakis eQTL	0.4950	0.4917	0.5184	0.5872	0.5938	0.6008	0.6017	0.5991	0.6465	0.6617
Osborne CHiC		0.4753	0.4930	0.5329	0.5453	0.5446	0.5408	0.5408	0.6071	0.6074
Aiden HiC	0.4949	0.5405	0.5583	0.5963	0.6139	0.6104	0.6350	0.6082	0.6823	0.6825
Average	0.4915	0.4981	0.5108	0.5461	0.5687	0.5610	0.5606	0.5571	0.5993	0.6190

Table 3.4 | Feature importance for basic Random Forest Model

	POLII ChIA-PET	CTCF ChIA-PET	RAD21 ChIA-PET	GTEx eQTLs	Dermitzakis eQTL	Osborne CHi-C	Aiden Hi-C	Average
Distance	0.2578	0.1667	0.1074	0.1777	0.1486	0.1532	0.1030	0.1592
Promoter DNase	0.0706	0.0611	0.0723	0.0602	0.0611	0.0636	0.0674	0.0652
Enhancer H3K27ac Mean	0.0446	0.0528	0.0602	0.0553	0.0618	0.0615	0.0655	0.0574
Promoter Conservation	0.0395	0.0515	0.0605	0.0601	0.0581	0.0611	0.0676	0.0569
Promoter H3K27ac	0.1027	0.0578	0.0666	0.0633	0.0630	0.0597	0.0715	0.0692
Enhancer H3K27ac SD	0.0418	0.0516	0.0547	0.0571	0.0578	0.0579	0.0640	0.0550
Promoter H3K27ac Mean	0.0339	0.0466	0.0506	0.0495	0.0486	0.0554	0.0575	0.0489
Promoter H3K27ac SD	0.0352	0.0456	0.0543	0.0487	0.0487	0.0553	0.0509	0.0484
H3K27ac Correlation	0.0416	0.0536	0.0656	0.0511	0.0611	0.0525	0.0557	0.0545
DNase Correlation	0.0564	0.0703	0.0853	0.0536	0.0515	0.0474	0.0593	0.0605
K-mer Correlation	0.0413	0.0508	0.0705	0.0542	0.0527	0.0461	0.0565	0.0532
Enhancer H3K27ac	0.0376	0.0432	0.0369	0.0440	0.0415	0.0427	0.0398	0.0408
Promoter DNase Mean	0.0344	0.0436	0.0373	0.0377	0.0440	0.0424	0.0406	0.0400
Enhancer Conservation	0.0323	0.0417	0.0381	0.0371	0.0408	0.0421	0.0437	0.0394
Promoter DNase SD	0.0331	0.0404	0.0309	0.0451	0.0454	0.0420	0.0413	0.0397
Enhancer DNase	0.0357	0.0405	0.0378	0.0386	0.0385	0.0412	0.0397	0.0388
Enhancer DNase SD	0.0313	0.0424	0.0392	0.0336	0.0385	0.0387	0.0380	0.0374

Enhancer DNase Mean	0.0302	0.0397	0.0320	0.0333	0.0383	0.0373	0.0380	0.0355

Table 3.5 | Feature importance for expression Random Forest model

	POLII ChIA-PET	CTCF ChIA-PET	RAD21 ChIA-PET	GTEx eQTLs	Dermitzakis eQTL	Osborne CHi-C	Aiden Hi-C	Average
Distance	0.2512	0.1633	0.1016	0.1765	0.1470	0.1521	0.0999	0.1559
Expression	0.0759	0.0518	0.0620	0.0713	0.0828	0.0530	0.0571	0.0648
Promoter H3K27ac	0.0875	0.0522	0.0580	0.0554	0.0552	0.0542	0.0655	0.0611
Promoter DNase	0.0583	0.0561	0.0664	0.0535	0.0540	0.0547	0.0624	0.0579
DNase Correlation	0.0507	0.0658	0.0814	0.0478	0.0453	0.0485	0.0546	0.0563
Promoter Conservation	0.0365	0.0490	0.0572	0.0596	0.0549	0.0529	0.0647	0.0535
Promoter DNase Mean	0.0399	0.0496	0.0556	0.0498	0.0542	0.0536	0.0600	0.0518
Promoter DNase SD	0.0389	0.0492	0.0522	0.0517	0.0531	0.0506	0.0596	0.0508
K-mer Correlation	0.0383	0.0477	0.0657	0.0488	0.0474	0.0474	0.0527	0.0497
H3K27ac Correlation	0.0372	0.0491	0.0592	0.0444	0.0518	0.0504	0.0532	0.0494
Promoter H3K27ac Mean	0.0322	0.0439	0.0480	0.0472	0.0454	0.0481	0.0556	0.0458
Promoter H3K27ac SD	0.0325	0.0428	0.0520	0.0450	0.0453	0.0474	0.0485	0.0448
Enhancer H3K27ac	0.0353	0.0418	0.0340	0.0401	0.0382	0.0415	0.0376	0.0384
Enhancer H3K27ac Mean	0.0325	0.0423	0.0362	0.0350	0.0404	0.0416	0.0388	0.0381
Enhancer H3K27ac SD	0.0309	0.0387	0.0299	0.0416	0.0415	0.0414	0.0386	0.0375
Enhancer Conservation	0.0302	0.0394	0.0371	0.0341	0.0373	0.0428	0.0414	0.0375
Enhancer DNase	0.0340	0.0386	0.0357	0.0360	0.0351	0.0415	0.0375	0.0369
Enhancer DNase SD	0.0295	0.0407	0.0367	0.0312	0.0359	0.0395	0.0362	0.0357
Enhancer DNase Mean	0.0286	0.0381	0.0313	0.0308	0.0352	0.0388	0.0361	0.0341

Table 3.6 | Feature importance for CTCF Random Forest model

	POLII ChIA-PET	CTCF ChIA-PET	RAD21 ChIA-PET	GTEx eQTLs	Dermitzakis eQTL	Osborne CHi-C	Aiden Hi-C	Average
Distance	0.2436	0.1559	0.0934	0.1699	0.1409	0.1460	0.0934	0.1490
Promoter CTCF	0.0346	**0.0745**	**0.1057**	0.0462	0.0521	0.0481	**0.0562**	0.0596
Expression	0.0709	0.0451	0.0512	0.0660	0.0765	0.0489	0.0521	0.0587
Promoter H3K27ac	0.0820	0.0464	0.0499	0.0503	0.0502	0.0496	0.0603	0.0555
Promoter DNase	0.0559	0.0490	0.0560	0.0501	0.0490	0.0496	0.0567	0.0523
DNase Correlation	0.0477	0.0589	0.0692	0.0438	0.0418	0.0440	0.0500	0.0508
Promoter Conservation	0.0333	0.0426	0.0508	0.0546	0.0493	0.0479	0.0591	0.0482
Promoter DNase Mean	0.0377	0.0439	0.0460	0.0458	0.0489	0.0485	0.0542	0.0464
K-mer Correlation	0.0354	0.0418	0.0583	0.0459	0.0431	0.0430	0.0484	0.0451
Promoter DNase SD	0.0357	0.0429	0.0420	0.0471	0.0491	0.0457	0.0528	0.0450
H3K27ac Correlation	0.0343	0.0431	0.0518	0.0410	0.0475	0.0461	0.0485	0.0446
Promoter H3K27ac Mean	0.0295	0.0379	0.0420	0.0431	0.0407	0.0437	0.0495	0.0409
Promoter H3K27ac SD	0.0299	0.0374	0.0438	0.0414	0.0411	0.0431	0.0431	0.0400
Enhancer H3K27ac	0.0328	0.0361	0.0291	0.0370	0.0350	0.0377	0.0337	0.0345
Enhancer H3K27ac Mean	0.0298	0.0356	0.0313	0.0318	0.0367	0.0377	0.0351	0.0340
Enhancer H3K27ac SD	0.0280	0.0340	0.0261	0.0379	0.0375	0.0375	0.0349	0.0337
Enhancer Conservation	0.0277	0.0344	0.0318	0.0313	0.0342	0.0387	0.0373	0.0336
Enhancer CTCF	0.0272	0.0405	0.0328	0.0284	0.0317	0.0362	0.0368	0.0334
Enhancer DNase	0.0313	0.0337	0.0310	0.0324	0.0320	0.0376	0.0337	0.0331
Enhancer DNase SD	0.0268	0.0344	0.0306	0.0283	0.0317	0.0354	0.0325	0.0314

Table 3.7 | Feature importance for comprehensive Random Forest model

	POLII ChIA-PET	CTCF ChIA-PET	RAD21 ChIA-PET	GTEx eQTLs	Dermitzakis eQTL	Osborne CHi-C	Aiden Hi-C	Average
Distance	0.1903	0.1212	0.0597	0.1307	0.1080	0.1114	0.0600	0.1116
Promoter CTCF	0.0152	0.0471	0.0714	0.0239	0.0250	0.0227	0.0272	0.0332
Expression	0.0392	0.0233	0.0260	0.0371	0.0433	0.0281	0.0252	0.0317
DNase Correlation	0.0261	0.0359	0.0426	0.0236	0.0216	0.0218	0.0255	0.0282
Promoter H2AFZ	0.0284	0.0249	0.0278	0.0231	0.0275	0.0240	0.0281	0.0262
Promoter H3K4me1	0.0222	0.0244	0.0297	0.0211	0.0241	0.0230	0.0331	0.0254
Promoter H3K27ac	0.0393	0.0208	0.0219	0.0223	0.0214	0.0222	0.0266	0.0249
K-mer Correlation	0.0191	0.0223	0.0347	0.0251	0.0217	0.0212	0.0249	0.0241
Promoter Conservation	0.0154	0.0205	0.0257	0.0297	0.0233	0.0237	0.0281	0.0238
Promoter DNase	0.0242	0.0221	0.0273	0.0225	0.0214	0.0224	0.0246	0.0235
Promoter H3K9ac	0.0297	0.0202	0.0226	0.0221	0.0226	0.0227	0.0236	0.0234
H3K27ac Correlation	0.0163	0.0221	0.0284	0.0206	0.0250	0.0228	0.0245	0.0228
Promoter EP300	0.0237	0.0194	0.0220	0.0213	0.0215	0.0233	0.0282	0.0228
Promoter POLII	0.0235	0.0194	0.0224	0.0248	0.0228	0.0210	0.0233	0.0224
Promoter H3K27me3	0.0146	0.0199	0.0245	0.0220	0.0268	0.0224	0.0256	0.0223
Promoter H3K4me2	0.0198	0.0196	0.0276	0.0203	0.0236	0.0215	0.0229	0.0222
Promoter H3K4me3	0.0227	0.0198	0.0226	0.0212	0.0215	0.0227	0.0238	0.0220
Promoter DNase SD	0.0165	0.0205	0.0198	0.0246	0.0247	0.0211	0.0243	0.0217
Promoter H3K79me2	0.0247	0.0177	0.0196	0.0224	0.0211	0.0208	0.0244	0.0215
Promoter H3K9me3	0.0155	0.0198	0.0233	0.0212	0.0191	0.0209	0.0288	0.0212
Promoter H4K20me1	0.0148	0.0196	0.0214	0.0213	0.0235	0.0212	0.0257	0.0211
Promoter H3K36me3	0.0164	0.0208	0.0222	0.0205	0.0217	0.0215	0.0222	0.0208

Promoter DNase Mean	0.0148	0.0195	0.0211	0.0217	0.0225	0.0215	0.0236	0.0207
Promoter H3K27ac Mean	0.0132	0.0181	0.0212	0.0214	0.0195	0.0209	0.0220	0.0195
Promoter RAMPAGE	0.0220	0.0164	0.0204	0.0184	0.0223	0.0147	0.0203	0.0192
Promoter H3K27ac SD	0.0133	0.0177	0.0223	0.0197	0.0193	0.0206	0.0195	0.0189
Enhancer H3K79me2	0.0141	0.0152	0.0146	0.0158	0.0171	0.0177	0.0186	0.0162
Enhancer CTCF	0.0128	0.0204	0.0157	0.0126	0.0146	0.0162	0.0177	0.0157
Enhancer H3K27ac SD	0.0131	0.0156	0.0118	0.0178	0.0179	0.0169	0.0154	0.0155
Enhancer H3K9me3	0.0141	0.0174	0.0138	0.0142	0.0154	0.0176	0.0156	0.0154
Enhancer Conservation	0.0127	0.0161	0.0144	0.0140	0.0158	0.0175	0.0175	0.0154
Enhancer H3K27ac Mean	0.0134	0.0157	0.0142	0.0142	0.0171	0.0165	0.0156	0.0152
Enhancer H3K36me3	0.0144	0.0159	0.0134	0.0152	0.0150	0.0173	0.0151	0.0152
Enhancer H2AFZ	0.0146	0.0170	0.0131	0.0137	0.0142	0.0159	0.0168	0.0151
Enhancer POLII	0.0137	0.0160	0.0127	0.0168	0.0143	0.0173	0.0140	0.0150
Enhancer EP300	0.0162	0.0151	0.0142	0.0139	0.0138	0.0168	0.0144	0.0149
Enhancer H4K20me1	0.0134	0.0161	0.0117	0.0154	0.0143	0.0177	0.0156	0.0149
Enhancer H3K4me1	0.0138	0.0156	0.0117	0.0149	0.0155	0.0168	0.0159	0.0149
Enhancer H3K27me3	0.0131	0.0158	0.0124	0.0172	0.0146	0.0173	0.0137	0.0149
Enhancer DNase	0.0145	0.0148	0.0139	0.0141	0.0137	0.0165	0.0146	0.0146
Enhancer H3K27ac	0.0145	0.0153	0.0120	0.0156	0.0141	0.0158	0.0145	0.0145
Enhancer DNase SD	0.0126	0.0166	0.0144	0.0130	0.0150	0.0157	0.0141	0.0145
Enhancer DNase Mean	0.0122	0.0151	0.0125	0.0128	0.0144	0.0153	0.0146	0.0138
Enhancer H3K4me3	0.0126	0.0147	0.0128	0.0126	0.0133	0.0159	0.0135	0.0136

Enhancer H3K9ac	0.0122	0.0145	0.0117	0.0123	0.0134	0.0151	0.0145	0.0134
Enhancer H3K4me2	0.0117	0.0147	0.0120	0.0130	0.0128	0.0151	0.0124	0.0131
Enhancer RAMPAGE	0.0094	0.0094	0.0084	0.0083	0.0090	0.0091	0.0100	0.0091

Table 3.8a | AUROC for cross cell type comparisons

Cell Type	ChIA-PET Target	Average Rank Distance & DNase Correlation	GM12878 Trained RF Model	HeLa Trained RF Model
GM12878	POLII	0.8327	0.9316	0.8654
HeLa	POLII	0.8072	0.8801	0.9193
GM12878	CTCF	0.8065	0.8936	0.8055
HeLa	CTCF	0.8346	0.8611	0.9047

Table 3.8.b | AUPR for cross cell type comparisons

Cell Type	ChIA-PET Target	Average Rank Distance & DNase Correlation	GM12878 Trained RF Model	HeLa Trained RF Model
GM12878	POLII	0.4553	0.7458	0.5465
HeLa	POLII	0.2658	0.3845	0.5663
GM12878	CTCF	0.2399	0.4380	0.2681
HeLa	CTCF	0.2259	0.2827	0.4074

METHODS

Creating Our Benchmark of ELS-Gene Pairs

Defining ELS cREs

We selected all cREs as defined as having enhancer-like signatures in GM12878 per own registry of regulatory elements pipeline. To be classified as an ELS cRE in GM12878, the cRE must have a DNase and H3K27ac Z-scores > 1.64 in GM12878 and either be 1) distal from an annotated TSS or 2) not have a H3K4me3 Z-score > 1.64 in GM12878. In total, there are 27,739 ELS cREs in GM12878.

Processing ChIA-PET Data

We downloaded Ruan lab ChIA-PET data from NCBI's Gene Expression Omnibus (GEO) under the accession GSE72816. We used links from *GSM1872886_GM12878_CTCF_PET_clusters.txt* for CTCF and *GSM1872887_GM12878_RNAPII_PET_clusters.txt* for POLII. We also downloaded ChIA-PET data produced by the Snyder from the ENCODE DCC (experiment ENCSR752QCX). We used links called from both replicates in ENCFF002EMO and ENCFF002EMQ.

To generate ELS-Gene pairs, we intersected the ends of the ChIA-PET links with GM12878 ELS cREs and TSSs from GENCODE 19 genes. We selected all links for which one end of the link overlapped an ELS-cRE and the other end fell within 2 kb of an annotated TSS. We classified links that overlapped an ELS-cRE but linked to more than one TSS as "ambiguous" and added them to a blacklist.

Processing Hi-C Data

We downloaded Hi-C loops generated by the Aiden lab from GEO under the accession GSE63525. We used the lab's called loops from *GSE63525_GM12878_primary+replicate_HiCCUPS_looplist.txt.* We also downloaded CHi-C links generated by the Osborne lab from ArrayExpress under the accession E-MTAB-2323. We used the lab's called links from *TS5_GM12878_promoter-other_significant_interactions.txt*

To generate ELS-Gene pairs, we intersected the ends of the (C)Hi-C links with GM12878 ELS cREs and TSSs from GENCODE 19 genes. We selected all links for which one end of the link overlapped an ELS-cRE and the other end fell within 2 kb of an annotated TSS. We classified links that overlapped an ELS-cRE but linked to more than one TSS as "ambiguous" and added them to a blacklist.

Processing eQTLs

We downloaded eQTLs curated from HaploReg, a database curated by the Kellis lab. To generate the GTEx eQTL pairs, we intersected ELS-cREs with eQTLs from lymphoblastoid cell lines. To generate the Derm eQTL pairs we intersected ELS-cREs with eQTLs from lymphoblastoic cell liens. For each overlap generated an ELS-cRE pair using the overlapping ELS and the gene reported by the eQTL.

Generating negative pairs

For each of the seven datasets, we calculated the 95th percentile of distance between the ELS-gene pairs. We defined distance and the minimum linear distance between a ELS-cRE and any TSS of the linked gene. We then selected all ELS-gene pairs that fell within this distance. For each ELS cRE in this pool, we generated a list of all genes within the 95th percentile distance. For datasets with blacklisted links (ChIA-PET and (C)Hi-C) we removed all genes that appeared on the black list connected to the ELS cRE. We considered all remaining genes negatives.

Generating training, validation, and test sets

After generating positive and negative ELS-gene pairs, we assigned them to training, validation and test sets. For training sets, we randomly selected half of the ELS-cREs and assigned all of their positive and negative paris to the training sets. For the validation and test sets we split the remaining cREs in half and assign the pairs of each to validation and test sets respectively. This results in training, validation and test sets containing roughly 50%, 25%, and 25% of the total number of ELS-gene pairs respectively. All of our analysis was currently evaluated using the validation datasets. We plan on using the test dataset for final comparisons of models.

Predicting Target Genes Using Distance

To identify the closest gene to each ELS cRE, we used the Bedtools command *closest* with ELS cREs as file a and GENCODE V19 TSSs as file b. We repeated this process using a filtered set of TSSs to identify the closest protein coding genes. We calculated the overall precision (TP/(TP+FP)) and recall (TP/(TP+FN)) for each dataset using both sets of TSSs.

To test using distance as a ranking scheme, for each ELS-gene pair we calculated the minimum linear distance between the ELS cRE and every annotated TSS for the gene. We used a custom python script. We calculated the AUROC and AUPR using custom the ROCR package and custom R scripts. Because prediction with larger values are considered a higher rank by the ROCR package, we used the inverse of distance to generated the ROC and PR curves

Correlation Methods

For correlation based methods we tested DNase vs. H3K27ac signal, using raw signal (reported directly from ENCODE bigwig files) vs. Z-score normalized signal (see Chapter II for explanation), and Pearson vs Spearman correlation coefficients. To determine signal at each cRE, we used Z-score and raw signals generated during the creation of the Registry of cREs. For these cREs, Z-score signals were calculated across all rDHSs. For TSSs, we used bigwigaverageoverbed to calculate the average signal in a +/- 500bp window around each TSS. We converted this signal to a Z-score relative to all other TSSs.

Note – these Z-scores are different from PLS-cRE Z-scores. Then using custom python scripts we calculated the correlation between each ELS-gene pair. If a gene had multiple TSSs, we selected the highest correlation coefficient for the pair. We calculated the AUROC and AUPR using custom the ROCR package and custom R scripts.

Random Forest Model

We implemented the Random Forest algorithm using the python package scikit learn. For each test, we ran the algorithm 25 times, generating 100 trees each run. For each ELS-gene pair we reported the average class probability across the 25 runs.

We included the following features in our Random Forest model:

1. Distance = the minimum linear distance between ELS cRE and any of the genes TSSs
2. Expression = expression of gene in GM12878 measured in transcripts per million
3. DNase correlation = Spearman correlation of DNase Z-score normalized signal across 460 cell types
4. H3K27ac correlation = Spearman correlation of DNase Z-score normalized signal across 136 cell types

5. Kmer correlation = Pearson correlation of k-mer correlation (default is 3-mer) for k-mer counts across 1 kb sequences centered at gene TSS and ELS cRE
6. Enhancer DNase Mean = the average DNase Z-score for ELS cRE across all cell types
7. Enhancer DNase SD = the standard deviation of ELS DNase Z-scores across all cell types
8. Enhancer H3K27ac Mean = the average H3K27ac Z-score for ELS cRE across all cell types
9. Enhancer H3K27ac SD = the standard deviation of ELS H3K27ac Z-scores across all cell type Enhancer DNase Mean = the average DNase Z-score for ELS cRE across all cell types
10. Promoter DNase Mean = the average DNase Z-score for +/- 500 bp around surrounding the TSS across all cell types
11. Promoter DNase SD = the standard deviation of ELS DNase Z-score for +/- 500 bp around surrounding the TSS across all cell types
12. Promoter H3K27ac Mean = the average H3K27ac Z-score for +/- 500 bp around surrounding the TSS across all cell types
13. Promoter H3K27ac SD = the standard deviation of ELS H3K27ac Z-score for +/- 500 bp around surrounding the TSS across all cell types

14. Enhancer signal = Signal at ELS cREs including DNase, H3K27ac, POLII, CTCF, EP300, RAMPAGE, histone modifications, and conservation
15. Promoter signal = Signal at ELS cREs including DNase, H3K27ac, POLII, CTCF, EP300, RAMPAGE, histone modifications, and conservation for +/- 500 bp around surrounding the TSS across all cell types

For each benchmark, we generated a model from the training data and validated using the benchmark validation sets. We calculated the AUROC and AUPR using custom the ROCR package and custom R scripts.

Scripts

Scripts for this analysis can be found on GitHub: https://github.com/Jill-Moore/Dissertation/tree/master/Chapter-III/

CHAPTER IV: Functional annotation of noncoding variants reveals role of neural and immune pathways in psychiatric disorders

PREFACE

Results from this chapter were adapted from

Moore and Weng. "Functional annotation of noncoding variants reveals role of neural and immune pathways in psychiatric disorders."

which is currently in preparation. I performed all analysis and generated all the figures that were used in the chapter.

ABSTRACT

Schizophrenia, bipolar disorder, and major depressive disorder are debilitating psychiatric disorders that affect a significant percentage of the population. While the etiologies of these psychiatric disorders are unknown, each have strong hereditary components and studies have demonstrated that they share common genetic risk factors. Genome wide association studies (GWAS) have associated over one hundred single nucleotide polymorphisms (SNPs) with these disorders and a majority of the associated SNPs lie in noncoding regions of the genome. Our aim was to functionally characterize these noncoding psych SNPs. We determined psych SNPs were enriched in ELS cREs in active in brain tissues as well as immune related tissues such as T-cells and the thymus. We also determined that these SNPs regulate genes expressed in brain tissue with roles in

neural pathways. Under the hypothesis that psych SNPs alter gene expression by disrupting transcription factor (TF) binding, we analyzed TF ChIP-seq data and observed the SNPs are enriched in SP4 motifs and binding sites for TFs with enriched expression in developing brain tissue. Finally, we characterized four cases of allele specific binding, demonstrating that specific psych SNPs disrupt TF binding sites. Our findings demonstrate that common genetic variants affect both neural and immune pathways.

INTRODUCTION

Schizophrenia (SCZ), bipolar disorder (BPD), and major depressive disorder (MDD) are three prevalent and debilitating psychiatric disorders that affect millions of people every year. While the causes of these disorders are unknown, studies demonstrate both genetic and environmental factors contribute to their onset. SCZ and BPD are highly heritable (~60% estimated heritability)[76,80,140] and through large scale analyses of national medical records and correlation of genetic risk variants studies have reported that these disorders share common genetic risk factors[83,84]. Though MDD has a lower estimated heritability (30-40%)[82], studies demonstrated that it also shares common genetic risks with SCZ and BPD[84].

Genome wide association studies (GWAS) have successfully identified hundreds of single nucleotide polymorphisms (SNPs) associated with SCZ, BPD, and MDD. The majority of these SNPs are in noncoding regions of the genome, and our understanding of how these variants contribute to disease onset is limited.

There have been efforts to characterize noncoding variants associated with SCZ but their scope was limited to characterizing individual variants, resulting in only a handful of annotated variants[141].

During the third phase of the Encyclopedia of DNA Elements (ENCODE) project we generated the ENCODE Encyclopedia, a collection of high throughput experiments (e.g. DNase-seq, RNA-seq, histone modification and transcription factor ChIP-seq) and higher-level analyses aimed at annotating the human and mouse genomes. In addition to assembling this resource, we also integrated DNase-seq data with ChIP-seq data to create the Registry of candidate Regulatory Elements (cREs), a collection of putative regulatory regions across human and mouse (described in Chapter II). In total, we identified over 1.3M human and 400k mouse cREs each annotated with cell-type specific signatures (e.g., promoter-like, enhancer-like) in over 400 human and 100 mouse cell types. Our goal was to use the Registry of cREs and supporting data from the ENCODE Encyclopedia to functionally characterize noncoding SNPs associated with SCZ, BPD, and MDD. By analyzing enrichments in cRE activity in over 500 cell types and integrating RNA-seq, TF-ChIP-seq data, we hoped to learn more about the genetic contributions of these diseases.

Here we report that SNPs associated with psychiatric diseases (psych SNPs) are enriched in candidate regulatory elements active in brain tissues and neural cells. We also curated lists of potential target genes for these SNPs and determined that these genes are enriched for expression in brain tissue. Using orthologous cREs in mouse, we analyzed temporal patterns of cRE activity during brain development and found specific examples of psych cREs active during brain development. We also determined that psych SNPs are enriched in motifs for neural TFs, particularly SP4, as well as immune related TFs such as IRF1. Additionally, we discovered specific instances of psych SNPs in regulatory elements that alter TF binding by disrupting TF motifs. Our analysis supports a genetic foundation for neural pathways in psychiatric disorders and also suggests a role for immune pathways.

RESULTS

The majority of GWAS signal for psych SNPs is explained by cREs

As of July 2017, there were 139 studies in the NHGRI-EBI GWAS catalog tagged with the terms "schizophrenia," "bipolar disorder," or "major depressive disorder." Due to variations in methodology and sampled populations, we selected one representative study for each disorder, prioritizing studies with the largest number of associated variants (Table 4.1). We also considered variants reported by the Cross-Disorder Group of the Psychiatric Genomics Consortium (PGC) who analyzed a mixed cohort of SCZ, BPD, MDD, autism spectrum disorder (ASD) and attention deficit hyperactivity disorder (ADHD) patients. In total, we curated 96 variants for SCZ, 23 for BPD, 43 for MDD, and 73 for cross-disorders (CD). These 235 associations, along with 6,479 SNPs in high ($r^2 > 0.7$) linkage disequilibrium

(LD), amount to 6,714 SNPs associated with psychiatric disorders (psych SNPs) in 233 regions of high LD, which we refer to as LD blocks.

`We began by determining the genetic context of psych SNPs using GENCODE V19 gene annotations. As expected, the majority of psych SNPs (99%) are in noncoding regions of the genome; only ~1% overlap coding exons (Table 4.2). Of these, only two variants (rs4584886 and rs678) are predicted by PROVEAN and SWIFT to be deleterious and damaging to the resulting protein (Table 4.3). While these two SNPs are likely causal, the majority of GWAS signal for psych SNPs is from noncoding regions of the genome.

To annotate these noncoding regions, we overlapped psych SNPs with human and mouse cREs. On average, 20% of psych SNPs overlapped a human cRE accounting for 79% of LD blocks (Table 4.4a,b). Of these, the majority, 76%, overlap cREs with enhancer-like signatures (ELS), while 20% overlap cREs with promoter-like signatures (PLS) and 4% overlap CTCF-only cREs (Table 4.5). An average of 7% of SNPs overlapped orthologous mouse cREs accounting for 46% of LD blocks (Table 4.4c,d). Individually, psych SNPs were slightly more enriched at cREs compared to controls (~1.2 and 1.4 fold enrichment for human and mouse cREs respectively), but there was no significant difference when we performed this analysis with the LD blocks. These results suggest the majority of signal (~79%) from these GWAS can be explained by variants in cREs and that additional annotation of these regions will give us further insight into the mechanisms that underlie these disorders.

Psych SNPs are enriched in cREs active in brain regions

Previous work has demonstrated that variants associated with human disease are enriched in regulatory elements active in disease-relevant cell and tissue types. For example, SNPs associated with Crohn's disease were enriched in T-helper cell DNase peaks[6], and SNPs associated with cholesterol levels are enriched in liver H3K27ac peaks[26]. The Schizophrenia Working Group of the PGC reported their SCZ associated loci were enriched in brain-specific and B cell-specific H3K27ac peaks when looking across 35 tissues from the Roadmap epigenomics project[105]

We wanted to extend this work by analyzing DNase and H3K27ac signals at cREs overlapping psych SNPs. Using data from the Registry of cREs, we calculated activity enrichments using 540 cell and tissue types (462 DNase and 136 H3K27ac). Additionally, during the third phase of the ENCODE project, production labs generated H3K27ac, DNase and RNA-seq data for twelve tissues across mouse embryonic development (eight surveyed time points). These experiments enabled us to analyze enrichment for specific temporal patterns of activity in mouse across development. To calculate enrichment, we counted the number of overlapping cREs with a signal (DNase or H3K27ac) Z-score greater than 2, indicating high signal in that cell type. To prevent over counting of SNPs in LD, we pruned our results by only reporting one hit per LD block for each cell type (see methods). We calculated enrichment using Fisher's exact test. Additionally, differing number of SNPs between the studies makes using a uniform p-value cutoff difficult. P-values from Fisher's exact test inflate as sample size increases, and therefore it is ideal to use more stringent cutoffs for studies with more reported SNPs. We also used a more conservative method for calculating our FDR than previous studies[26] (see methods). Therefore, for our analysis, we focused the top 5 most enriched tissues with at least $p<0.05$ for each study, noting those that meet our FDR threshold of 5%.

Overall, we detected significant enrichments for SCZ and CD SNPs in cREs active in neural cells and brain tissues (Figures 4.1 and 4.2, External Table 4.1). For SCZ SNPs, when we filtered cREs using H3K27ac Z-scores, the five most significantly enriched cell types were temporal lobe, angular gyrus, middle frontal area, iPS DF 19.11, and caudate nucleus (Figure 4.1a). The enrichment in neural tissues was so pronounced that when we expanded our search to the top ten most significantly enriched cell types, seven were from brain tissues or neural cells. Of the remaining three, two were iPSCs, and one was from fetal thymus (External Table 4.1). Even after removing variants on chromosome 6 to account for SNPs in the major histocompatibility complex (MHC), we still observed enrichment in fetal thymus (External Table 4.1). When we filtered cREs using DNase Z-scores, we also observed enrichment in brain tissues (fetal brain and superior temporal gyrus) and neural cells (neuronal progenitor and stem cells) (External Table 4.2). Interestingly, the most significant enrichment was for NCI-H226, a lung cancer cell line. When we clustered DNase cell types by their activity in cREs overlapping SCZ

SNPs, we observed that NCI-H226 did not cluster with lung tissue but rather in a large block consisting of various primary cells and cell lines suggesting that the enrichment is not due to lung-related factors but rather properties unique to the immortalized cell NCI-H226 (External Figure 4.1). For CD SNPs, when we filtered using H3K27ac, the top five most enriched tissues were brain (middle frontal area), neural cells, OCI-LY7 (lymphoma cell line), neural progenitor cells, and iPS-20b. When we excluded cREs on chromosome 6, OCI-LY7 dropped in ranked from third most enriched tissue to tenth and was replaced by temporal lobe tissue (External Table 4.1). When we filtered by DNase four of the top five most enriched tissues were from the brain (occipital lobe, superior temporal gyrus, middle frontal gyrus and the cerebellar cortex) with other being L1-S8R, an iPS cell line (External Table 4.2).

When we analyzed H3K27ac signal orthologous mouse cREs, we also observed enrichment in brain tissues for SCZ and CD SNPs. SCZ SNPs were enriched in midbrain, forebrain, and hindbrain regions particularly at later developmental time points (Figure 4.1b, External Table 4.3). CD SNPs were enriched primarily in forebrain. These tissues were also highly ranked when filtered cREs using DNase signal but none of the tissues met our FDR threshold of 5%.

With BPD and MDD SNPs we did not observe any enrichments that met our FDR threshold of 0.05, however, due to the small number of SNPs reported in each study we still analyzed top-ranked tissues to look for general patterns of enrichment (External Tables 4.1-4.4). In the case of MDD, there were no enrichments for human cell types ($p < 0.05$) thresholding by either H3K27ac or DNase. However, when we filtered with H3K27ac in mouse, we observed enrichments in hindbrain and neural tube tissue at time points e13.5 and e14.5 (Figure 4.1b, External Table 4.3). When we used DNase for filtering, we observed enrichment in CD-1 mesoderm tissue (External Table 4.4). For BPD SNPs, we only observed enrichment for human cell types with H3K27ac data. Four of the top five most enriched tissues were from blood (T cell subtypes and mononuclear blood cells) while the fifth was iPS cell line 20b (External Table 4.1). We repeated the analysis filtering out variants on chromosome 6 to account for SNPs in the MHC. We still observed enrichments in T helper cells and iPSCs but now observed enrichments in SK-N-MC, a neuroblast cell line, HUES64 ESCs and fetal adrenal gland (External Table 4.1). In mouse, the only tissue with $p<0.05$ for CD SNPs was bone marrow (External Table 4.3).

Overall, psych SNPs, particularly SCZ and CPD SNPs, are enriched in cREs that are active in brain tissue and neural cells in both human and mouse. We also observed enrichment of cREs in immune-related tissues such as fetal thymus for SCZ SNPs, T cells for BPD SNPs, and lymphoma cell line for CD SNPs. While these enrichments became less significant when we removed cRES on chromosome 6, they were still some of the top-ranked tissues.

SCZ and CD SNPs regulate genes expressed in the brain involved in neural pathways

Since we determined that the majority of signals for the psych GWAS can be explained by SNPs in cREs, we wanted to determine the genes regulated by these regions. We began by analyzing expression quantitative trait loci (eQTLs) generated by the Genotype-Tissue Expression (GTEx) project. In their 2015 release, the GTEx project reported over one milion eQTLs surveyed across 44 tissues, ten of which are from the brain. We intersected our psych SNPs with a list of GTEx tissue specific eQTLs curated by HaploReg[106]. LD blocks containing SCZ and CD SNPs were enriched for eQTLs with about 43% of LD blocks overlapping at least one eQTL (p=4.5E-3, p=8.5E-3, fisher's exact test) (Table 4.6). BPD and MDD LD blocks were neither enriched nor depleted.

When we analyzed the tissue specificity of these eQTLs we did not observe any enrichments with an FDR < 5% (External Table 4.5). Unlike cRE enrichments, most of the top ranked tissues for SCZ and CD genes were not from the brain. For SCZ only two of the top five tissues were from the brain (cerebellum and frontal cortex) with the others from thyroid, whole blood, and testis. For CD eQTLs, there were no brain regions in the top five tissues. BPD and MDD eQTLs did not have any tissues with p< 0.05. In order to further investigate these eQTL genes, we first looked for enrichment of gene ontology (GO terms) using PantherDB[142]. For all four eQTL genes sets, we did not observe any significant enrichments for GO terms. Using GTEx expression data we compared the expression of these eQTL linked genes to those linked with control SNPs. When we analyzed the top ten most enriched tissues, we only observed three brain regions for SCZ SNPs; the other tissues were testis, female reproductive tissues and colon (Figure 4.3, External Tables 4.6). CD eQTL genes were enriched for expression in immune tissues such as EBV lymphocytes, spleen and whole blood. MDD and BPD eQTLs genes did not have enrichments with $p < 0.05$.

These results lead us to believe that many of these eQTL links were tissue specific and LD may prevent us from identifying the direct target of the cRE. For example, rs9936474 overlaps cRE EH37E1142914 which has high H3K4me3, H3K27ac, and DNase signal in human brain tissues (External Table 4.13). While this cRE overlaps a noncoding RNA (CTD-2574D22.4), its closest protein coding gene (1.6 kb away) is *KCTD13*, a gene which encodes a potassium channel tetramerization domain protein. *KCTD13* is highly expressed in developing brain and neural cells (External Table 4.9) and has previously been linked with psychiatric disorders and brain development pathways[143,144]. If we only consider genes linked via eQTLs as potential target genes, *KCTD13* is not one of the 16 genes on the list. This example highlights two major problems with solely using eQTLs to predict target genes. First, none of the eQTL links for rs9936474 are in brain; some of the eQTLs may be tissue specific and therefore are not relevant to SCZ. If we restrict ourselves to only using eQTLs from brain tissue, however, we cover less than 20% of GWAS LD blocks. Second, of these 16 potential gene targets, only five are within 100kb of EH37E1142914. The majority of these links are likely due to indirect regulatory effects or SNPs in LD with rs9936474 rather than direct regulation by EH37E1142914. Because of these two issues, we believe using only eQTL genes for target gene prediction results in a gene list full of false positives and negatives. Therefore, we decided to curate our own lists of putative target genes for further analysis.

In chapter III, we demonstrated that 61% of high resolution GM12878 POLII ChIA-PET data links connected ELS cREs with the promoter of the closest protein coding gene; additionally, 86% of ELS-promoter links are with 100kb (Figure 4.4a). As we did not have POLII ChIA-PET in human or mouse brain tissues, we curated potential target genes based the observed ranges in GM12878. For each study, we generated two lists of genes: 1) all protein coding genes with a TSS within 100kb of the SNP, 2) the closest protein coding gene using linear distance from TSS for each SNP. These gene lists have limited overlap with eQTL genes (Figure 4.4b); in fact, approximately 41% of eQTL genes are unique.

We began by analyzing the expression of each of the genes sets comparing them to gene sets generated using control SNPs. For SCZ and CD, both the closest gene set and 100kb gene sets were enriched for expression in brain tissues surveyed by GTEx and ENCODE (Figure 4.3b,c, External Tables 4.7-4.8). Specifically, the closest genes for SCZ (N=157) had the most significant enrichment in the frontal cortex, cortex and anterior cingulate cortex. When we compared the expression of these genes using ENCODE RNA-seq data, we also observed enrichment in brain tissue, such as the parietal, occipital and temporal lobes of fetal brain (Figure 4.3d). We also analyzed the expression of orthologous mouse genes using gene expression data from the developmental time series. SCZ genes were enriched for expression in brain tissue (forebrain, midbrain, and hindbrain) particularly at later time points (Figure 4.3e). CD genes also were enriched for expression in brain tissues for both human GTEx and ENCODE samples, but these enrichments did not meet our FDR threshold of 5% (External Tables 4.7 and 4.8). We did not observe any significant (p<0.05) tissue-specific enrichments in expression for the BPD and MDD gene lists (External Tables 4.7 and 4.8). This is not surprising since we were also unable to detect strong tissue-specific enrichments in cRE activity for these disorders.

We also performed gene ontology analysis using these genes lists. For SCZ genes, we observed overwhelming enrichment in neural pathway terms (Figure 4.4f, Table 4.7, External Table 4.10) such as neuron components, synaptic transmission and gated channel activity. This suggests that these SCZ genes have primary roles in neural pathways. For the other disorders, there were not many enriched terms (External Table 4.10). For CD genes, we observed significant enrichments in immune related terms such as T-cell activation and MHC Class II receptor activity. However, these terms were no longer significant after removing genes on chromosome 6; therefore, these enrichments were only driven by SNPs in the MHC. The one enrichment that did remain significant was for genes in the Alzheimer disease-amyloid secretase pathway. For MDD, there were two genes, *IL12A* and *IL12B*, that were components of the interleukin-12 complex. Finally, there were three BPD genes involved with mannose metabolic processes.

Overall, we determined that potential target genes of SCZ SNPs have enriched expression in brain tissues and have roles in neural pathways. We also determined that solely using eQTLs to predict target genes results in enrichments for unrelated cell and tissues types possibly due to tissue specificity and LD.

Temporal activity of cREs containing psych SNPs reveals biological role

Since psych SNPs are enriched in orthologous mouse cRES active in developing brain tissues and are near genes expressed during these timepoints, we wanted to know if there was a temporal dependence on this enrichment. Using K-means clustering (K=4), we grouped cREs using H3K27ac signal across embryonic development in mouse forebrain, midbrain, and hindbrain. The resulting four clusters had nearly identical temporal patterns across the three brain subregions: cluster 1 cREs increased in activity overtime, cluster 2 cREs decreased in activity over time, cluster 3 cREs increased in activity until ~e12.5 then decrease in activity, and cluster 4 cREs increased in activity and tapered off/slightly decreased just before birth (Figure 4.5a). Comparing across the tissues, we determined that cREs tended to belong to the same clusters in all three tissues (Figure 4.5b). For each cRE group, we generated a list of linked genes by selecting the nearest protein coding gene defined using linear distance to the closest TSS. We observed the gene expression patterns of these linked genes followed the same trends as H3K27ac signal patterns across forebrain, midbrain, and hindbrain demonstrating that trends we observed in cREs hold for gene expression (Figure 4.5c). We then performed GO analysis with these gene lists and selected all enriched terms with a bonferroni corrected p-value > 0.05 (Figure 4.5d, External Table 4.11). Cluster 1 cREs were near genes involved with basic cellular processes such as translation, nuclear transport and metabolism. Cluster 2 cREs were near genes involved with embryonic development including neural crest cell differentiation, regulation of cell cycle arrest and embryonic pattern specification. Cluster 3 cREs, were enriched in terms related to CNS development such as neuron fate specification, cranial nerve development, and pallium development. Cluster 4 cREs, were enriched GTPase signaling, axon development, and neuron projection morphogenesis.

Using these clusters, we tested whether psych SNPs were enriched in cREs with a specific temporal activity pattern. To account for LD structure for both psych SNPs and controls, we randomly selected one representative cRE per LD block and averaged the results over 500 trials. In general, psych SNPs were not enriched for cREs with a specific temporal pattern. SCZ, MDD, and CD SNPs were evenly distributed across cREs clusters in the three brain regions (External Table 4.12). BPD SNPs on the other hand were enriched in cREs in cluster 2 cREs in midbrain and hindbrain (Chi-Square test p=2.83E-02, 5.47E-03). While only a small number of LD blocks overlap mouse cREs, on average 71 and 91% of the cREs are in cluster 2 in midbrain and hindbrain compared to 20 and 21% in controls.

While the majority of psych SNPs were not enriched in a particular temporal cluster, this type of analysis leads to new biological insights of how these SNPs contribute to disease. For example, CD SNP rs12424245 overlaps an ELS cRE in both human and mouse (EH37E0250841 and EM10E0283811). EH37E0250841 lies within CACNA1C, a well-documented SCZ and BPD associated gene[88-90,92]. CACNA1C encodes a calcium voltage-gated channel subunit and expressed in many tissues such as heart, muscle, and brain (External Table 4.9). In humans, EH37E0250841 has high H3K27ac signal (z-score > 1.64) in 28 cell types including adult brain, heart, and GI tissues and high DNase activity in 168 cell types including fetal brain, spinal cord, and kidney (External Table 4.13). Interestingly, the orthologous mouse cRE, EM10E0283811, only has high H3K27ac signal in neural tissues: forebrain, midbrain, hindbrain and neural tube (Figure 4.6a). In these tissues, EM10E0283811's H3K27ac z-score increases over time, plateauing just before birth; therefore, the K-means algorithm classified EM10E0283811 as a group 4 cRE in forebrain, midbrain, and hindbrain. This pattern of H3K27ac activity also correlates with CACNA1C expression these brain tissues (*r*=0.79) (Figure 4.6b). Though CACNA1C is also highly expressed in heart and lung tissue during embryonic development, EM10E0283811 is a neural specific enhancer and therefore would only control CACNA1C in developing brain tissue. Because of this temporal pattern and EH37E0250841's high H3K27ac signal in human adult brain tissue and high DNase signal in fetal brain tissue, we predict that this ELS cRE turns on during brain development and remains active throughout adulthood.

Therefore, we predict rs12424245 may contributes to the onset of psychiatric disorders by altering the regulation of CACNA1C expression.

Additionally, this temporal analysis enables us to more preciously determine target genes of SNPs. For example, rs7959408, a SCZ SNP, overlaps human cRE EH37E1112284 and mouse cRE EM10E0066315. EH37E1112284 is classified as a distal PLS cRE because it only has high H3K4me3 signal in two cell types: WERI-Rb-1 and bipolar spindle neurons. However, it does have high DNase signal in fetal brain tissue (External Table 4.13). In mice, EM10E0066315 has high DNase and H3K27ac signals across brain tissues (External Table 4.13). Because, EM10E0066315 H3K27ac z-score increases then decreases in forebrain and midbrain (Figure 4.6c), the K-means algorithm classified EM10E0066315 as a group 3 cRE; because of the steady decrease of H3K27ac signal in hindbrain, the K-means algorithm classified EM10E0066315 as a group 2 cRE. These temporal patterns in mouse embryonic brain and the complementary DNase data from human fetal brain suggests that this ELS cRE in only active during brain development. In humans, the closest protein coding gene (GENCODE V19 annotations) to EH37E1112284 is RP11-552I14.1, which has no mouse ortholog and is only expressed in the testis and prostate gland (External Table 4.9). More recent annotations of GENCODE genes (on hg38 genome) reclassify RP11-552I14.1 as a lincRNA gene. Therefore, we sought to identify the potential gene target of this ELS cRE. Using a +/- 300 kb window we analyzed expression levels of all protein coding, human-mouse orthologous genes. The two next closest protein coding genes, C12orf42 and PAH are not expressed in brain tissue or neural cells. C12orf42 overall has very low expression except in testis and PAH is almost exclusively expressed in the liver tissues (Figure 4.6d, External Table 4.9). ASCL1, a gene involved with neuronal commitment and differentiation, is highly expressed in developing brain and has correlated gene expression with EM10E0066315 H3K27ac signal (r=0.78) (Figure 4.6e, External Table 4.9). Therefore, we predict the target gene of EH37E1112284 and rs7959408 is ASCL1. This analysis demonstrates how analyzing temporal patterns of activity during embryonic development can lead to new biological insights for SCZ and CD.

Psych SNPs overlap putative binding sites for TFs involved in neural and immune pathways

Since we determined psych SNPs are in cREs active in neural and brain tissues and likely regulate genes expressed in these tissues, we wanted to determine the mechanism by which these SNPs alter gene expression. One possibility is that psych SNPs disrupt transcription factor (TF) binding sites, thus altering the regulation of target genes. To test this hypothesis, we analyzed TF binding motifs and ChIP-seq peaks that overlap psych SNPs.

We began by searching for TF sequence motifs overlapping each SNP using experimentally derived motifs from Cis-BP[145] and FIMO, a motif search software[146]. In general, SCZ, MDD, and CD SNPs were enriched for TF motifs but this enrichment was only significant for individual SNPs, not LD blocks as a whole

(External Table 4.14). Conversely, while individual BPD SNPs were not enriched for TF motifs, BPD LD blocks were with 22 of 23 LD blocks (95%) containing a SNP that overlaps a motif instance. We then tested for enrichment for specific TFs using fisher's exact test. After selecting all enrichments with FDR < 5%, we clustered TFs with similar motifs (e.g. GC rich SP family motifs) using overlap coefficients and selected one representative TF per group based on the most significant p-value (Figure 4.7).

For SCZ, MDD and CD SNPs, we observed enrichments in motifs corresponding to TFs with important roles in neural differentiation and brain development. For both SCZ SNPs and CD SNPs, SP4 was the most significantly enriched motif (Figure 4.8a). SP4 is primarily expressed in the brain (External Table 4.9) and is thought to play a role in central nervous system development[147-149]. SP4 has also been previously reported as a disease susceptibility gene in schizophrenia[150] and major depressive disorder[151,152]. In addition to SP4, SCZ SNPs were also enriched in motifs for FOXJ3, NR2F2, and LHX9 all of which are highly expressed in embryonic mouse brain tissue and human neural cells (External Table 4.9) and have roles in CNS (central nervous system) function development[153-156]. MDD SNPs were enriched for neural related transcription factors HOXA1 and TCF4. HOXA1 is involved with hindbrain and neural tube development and subsequently is highly expressed in these tissues during embryonic development (External Table 4.9). TCF4 is involved in initiating neuronal differentiation and is primarily expressed in the brain (External Table 4.9). It is also a well-established SCZ susceptibility risk gene[89,93-96,105,150,157-159].

In addition to neural related TFs, psych SNPs were enriched for motifs related to immune response and blood cell development (Figure 4.8a). SCZ SNPs were enriched for MGA motifs. MGA is highly expressed in hematopoietic stem cells (External Table 4.9) and is linked with development cell proliferation and development[160]. CD SNPs were enriched for IRF1 motifs. IRF1, encodes an interferon regulatory factor, which is primarily expressed in blood cells and immune related tissues such as the thymus (External Table 4.9). BPD SNPs were significantly enriched for the SPI1 motif, a transcription factor known to regulate blood cell development[161]. MDD SNPs were enriched in motifs for RREB1, which is involved with cell differentiation and is a negative regulator of HLA complex[162] as well as MEIS1, which regulates hematopoietic development[163].

We then compared these enrichments with TF ChIP-seq data by intersecting psych SNPs with ChIP-seq peaks from 914 ENCODE experiments. We calculated enrichment for general TF peak overlap as well as for individual TFs (N=303) and cell types (N=85) using fisher's exact test, removing SNPs on chromosome 6 due to the overwhelming enrichment of RNA POLIII machinery at the MHC. Overall, all SNPs were enriched for overlapping TF peaks but only SCZ SNPs were enriched for specific TFs and cell types. SCZ SNPs were enriched for ChIP-seq peaks for 21 TFs and 14 cells types (Table 4.8). Six of the top ten most enriched cell types were lymphoblastoid cell lines (LCLs); only one, SK-N-SH, was neural related. While this enrichment for TFs in LCLs may have biological significance, such as the enrichment Ripke *et al*. observed for SCZ SNPs in B cell H3K27ac peaks[105], it is most likely due to fact that ENCODE TF ChIP-seq experiments were performed in a different subset of cell types. For example, 22 (25%) of the cell types surveyed for TF ChIP-seq are LCLs, which is significantly higher compared to DNase (0.6%) and H3K27ac (0.7%) experiments. Additionally, unlike DNase and H3K27ac experiments, there are no ENCODE TF ChIP-seq experiments in human brain tissue.

Therefore, in order to gain a better of understanding of the role of cell type specificity with TFs, we decided to compare the expression patterns of the 21 enriched TFs against the other 282 TFs. While we did not observe an enrichment in human cell types, we observed significantly enriched expression in mouse embryonic brain tissues and blood cells (Figure 4C, Table 4.12). In mouse brain, we observed the most significant enrichments in hindbrain and midbrain at timepoints e11.5 and e13.5. In addition to brain, we observed enrichment for expression in mouse B cell and megakaryocytes (Table 4.10). This enrichment in both neural and immune cells compliments both the enrichments we observed in motifs for both neural and immune related TFs.

One possibility for the connection between neural and immune enrichments is due to the dual role of some transcription factors in immune and neural pathways. For example, SCZ SNPs are enriched 2.5-fold in POU2F2 ChIP-seq peaks in lymphoblastoid cell lines GM12878 and GM12891. POU2F2 has well documented roles in the immune system[164] but has also been identified as a regulator of neuronal differentiation[165]. In humans, *POU2F2* is highly expressed (30-40 tpm) in lymphoblastoid cells and bipolar spindle neurons (External Table 4.9) and in mouse, Pou2f2 is highly expressed (30-50 tpm) during brain development. Therefore, the dual roles of POU2F2 and other enriched TFs in both neural and immune pathways may explain the enrichment that we observe for neural and immune system factors. Overall, psych SNPs are enriched in regions bound by TFs involved in neural and immune pathways.

Psych SNPs disrupt sequence motifs resulting in the disruption of TF binding

After identifying global enrichment patterns for psych SNPs, we wanted to identify specific psych SNPs that are likely to be causal. We decided to analyze SNPs that result in allele specific binding of TFs. Using mapped reads directly from the ENCODE processing DNase pipeline, we performed *in silico* genotyping using a method adapted from Maurano *et al.*[102]. We classified a SNP as heterozygous in a given cell-type if there were at least 15 non-redundant DNase reads at the locus and a ratio of minor allele reads to major allele reads greater than 0.05. Similar to Maurano *et al*, we evaluated our method using SNPs which were genotyped using Illumina Human 1M-Duo arrays during phase 2 of the ENCODE project. Though only 4.7% of the psych SNPs had more than 15 reads, our method had perfect sensitivity (100%) and specificity (100%). To test for allelic imbalance of DNase reads at heterozygous SNPs, we used a binomial test with an FDR threshold of 5%. In total, we identified 263 allele specific SNPs: 129 for SCZ, 38 for BPD, 24 for MDD and 78 for CD (External Table 4.15). We further annotated these psych SNPs filtering for those that also overlap TF motifs (21%) (External Table 4.15). Within these lists there were several examples of allelic imbalance with interesting biological implications -- four of which we describe in detail below.

MDD SNP rs12552369 overlaps ELS cRE EH37E1025241, which lies in *AK8*, which encodes an adenylate kinase. *AK8* is expressed in human adult hepatocytes and mouse developing lung and brain and has also been reported as a susceptibility genes for ADHD[166]. EH37E1025241 has high H3K27ac signal in endodermal cells and high DNase signal in fetal brain tissues. We observed allelic imbalance for DNase reads in fetal brain tissue (day 58) with reads favoring the alternative allele A over the reference allele G (Figure 4.9a). Rs12552369 overlaps a RFX2 motif site, and the reference allele results in lower log odds score compared to the alternative allele (Figure 4.9a). We hypothesize rs12552369 disrupts RFX2 binding resulting in altered expression of AK8.

SCZ SNP rs12895055 overlaps ELS cRE EH37E0354132, which lies within *BCL11B* (Figure 4.9b). BCL11B is a transcriptional repressor that has been linked to T cell and neuronal development[167]. *BCL11B* is expressed in skin and brain tissue and its murine ortholog, Bcl11b, is highly expressed in embryonic thymus and brain tissue (External Table 4.9). EH37E0354132 has high DNase signal and H3K27ac across many different cell and tissue types (T-cells, neural cells, and fetal brain) as does its orthologous mouse cRE EM10E0105028 (External Table 4.13). We observed allelic imbalance at rs12895055 in both fetal arm muscle and eye tissues. DNase reads favored the reference allele, C (72%), over the alternative allele T (Figure 4.9b). Rs12895055 overlaps a SP4 motif, and the alternative allele results in lower log odds score compared to the reference allele (Figure 4.9b). We hypothesize rs12895055 disrupts SP4 binding therefore altering expression of BCL11B during brain development.

BPD SNP rs2861405 overlaps PLS cRE EH37E1171401, which overlaps the TSSs of two genes ZNF490 and ZNF791 (Figure 4.9c). Rs2861405 overlaps a motif for IKZF1, a tumor suppressor TF linked with lymphocyte differentitation[168]. We observed allelic imbalance at rs2861405 in kidney, blood vessel and cerebellar cortex DNase reads all favoring the reference allele C. Interestingly, we also observed allelic imbalance favoring the alternative allele A in fetal muscle. While both ZNF490 and ZNF791 are highly expressed in many tissues, they genes are not expressed equally across all cell types suggesting differing regulation between the genes. Rs2861405 overlaps a motif site for a IKZF1 motif site that slightly favors reference allele C which is reflected in the observed allelic imbalance. We also observe binding of IKZF1 in five of the eight surveyed cell types. Therefore, we hypothesize that rs2861405 disrupts the binding of IKZF1 resulting changes in expression of ZNF490 and/or ZNF791.

Finally, CD SNP rs73048919 overlaps EH37E0884523 a ELS cRE in 22 cell types and a CTCF-only cRE in 18 cell types. We detected significant allelic imbalance across rs73048919 in 14 cell types with 80% of DNase reads favoring the reference allele, C, over the alternative allele, A (Figure 4.9d). Rs73048919 overlaps CTCF ChIP-seq peaks in 81 ENCODE experiments including the brain cancer cell line SK-N-SH (Figure 4.9d). After confirming our *in silico* genotyping of rs73048919 in SK-N-SH using array genotyped SNPs in LD (rs12666575 and rs6461049), we observe almost complete imbalance of CTCF reads favoring the reference allele (Figure 4.9d). We believe this is because the minor allele of rs73048919 disrupts a CTCF motif site reducing the log odds score of matching the motif. Unlike the previous examples it is difficult to determine the biological consequence of this imbalance. CTCF-only cREs may have different biological functions such as insulators, repressors or anchors of three-dimensional chromatin loops. Since we also observed allelic imbalance of RAD21, a component of the cohesion complex, we propose that in SK-N-SH EH37E0884523 may be an anchor for chromatin loops mediate by cohesion. However, in order to truly elucidate the function of EH37E0884523 we need to perform additional experiments such as CRISPR-CAS9.

DISCUSSION

In this chapter, we used the Registry of cREs and data from the ENCODE, Roadmap and GTEx consortia to annotate noncoding variants associated with psychiatric disorders. We observed overwhelming evidence for the role of psych SNPs in neuronal development and function. We demonstrated that SCZ and CD

SNPs are enriched in cREs active in brain tissue and neuronal precursor cells and that these cREs likely target genes expressed in these tissues. While these results are not surprising, they do reaffirm the role for neural development pathways in the onset of psychiatric disorders.

The most significantly enriched TF motif for both SCZ and CD SNPs was for SP4, which we demonstrated has high expression in human neural cells and mouse developing brain. We were also able to identify a potential SP4 binding site within a ELS cRE that we believe is disrupted by the alternative allele ultimately affecting the expression of *BCL11B*. This strong enrichment for SP4 is of particular interest due to previously established link between SP4 and psychiatric disorders. SP4 has previously been reported as a disease susceptibility gene for BPD[152,169] and groups have reported altered SP4 levels in the brain of patients with SCZ and BPD[170,171] However, some of the most striking evidence for the role of SP4 in neural development is from Zhou et al who demonstrated hypomorphic SP4 mice undergo changes in behavior and memory formation analogous to symptoms of psychiatric disorders[172,173].

Additionally, we demonstrated that we can gain additional insight into the temporal dynamics of cRE activity during brain development. In order to properly study the developing human brain, we would need to collect fetal brain samples for DNase and ChIP-seq experiments at uniform time intervals. Even if we could accomplish this very difficult task there are many factors that may bias results such as gender, genetic background, and gestational conditions. During ENCODE3, production groups were about to precisely harvest brain regions at specific time points from genetically identical mice. Because of this, we are able to clearly see patterns of activity over time for cREs and identity potential therapeutic targets. For example, EM10E0066315 and ASCL1 are only active during brain development and therefore would not likely be a therapeutic target for an adult presenting with SCZ. However, EH37E0250841 and CACNA1C are active throughout adulthood and may present a better potential target. Our findings also enable us to use mouse models in the future to understand the global response of these variants. Since we demonstrated that psych SNPs are enriched in cREs that are active in mouse brain tissues, we can further investigate the role of these cREs through mouse transgenic assays and CRISPR-Cas9 experiments.

While our analysis supports a strong link between psych SNPs and neural development, our data also suggests that factors linked with the immune system have a role in the genetic risk for psychiatric disorders. We observed enrichments for cREs active in immune tissues, fetal thymus for SCZ, T-cells for BPD SNPs, and lymphoma cells for CD SNPs. We also observed enrichment of motifs for TFs with strong links to the immune system such as IRF1, RREB1, and MEIS1.

There have been several studies suggesting that dysregulation in the immune system has a role in the onset of psychiatric disorders. Smith was one of the first to propose the involvement of the immune system with his macrophage theory of depression[174] and macrophage-T-lymphocyte theory of schizophrenia[175,176]. Additionally, Schwarz and colleagues proposed the Th2-

hypothesis of schizophrenia in which certain subgroups of schizophrenia patients had altered ratio of Th1 cells vs. Th2 cells resulting in differing ratios of cell specific cytokines[177,178]. There have been numerous studies looking at the levels of cytokines in blood samples of patients with psychiatric disorders. Results from some of these studies are inconsistent with one another due to complexity of psychiatric disorders. For example, difference in results have been attributed to the psychiatric state of the patient during sample collection (e.g. manic episode, first psychotic episode, latent period)[179,180]. However, there is a general pattern of dysregulation of many different cytokines. Imbalances in cytokines could directly lead to the onset of psychiatric disorders or could be the results of the dysregulation of immune pathways that have roles in regulating the CNS. Recently, Filiano et al. demonstrated that reducing the number of meningeal T cells in mice resulted in a decrease of social behaviors analogous to autism and schizophrenia[181].

One hypothesis is that genes involved with the onset of psychiatric disorders have duels roles in both immune and neural development pathways. For example, BCL11B and POU2F2, two genes we report as potential psych risk genes, have important roles in both the immune system and neural development. Therefore, it is possible that the dysregulation of these genes during embryonic development alters brain structure and connections leaving the patient more susceptible to developing a psychiatric disorder. These genes are also dysregulated in the adult immune system, but this is not causal towards developing a psychiatric disorder.

We hope to test this hypothesis by analyzing cell type specific activity patterns and integrating results from additional GWAS.

Even though the signal for the majority of LD blocks can be explained by SNPs in cREs, we currently cannot explain the signal for 21% of the LD blocks. This could be for a variety of reasons. First, SNPs in these LD blocks may overlap cREs that we have yet to annotate. While we have surveyed over 500 biosamples, we have not covered every type of cell in the human body. Additionally, even if we have surveyed the cell type, our current method of curating cREs requires a cRE to have both high DNase signal (Z-score > 1.64) and high H3K4me, H3K27ac, or CTCF signal in at least one cell type. Some cell types such as fetal brain tissue only have DNase signal and therefore we are unable to curate cREs specific to these tissues without additional experiments. Also, psych SNPs may overlap cREs specific to a particular neural cell type such as neurons or glial cells; tissues collected from different brain regions are comprised of many different cell types so signals from a particular cell type may become diluted. Second, the signal for some of these LD blocks may be explained by SNPs that function at the level of transcriptomic regulation versus genomic regulation. These variants may disrupt RNA binding proteins sites, splicing sites or microRNA target sites, altering final protein production. In the future, we can consider these option by analyzing eCLIP and RNA-seq data to further annotate these variants.

Overall our results suggest that genetic variants associated with psychiatric disorders effect the regulation of genes involved in neural development pathways and that while components of the immune system may also play a role, the nature of their contribution remains unknown.

FIGURES

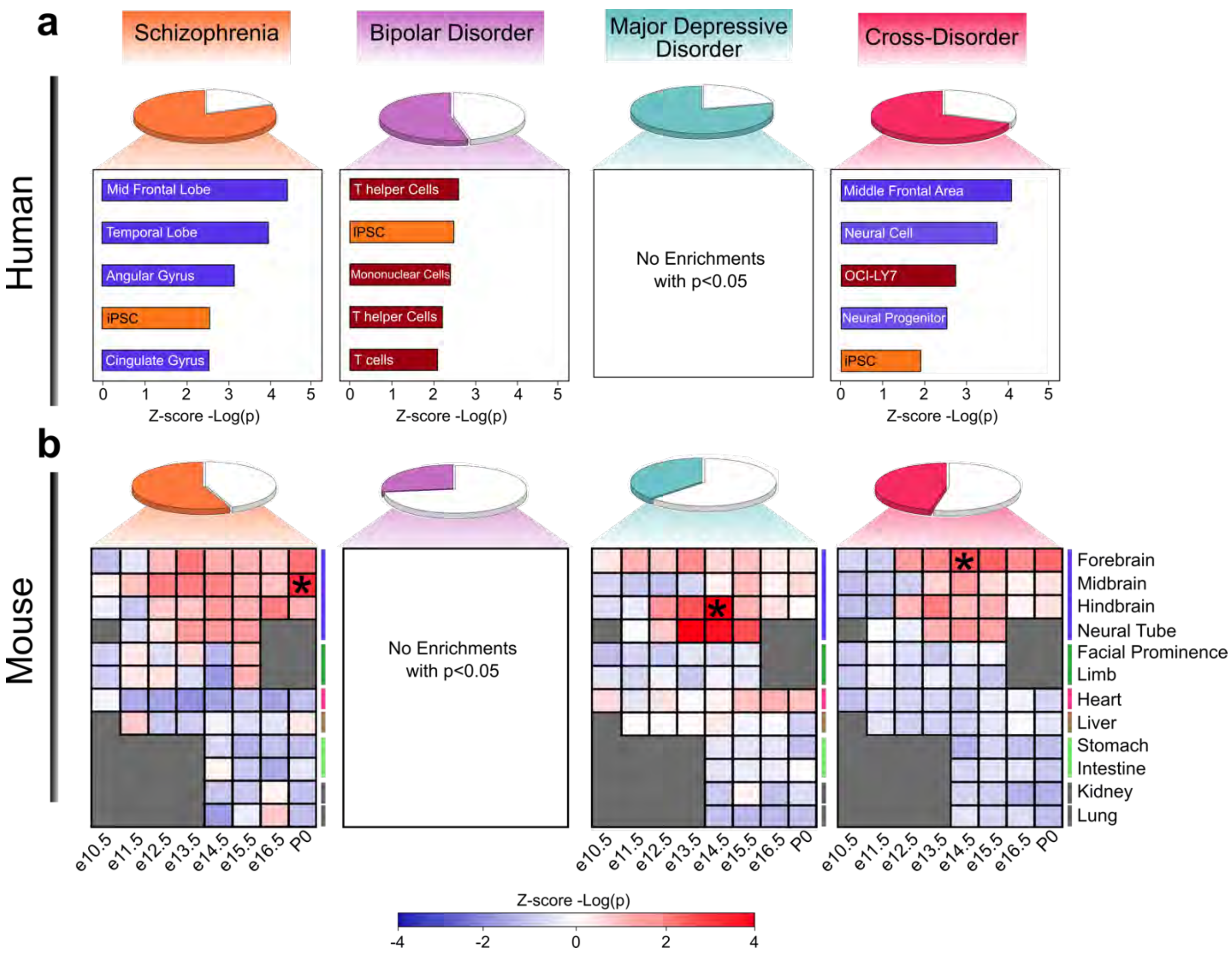
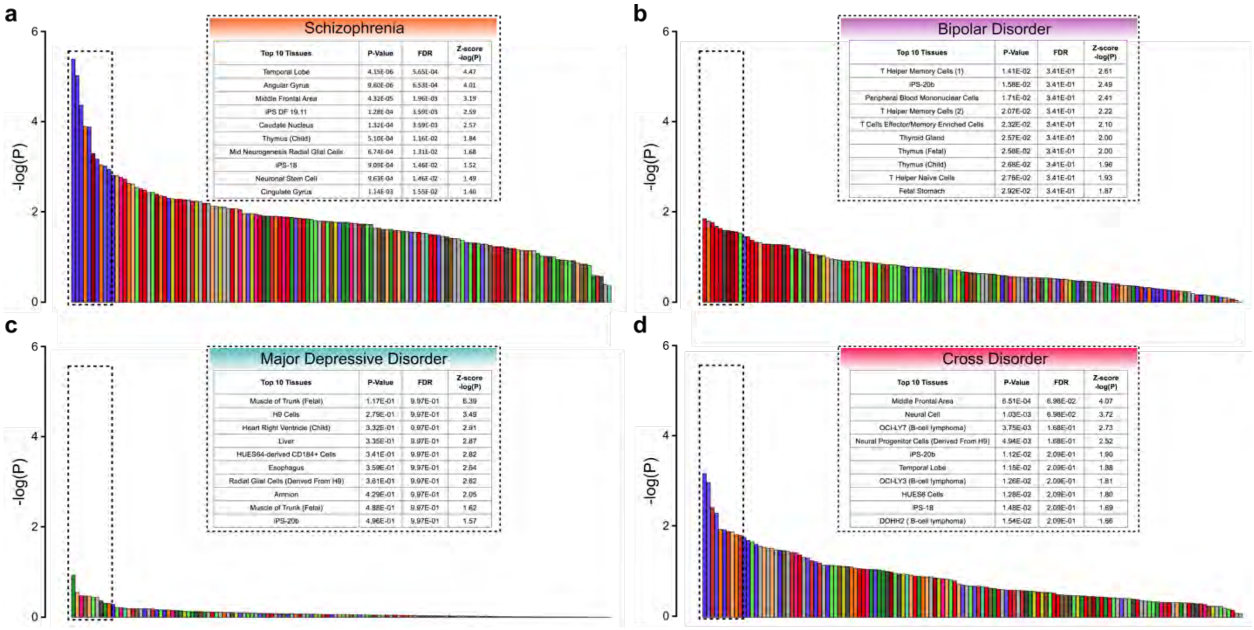


Figure 4.1 | Top tissues with enhancer-like cREs enriched for psych SNPs. Enrichment for **a,** Human cell types with cREs enriched for psych SNPs. Pie charts indicated the number of LD blocks that overlap cREs. Bars indicapte Z-score of –log(p) for each enrichment. MDD SNPs did not have any enrichments with p<0.05. **b,** Embryonic mouse tissues with orthologous cREs enriched for psych SNPs. Pie charts indicated the number of LD blocks that overlap orthologous mouse cREs. Colors indicate Z-score of –log(p) for each enrichment. BPD SNPs did not have any enrichments with p<0.05.



Schizophrenia

Top 10 Tissues	P-Value	FDR	Z-score -log(P)
Temporal Lobe	4.15E-06	5.65E-04	4.47
Angular Gyrus	9.60E-06	6.53E-04	4.01
Middle Frontal Area	4.32E-05	1.96E-03	3.19
iPS DF 19.11	1.28E-04	3.59E-03	2.59
Caudate Nucleus	1.32E-04	3.59E-03	2.57
Thymus (Child)	5.10E-04	1.16E-02	1.84
Mid Neurogenesis Radial Glial Cells	6.74E-04	1.31E-02	1.68
iPS-18	9.09E-04	1.46E-02	1.52
Neuronal Stem Cell	9.63E-04	1.46E-02	1.49
Cingulate Gyrus	1.14E-03	1.55E-02	1.40

Bipolar Disorder

Top 10 Tissues	P-Value	FDR	Z-score -log(P)
T Helper Memory Cells (1)	1.41E-02	3.41E-01	2.61
iPS-20b	1.58E-02	3.41E-01	2.49
Peripheral Blood Mononuclear Cells	1.71E-02	3.41E-01	2.41
T Helper Memory Cells (2)	2.07E-02	3.41E-01	2.22
T Cells Effector/Memory Enriched Cells	2.32E-02	3.41E-01	2.10
Thyroid Gland	2.57E-02	3.41E-01	2.00
Thymus (Fetal)	2.58E-02	3.41E-01	2.00
Thymus (Child)	2.68E-02	3.41E-01	1.96
T Helper Naive Cells	2.76E-02	3.41E-01	1.93
Fetal Stomach	2.92E-02	3.41E-01	1.87

Major Depressive Disorder

Top 10 Tissues	P-Value	FDR	Z-score -log(P)
Muscle of Trunk (Fetal)	1.17E-01	9.97E-01	5.39
H9 Cells	2.79E-01	9.97E-01	3.49
Heart Right Ventricle (Child)	3.32E-01	9.97E-01	2.91
Liver	3.35E-01	9.97E-01	2.87
HUES64-derived CD184+ Cells	3.41E-01	9.97E-01	2.82
Esophagus	3.59E-01	9.97E-01	2.64
Radial Glial Cells (Derived From H9)	3.61E-01	9.97E-01	2.62
Amnion	4.29E-01	9.97E-01	2.05
Muscle of Trunk (Fetal)	4.88E-01	9.97E-01	1.62
iPS-20b	4.96E-01	9.97E-01	1.57

Cross Disorder

Top 10 Tissues	P-Value	FDR	Z-score -log(P)
Middle Frontal Area	6.51E-04	6.98E-02	4.07
Neural Cell	1.03E-03	6.98E-02	3.72
OCI-LY7 (B-cell lymphoma)	3.75E-03	1.68E-01	2.73
Neural Progenitor Cells (Derived From H9)	4.94E-03	1.68E-01	2.52
iPS-20b	1.12E-02	2.09E-01	1.90
Temporal Lobe	1.15E-02	2.09E-01	1.88
OCI-LY3 (B-cell lymphoma)	1.26E-02	2.09E-01	1.81
HUES6 Cells	1.28E-02	2.09E-01	1.80
iPS-18	1.48E-02	2.09E-01	1.69
DOHH2 (B-cell lymphoma)	1.54E-02	2.09E-01	1.66

Figure 4.2 | Enrichment for Psych SNPs in cREs with high H3K27ac signal. Enrichment for **a,** schizophrenia, **b,** bipolar disorder, **c,** major depressive disorder, and **d,** cross disorder SNPs in cREs with H3K27ac activity. X axis indicates –log(p). Color indicates tissue of origin with purple representing brain tissue and neural cells and red indicating immune tissue and blood cells

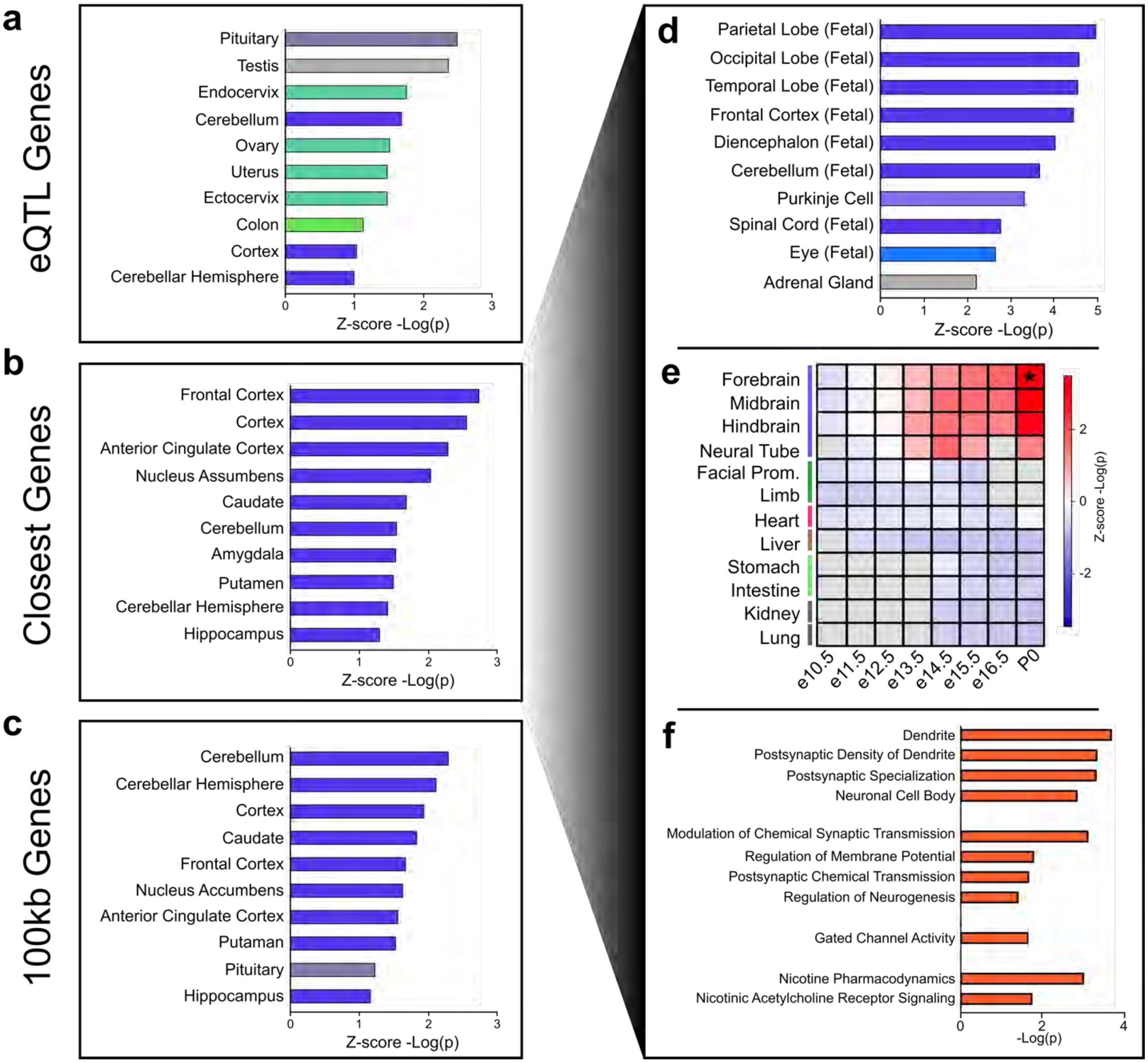


Figure 4.3 | SCZ SNP genes are enriched for brain expression and neural pathway terms. Enrichment for expression of **a,** eQTLs genes, **b,** closest genes, and **c,** genes within 100 kb for SCZ SNPs in GTEx tissues. Color indicates tissue of origin with purple representing brain tissue and neural cells. **d,** Enrichment for expression of SCZ closest genes in ENCODE **d,** human cell types and **e,** mouse embryonic tissues. **f,** Enriched gene ontology terms for SCZ closest genes.

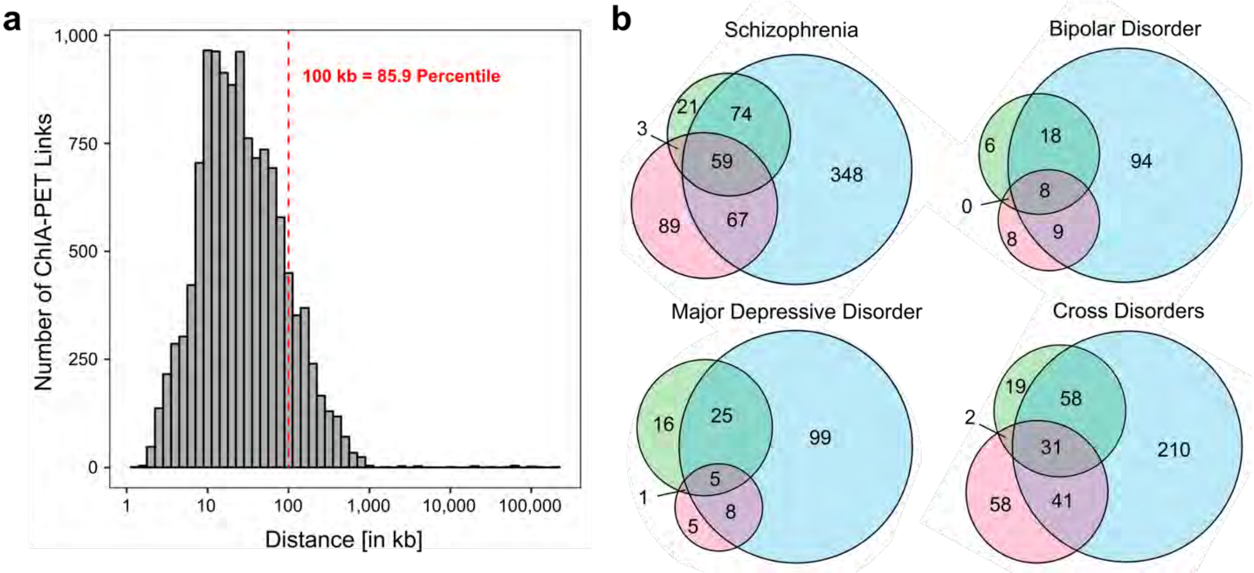


Figure 4.4 | Determining genes associated with psych SNPs. **a,** Distribution of distances between ELS-genein POLII ChIA-PET dataset. 86% of pairs occur within 100 kb. **b,** Overlap of genes called by three different methods: green = eQTLs, red=closest gene, blue= all genes within 100 kb.

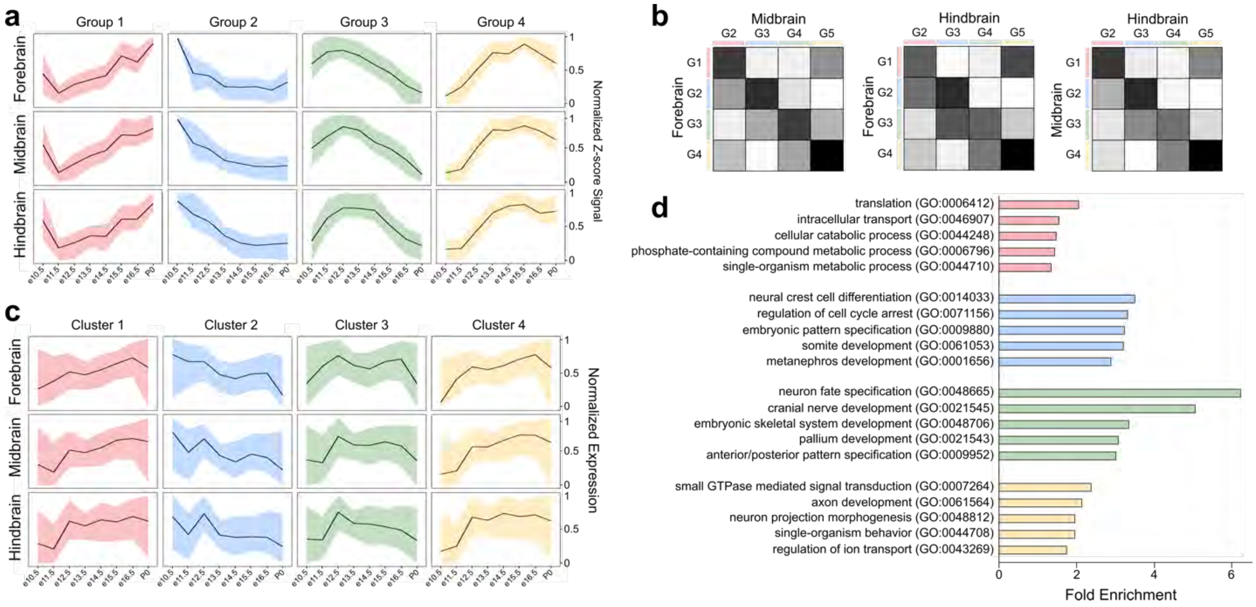


Figure 4.5 | Analysis of cRE activity across brain development. a, H3K27ac Kmeans clustering of H3K27ac signal across embryonic time points results in four clusters. **b,** Overlap of cREs in each group between brain subregions. cREs tend to be in the same activity group in all three brain subregions. Color indicates overlap coefficient. **c,** Expression of closest protein coding genes linked with cREs in each group. **d,** Enriched gene ontology terms for genes linked with cREs in each group.

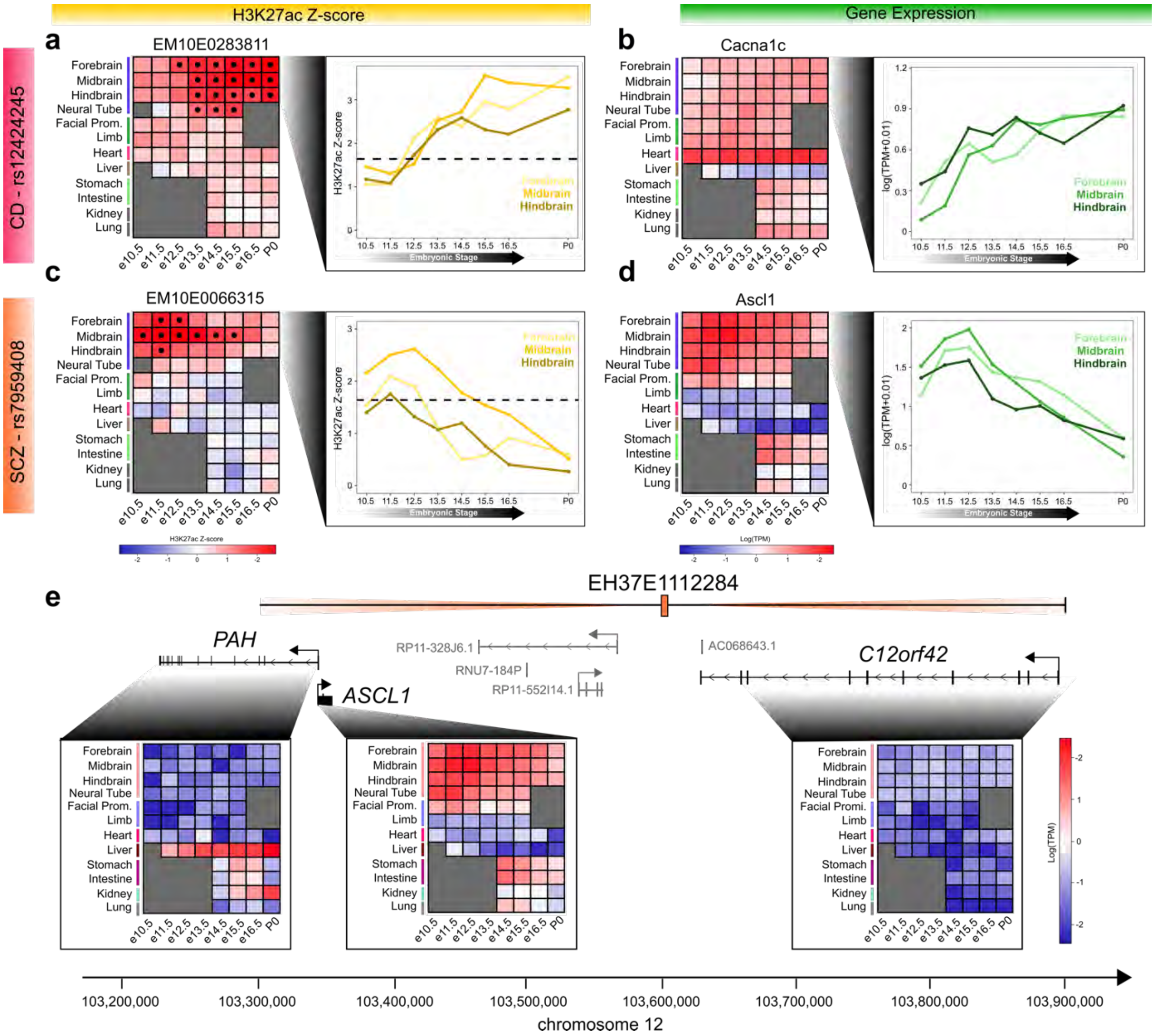


Figure 4.6 | Psych SNPs overlap orthologous mouse cREs active throughout brain development. **a,** H3K27ac Z-score signal at EM10E0283811, whose orthologous cRE overlaps CD SNP rs12424245. Color in heatmap indicates H3K27ac Z-score. **b,** Expression of *Cacna1c* across embryonic development with focus on brain subregions. Color in heatmap indicates Log(TPM). **c,** H3K27ac Z-score signal at EM10E00066315, whose orthologous cRE overlaps SCZ SNP rs7959408. Color in heatmap indicates H3K27ac Z-score. **d,** Expression of *Ascl1* across embryonic development with focus on brain subregions. Color in heatmap indicates Log(TPM). **e,** Expression of protein coding genes near EM10E00066315 across mouse embryonic development. Noncoding genes are shown in gray. Color in heatmap indicates Log(TPM).

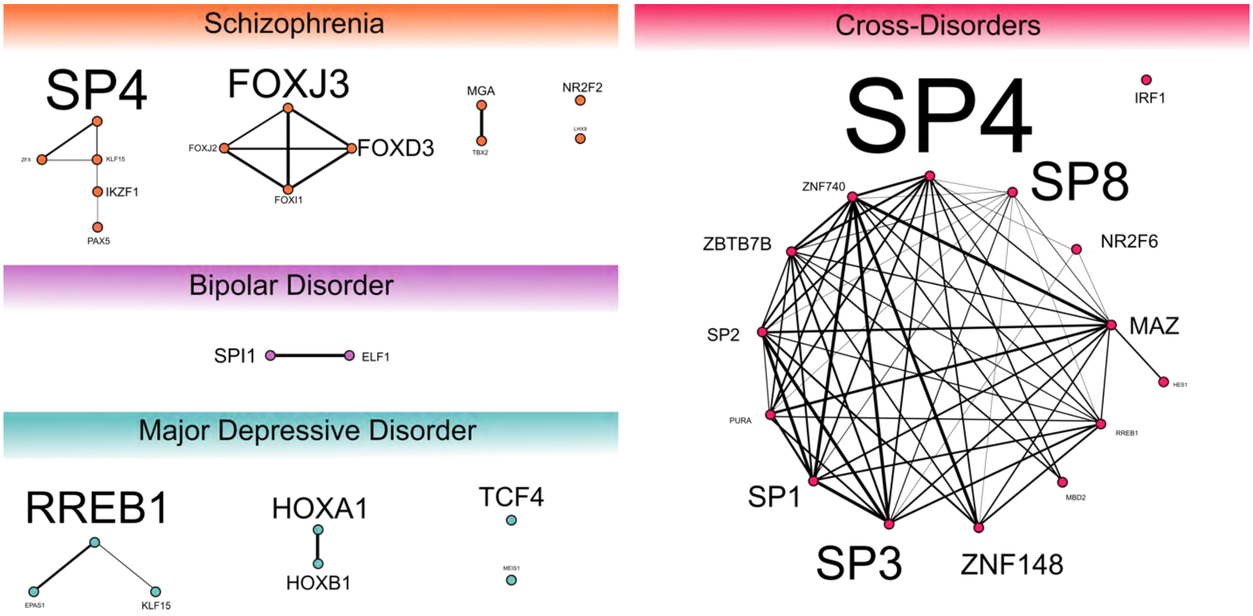


Figure 4.7 | Clustering of enriched transcription factor motifs. We connected motifs if they overlapped the same SNP with the thickness of the line indicating the number of common SNPs. Size of each motif is relative to its -log(FDR). For each cluster, we reported the most significant motif.

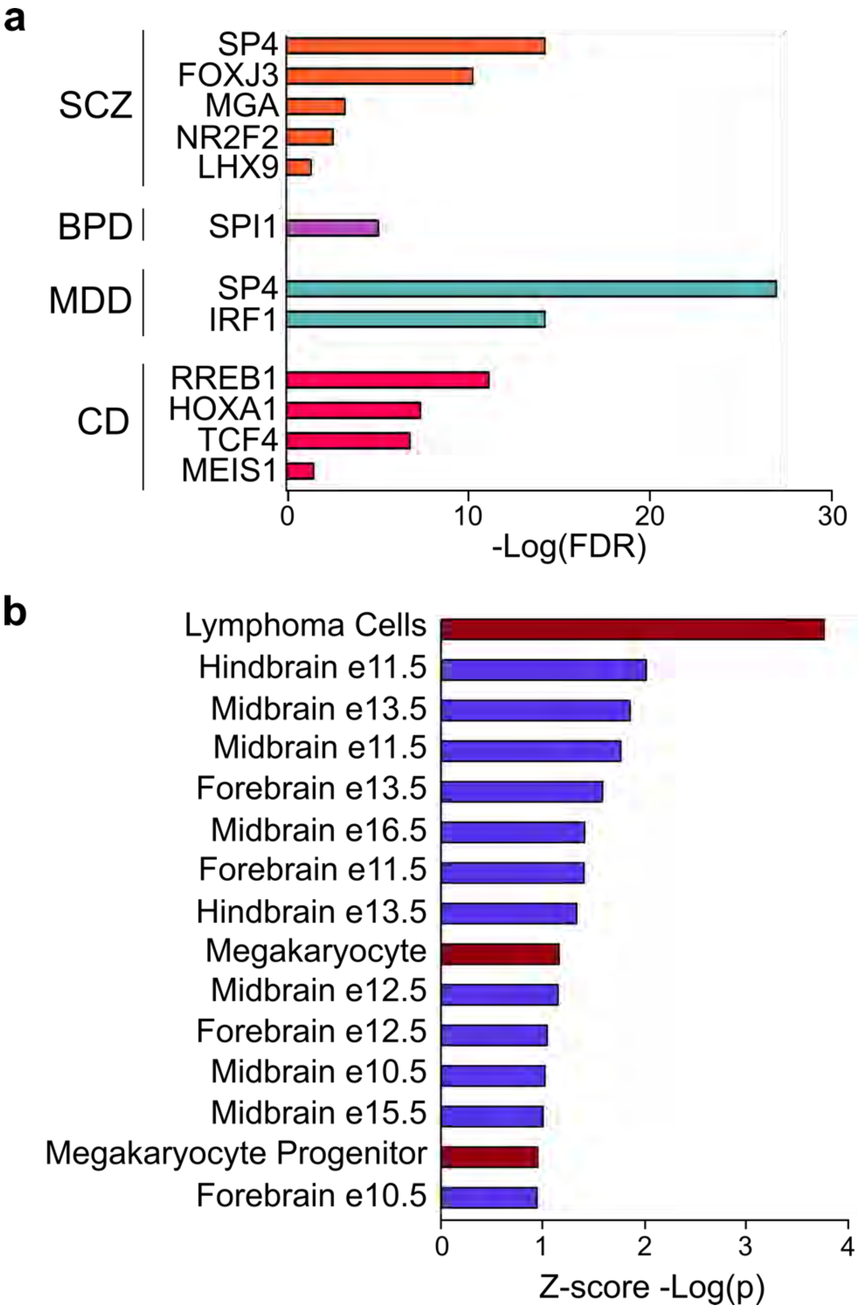


Figure 4.8 | Psych SNPs are enriched for motifs and binding sites of TFs involved in neural and immune pathways. **a,** Enriched TF motifs at psych SNPs. **b,** Enriched expression in mouse tissues and cell types for TF peaks enriched at SCZ SNPs. Colors indicate tissue of origin with purple for brain tissue and red for blood.

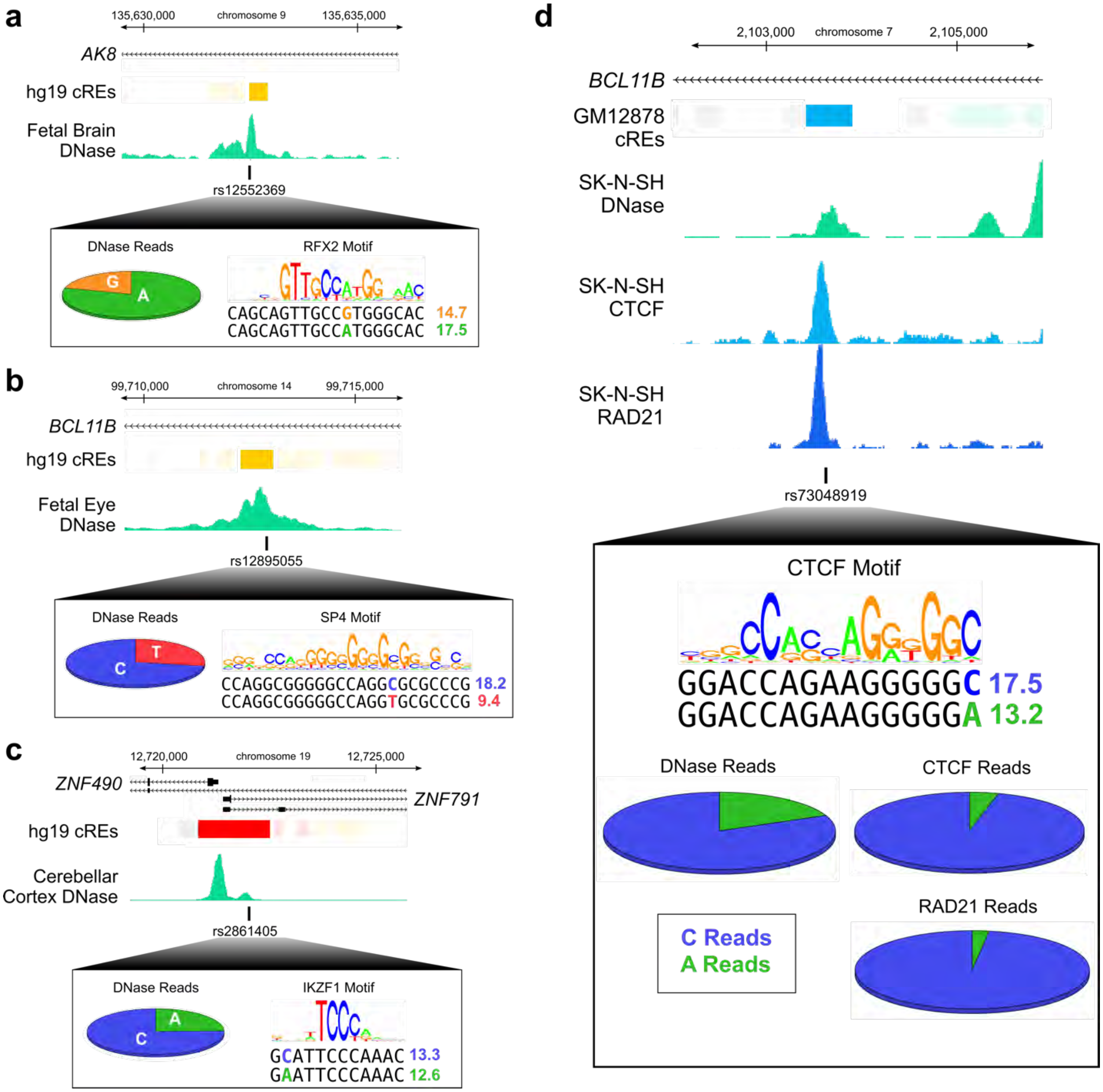


Figure 4.9 | Examples of allele specific chromatin accessibility and TF binding at psych SNPs. Allele specific chromatin accessibility at **a,** MDD SNP rs12552369, **b,** SCZ SNP rs12895055, **c,** BPD SNP rs2861405 and **d,** CD SNP rs73048919. Pie charts indicate percentage of reads with each allele. Numbers next to motif sequences indicate FIMO score for the TF motif at each sequence.

TABLES

Table 4.1 | GWAS studies included in analysis

Disorder	Authors	PMID	Publication Date	# Reported Variants
Schizophrenia (SCZ)	Schizophrenia Working Group of the Psychiatric Genomics Consortium	25056061	July 2014	98
Bipolar Disorder (BPD)	Jiang and Zhang	21254220	Feb 2011	24
Major Depressive Disorder (MDD)	GENDEP	23377640	Feb 2013	49
Cross-disorders	Cross-Disorder Group of the Psychiatric Genomics Consortium.	23453885	April 2013	74

Table 4.2 | Genetic context of psych SNPs

	Psychiatric Disorders							
	Schizophrenia		**Bipolar Disorder**		**Major Depressive Disorder**		**Cross-Disorder**	
	# SNPs	**% SNPs**	**# SNPs**	**% SNPs**	**# SNPs**	**% SNPs**	**# SNPs**	**% SNPs**
Coding Exon	38	1.11%	3	0.41%	1	0.13%	24	1.35%
UTR	88	2.57%	18	2.45%	7	0.90%	34	1.91%
Intron	2,226	65.13%	450	61.22%	268	34.49%	1,274	71.41%
Intergenic	1,066	31.19%	264	35.92%	501	64.48%	452	25.34%
Proximal	505	14.77%	76	10.34%	50	6.44%	285	15.98%
Distal	2,913	85.23%	659	89.66%	727	93.56%	1,499	84.02%

Table 4.3 | Deleterious predictions for psych SNPs overlapping coding exons

Study	SNP	Protein	Reference AA	Alternative AA	PROVEAN Prediction	SIFT Prediction
SCZ	rs4584886	ENSP00000326870	R	W	Deleterious	Damaging
SCZ	rs2955365	ENSP00000205890	A	T	Neutral	Damaging
SCZ	rs2955367	ENSP00000205890	W	G	Neutral	Damaging
SCZ	rs12596883	ENSP00000457441	E	D	Neutral	Tolerated
SCZ	rs13107325	ENSP00000349174	A	T	Neutral	Tolerated
SCZ	rs20551	ENSP00000263253	I	V	Neutral	Tolerated
SCZ	rs3176443	ENSP00000354481	L	V	Neutral	Tolerated
SCZ	rs3617	ENSP00000415769	Q	K	Neutral	Tolerated
SCZ	rs950169	ENSP00000286744	T	I	Neutral	Tolerated
SCZ	rs10117	ENSP00000297185	L	L	Neutral	Tolerated
SCZ	rs10414643	ENSP00000394510	L	L	Neutral	Tolerated
SCZ	rs1047361	ENSP00000384899	S	S	Neutral	Tolerated
SCZ	rs1051431	ENSP00000445859	Y	Y	Neutral	Tolerated
SCZ	rs1143702	ENSP00000353030	Y	Y	Neutral	Tolerated
SCZ	rs13189822	ENSP00000261483	Q	Q	Neutral	Tolerated
SCZ	rs2074090	ENSP00000262815	S	S	Neutral	Tolerated
SCZ	rs216193	ENSP00000263073	A	A	Neutral	Tolerated
SCZ	rs2229193	ENSP00000332549	L	L	Neutral	Tolerated
SCZ	rs2274267	ENSP00000439065	T	T	Neutral	Tolerated
SCZ	rs2955355	ENSP00000268719	G	G	Neutral	Tolerated
SCZ	rs2955366	ENSP00000205890	P	P	Neutral	Tolerated
SCZ	rs3743739	ENSP00000219345	G	G	Neutral	Tolerated
SCZ	rs3745474	ENSP00000246794	F	F	Neutral	Tolerated
SCZ	rs3745475	ENSP00000394510	P	P	Neutral	Tolerated
SCZ	rs4368210	ENSP00000326870	L	L	Neutral	Tolerated
SCZ	rs4685	ENSP00000335321	V	V	Neutral	Tolerated
SCZ	rs5629	ENSP00000244043	R	R	Neutral	Tolerated
SCZ	rs6163	ENSP00000358903	S	S	Neutral	Tolerated
SCZ	rs62021888	ENSP00000260402	R	R	Neutral	Tolerated

SCZ	rs7148456	ENSP00000338814	P	P	Neutral	Tolerated
SCZ	rs749240	ENSP00000263073	Q	Q	Neutral	Tolerated
SCZ	rs769267	ENSP00000262815	P	P	Neutral	Tolerated
SCZ	rs788018	ENSP00000335321	G	G	Neutral	Tolerated
SCZ	rs788023	ENSP00000335321	K	K	Neutral	Tolerated
SCZ	rs8539	ENSP00000340019	K	K	Neutral	Tolerated
SCZ	rs9611519	ENSP00000216237	P	P	Neutral	Tolerated
SCZ	rs9806806	ENSP00000332549	R	R	Neutral	Tolerated
SCZ	rs4072738	record not found				
BPD	rs10458896	ENSP00000263181	I	V	Neutral	Tolerated
BPD	rs2297815	ENSP00000354623	V	V	Neutral	Tolerated
BPD	rs4804725	record not found				
MDD	rs4777035	ENSP00000403392	P	L	Neutral	Damaging
CD	rs678	ENSP00000273283	E	V	Deleterious	Damaging
CD	rs214967	ENSP00000341887	S	L	Neutral	Damaging
CD	rs3132580	ENSP00000417182	E	K	Neutral	NA
CD	rs1042779	ENSP00000273283	Q	R	Neutral	Tolerated
CD	rs41273537	ENSP00000358064	M	V	Neutral	Tolerated
CD	rs3094086	ENSP00000417182	S	S	Neutral	NA
CD	rs1058766	ENSP00000338629	R	R	Neutral	Tolerated
CD	rs11121172	ENSP00000338629	R	R	Neutral	Tolerated
CD	rs13596	ENSP00000338629	P	P	Neutral	Tolerated
CD	rs2071702	ENSP00000436786	N	N	Neutral	Tolerated
CD	rs2229193	ENSP00000332549	L	L	Neutral	Tolerated
CD	rs2230534	ENSP00000233027	P	P	Neutral	Tolerated
CD	rs2230535	ENSP00000233027	L	L	Neutral	Tolerated
CD	rs2275271	ENSP00000402831	S	S	Neutral	Tolerated
CD	rs2523721	ENSP00000391879	R	R	Neutral	Tolerated
CD	rs3740387	ENSP00000339479	D	D	Neutral	Tolerated
CD	rs6951493	ENSP00000265854	H	H	Neutral	Tolerated
CD	rs7107305	ENSP00000436786	L	L	Neutral	Tolerated

CD	rs72696841	ENSP00000358064	S	S	Neutral	Tolerated
CD	rs748002	ENSP00000264051	A	A	Neutral	Tolerated
CD	rs9324	ENSP00000273283	S	S	Neutral	Tolerated
CD	rs9332801	ENSP00000436786	I	I	Neutral	Tolerated
CD	rs943037	ENSP00000402831	A	A	Neutral	Tolerated
CD	rs9806806	ENSP00000332549	R	R	Neutral	Tolerated

Table 4.4a | Overlap of psych SNPs with cREs

	GWAS			Control				
Study	**Total # SNPs**	**# Overlap cREs**	**Percent**	**Total # SNPs**	**# Overlap cREs**	**Percent**	**Enrichment**	**P-value**
SCZ	3418	687	20.10%	1635798	298134	18.23%	1.10	5.19E-03
BPD	735	150	20.41%	371792	67310	18.10%	1.13	1.13E-01
MDD	777	159	20.46%	715619	114342	15.98%	1.28	1.01E-03
CD	1784	409	22.93%	1198932	211708	17.66%	1.30	1.76E-08

Table 4.4b | Overlap of LD Blocks with cREs

	GWAS			Control				
Internal ID	**Total # LD Blocks**	**# Overlap cREs**	**Percent**	**Total # LD Blocks**	**# Overlap cREs**	**Percent**	**Enrichment**	**P-value**
SCZ	96	78	81.25%	47408	37518	79.14%	1.03	7.06E-01
BPD	23	20	86.96%	11221	8901	79.32%	1.10	4.51E-01
MDD	43	34	79.07%	21276	16002	75.21%	1.05	7.24E-01
CD	71	49	69.01%	36298	27849	76.72%	0.90	1.24E-01

Table 4.5 | Type of overlapping cREs

	PLS cREs		ELS cREs		CTCF-only cREs	
Disorder	**#**	**%**	**#**	**%**	**#**	**%**
Schizophrenia	150	26.83%	387	69.23%	22	3.94%
Bipolar Disorder	21	19.63%	78	72.90%	8	7.48%
Major Depressive Disorder	15	12.82%	98	83.76%	4	3.42%
Cross-Disorder	68	20.86%	253	77.61%	5	1.53%

Table 4.6 | Overlap of eQTLs and psych SNPs

	GWAS		Control		
Disorder	**# eQTL SNPs**	**% Total SNPs**	**# eQTL SNPs**	**% Total SNPs**	**P-value**
SCZ	2171	63.52%	757632	46.32%	1.87E-90
BPD	189	25.71%	142694	38.38%	5.24E-13
MDD	217	27.93%	223867	31.28%	4.41E-02
CD	868	48.65%	495409	41.32%	4.34E-10

Disorder	**# eQTL LD Blocks**	**% Total LD Blocks**	**# eQTL LD Blocks**	**% Total LD Blocks**	**P-value**
SCZ	52	54.17%	18707	39.46%	4.52E-03
BPD	10	43.48%	3641	32.45%	2.70E-01
MDD	11	25.58%	5600	26.32%	1.00E+00
CD	35	49.30%	12403	34.17%	8.52E-03

Table 4.7 | Enriched gene ontology terms for SCZ closest genes

Category	Term	Fold Enrichment	P-Value
GO Cellular Component	postsynapse (GO:0098794)	5.81	3.82E-06
GO Cellular Component	neuron projection (GO:0043005)	3.44	1.43E-05
GO Cellular Component	somatodendritic compartment (GO:0036477)	4.16	5.46E-05
GO Cellular Component	neuron part (GO:0097458)	3.01	7.47E-05
GO Cellular Component	dendrite (GO:0030425)	4.71	2.13E-04
GO Cellular Component	postsynaptic density of dendrite (GO:0014069)	7.45	4.79E-04
GO Cellular Component	postsynaptic specialization (GO:0099572)	7.42	5.02E-04
GO Cellular Component	asymmetric synapse (GO:0032279)	7.31	5.80E-04
GO Cellular Component	neuron to neuron synapse (GO:0098984)	7.24	6.38E-04
GO Biological Process	modulation of chemical synaptic transmission (GO:0050804)	6.14	7.96E-04
PANTHER Pathways	Nicotine pharmacodynamics pathway (P06587)	19.93	1.02E-03
GO Cellular Component	synapse part (GO:0044456)	3.73	1.29E-03
GO Cellular Component	synapse (GO:0045202)	3.42	1.30E-03
GO Cellular Component	neuronal cell body (GO:0043025)	4.63	1.47E-03
GO Cellular Component	cell projection (GO:0042995)	2.36	3.80E-03
GO Cellular Component	plasma membrane bounded cell projection (GO:0120025)	2.37	5.10E-03
GO Cellular Component	cell body (GO:0044297)	4.06	7.20E-03
GO Cellular Component	postsynaptic membrane (GO:0045211)	6.08	9.84E-03
GO Biological Process	regulation of membrane potential (GO:0042391)	4.73	1.71E-02
PANTHER Pathways	Nicotinic acetylcholine receptor signaling pathway (P00044)	8.05	1.88E-02
GO Cellular Component	synaptic membrane (GO:0097060)	5.05	1.90E-02
GO Biological Process	chemical synaptic transmission, postsynaptic (GO:0099565)	11.86	2.27E-02
GO Molecular Function	gated channel activity (GO:0022836)	4.91	2.37E-02
GO Biological Process	biological regulation (GO:0065007)	1.31	3.45E-02
GO Biological Process	regulation of neurogenesis (GO:0050767)	3.49	4.13E-02

Table 4.8 | Enriched TF ChIP-seq peaks overlapping SCZ SNPs

TF	Fold Enrichment	P-value	FDR
BCLAF1	3.6255	2.38E-04	2.85E-02
TAF1	2.1440	3.06E-04	2.85E-02
TAF7	3.2253	3.53E-04	2.85E-02
ZNF263	2.2811	3.76E-04	2.85E-02
POU2F2	2.5088	5.81E-04	3.52E-02
CCNT2	2.5539	7.34E-04	3.71E-02
USF1	2.0150	9.89E-04	3.78E-02
PHF8	2.3665	1.08E-03	3.78E-02
SREBF1	3.0027	1.12E-03	3.78E-02
SAP30	2.8971	1.49E-03	4.40E-02
CEBPD	3.2675	1.83E-03	4.40E-02
eGFP-ATF1	1.9465	2.03E-03	4.40E-02
NBN	2.7484	2.25E-03	4.40E-02
RFX5	2.2771	2.28E-03	4.40E-02
ZNF207	2.8709	2.63E-03	4.40E-02
eGFP-ZBTB11	2.1681	2.64E-03	4.40E-02
FLAG-SSRP1	2.8535	2.74E-03	4.40E-02
eGFP-ELF1	2.3168	2.76E-03	4.40E-02
SIN3A	1.7325	2.85E-03	4.40E-02
TBP	1.8900	2.90E-03	4.40E-02
TARDBP	2.1932	3.26E-03	4.70E-02

Table 4.9 | Enriched cell types with TF ChIP-seq peaks overlapping SCZ SNPs

TF	Fold Enrichment	P-value	FDR
GM15510	2.97	1.51E-04	9.89E-03
GM10847	3.33	2.68E-04	9.89E-03
HEK293	1.82	4.02E-04	9.89E-03
GM19099	2.64	5.27E-04	9.89E-03
GM18526	2.88	5.82E-04	9.89E-03
SK-N-SH	1.59	1.81E-03	2.17E-02
PFSK-1	2.14	2.23E-03	2.17E-02
GM19193	2.28	2.28E-03	2.17E-02
H1-hESC	1.44	2.43E-03	2.17E-02
GM18505	2.34	2.55E-03	2.17E-02
Panc1	2.07	3.03E-03	2.31E-02
Raji	2.37	3.26E-03	2.31E-02
GM18951	2.07	4.13E-03	2.69E-02
Ishikawa	1.79	4.42E-03	2.69E-02

Table 4.10 | Top 10 mouse RNA-seq experiments with enriched expression for TF

Tissue/Cell Types	Enriched Expression	Background Expression	~ Fold Enrichment	P-value	FDR
B10.H-2aH-4bp/Wts CH12.LX	39.86	13.87	2.87	1.18E-03	3.78E-02
C57BL/6 hindbrain embryo (11.5 days)	31.75	18.65	1.70	4.60E-03	3.78E-02
C57BL/6 midbrain embryo (13.5 days)	39.86	18.26	2.18	5.22E-03	3.78E-02
C57BL/6 midbrain embryo (11.5 days)	35.90	19.29	1.86	5.60E-03	3.78E-02
C57BL/6 forebrain embryo (13.5 days)	38.17	20.49	1.86	6.45E-03	3.78E-02
C57BL/6 midbrain embryo (16.5 days)	24.05	15.55	1.55	7.40E-03	3.78E-02
C57BL/6 forebrain embryo (11.5 days)	46.56	22.09	2.11	7.43E-03	3.78E-02
C57BL/6 hindbrain embryo (13.5 days)	30.95	17.58	1.76	7.85E-03	3.78E-02
C57BL/6 megakaryocyte male adult (5-6 weeks)	23.19	13.56	1.71	8.99E-03	3.78E-02
C57BL/6 midbrain embryo (12.5 days)	41.95	25.41	1.65	9.06E-03	3.78E-02

EXTERNAL FIGURES AND TABLES

These figures and tables are too large to include in the PDF version of this thesis. They are available online at https://drive.google.com/drive/folders/0B07orkTYRj9pRy1IdE9JUVVYTzA?usp=sharing

METHODS

Generating Control Datasets

For each lead SNP from the NHGRI GWAS catalog[182,183], we selected a random SNP from the SNP chip used in the study that fell in the same minor allele frequency (MAF) quartile and distance from transcription start site (TSS) quartile as the lead SNP. For each of the control SNPs, we extracted all SNPs in LD (r^2>0.7) based on phasing analysis from the 1000 Genomes Project European population (EUR)[113] using the HaploReg Database[106]. This constituted one control dataset. We generated 500 control datasets for each GWAS. This method was adapted from the Understanding Enrichment Through Simulation (UES) algorithm from the Klein lab[101].

Testing for Enrichment in cRE Activity

Using Bedtools (v25.5.0) we intersected psych and control SNPs with cREs from the Registry of cREs. To assess whether the cREs in a cell type were enriched in psych SNPs, we select all cREs overlapping psych or control SNPs with cREs that have a H3K27ac/DNase Z-score > 2 in the cell type. To avoid over counting, we pruned the overlaps, counting each LD group once per cell type. We calculated enrichment between GWAS LD groups with the 500 matched controls using a one-sided Fisher's exact test. Finally, we applied an FDR of 5% to each study. To compare relative enrichments across studies, we calculated Z-scores of the -log(p-value).

Enrichment for Gene Expression

For each cell type, we tested whether the psych genes group had higher expression than the control group using a one-sided Wilcoxon-Rank Sum test. We used gene expression values generated by the GTEx consortium[108] and the ENCODE project. We also considered mouse datasets for which we compared the expression of genes with ortholgous mouse genes as defined by the Jackson laboratory (HOM_MouseHumanSequence.rpt)

Gene Ontology Analysis

We performed gene ontology analysis using the online tool Panther's statistical overrepresentation test (http://pantherdb.org/). For all analyses, we reported enrichments from five collections: GO Molecular Function, GO Biological Process, GO Cellular Component, Reactome Pathways, and Panther Pathways. We only report categories with a Bonferroni p-value < 0.05.

Temporal Clustering of Brain cREs

We selected all mouse cREs that had a DNase Z-score > 1.64 and H3K27ac Z-score > 1.64 for at least one brain region time point. For each of these brain cREs we extracted their Z-score signal for forebrain, midbrain, and hindbrain. For each cRE in each tissue, we normalized the range of Z-scores so that all values fell between 0 and 1. Using the elbow method, we determined the optimal number of clusters was four. We then implemented K-means clustering for each tissue using the python scikit learn package.

Determining Overlapping Motifs

For each psych SNP, we generated major and minor allele sequences for 41 bps window centered on the SNP. Using the FIMO algorithm, we searched for motifs from the CISBP transcription factor database[145]. We selected all motifs that overlapped a psych SNP with a q-value < 0.05. We tested for motif enrichment by comparing the fraction of psych SNPs overlapping each motif compared to the control SNPs using one-sided fisher's exact test and applying an FDR of 5%.

Enrichment for TF Binding Sites

We intersected psych and control SNPs with optimal IDR ChIP-seq peaks from 914 ENCODE experiments. To assess whether psych SNPs we enriched for a TF binding sites, we used a one-sided Fisher's exact test comparing the number psych SNPs overlapping peaks vs control SNPs after pruning for LD. We ran three tests: 1) analyzing TFs 2) analyzing cell types, and 3) analyzing individual experiments. In all three cases we applied an FDR of 5%.

Testing for Allele Specificity

To test for allele specificity, began by first identifying heterozygous loci. For each psych SNP, we determined for which cell types it overlapped an active cRE (DNase Z-score). For these cell types, we downloaded mapped reads DNase reads (bam files) and determined the allele at the SNP. We considered a locus heterozygous if it has at least 15 reads at the ratio of the alternative allele to reference allele was at least 0.05. If a SNP was heterozygous, we tested for allele specificity in that cell type using a two-sided Fisher's exact test. We evaluated our *in silico* genotyping method using genotyping results generated by the ENCODE consortium: wgEncodeHaibGenotypeBalleleSnp2015-03-04.tsv

Scripts

Scripts for this analysis can be found on GitHub: https://github.com/Jill-Moore/Dissertation/tree/master/Chapter-IV/

CHAPTER V: Conclusions and Future Directions

Introduction

In this thesis, we described the Registry of candidate Regulatory Elements, a collected of putative regulatory regions we curated across the human and mouse genomes. We demonstrated several biological applications of cREs, in particular their use in annotating variants reported by genome wide association studies (GWAS). In Chapter III, we evaluated methods for linking cREs with potential target genes. We developed a benchmark of ELS-gene links which we used to test correlation methods and Random Forest models. In Chapter IV, we demonstrated the usefulness of the Registry by annotating genetic variants associated with psychiatric disorders. Our results suggest that in GWAS variants disrupt neural pathways and may also play a role in the immune system.

The Registry of cREs and Future Plans for Expansion

Chapter II of this thesis detailed the creation and implementation of the Registry of cREs. We curated this registry by integrating hundreds of DNase-seq and ChIP-seq datasets, generating over 1.3 million cREs in human and 400 thousand cREs in mouse. We classified these cREs into groups using H3K4me3, H3K27ac, and CTCF ChIP-seq signals and extended these classification schemes to generate cell type specific annotations for every cRE. We based these classification schemes on unsupervised enhancer prediction methods that we developed using embryonic mouse data. In this analysis, we determined that combining DNase and H3K27ac data was the best performing method and we used this method to predict enhancers which were validated using transgenic mouse assays. Subsequent analyses of our classification schemes demonstrated they are concordant with other enhancer prediction methods, particularly those that also integrate epigenomic datasets. We further demonstrated the utility of the registry of cREs by annotating variants reported by GWAS, particularly in Chapter IV where we focused on SNPs associated with psychiatric disorders.

Unlike other collections of regulatory elements, The Registry of cREs has several unique features which make it useful for biological research. First, we accessioned all cREs so they can be accurately referenced in presentations and publications. Other tools generate run-specific identifiers for their called regions, but these are not maintained during subsequent uses of the programs. Additionally, the Roadmap Epigenomics Consortium generated genome segmentations for 3 different ChromHMM models (15 state, 18 state and 25 state models). Determining the exact elements a paper used may be difficult if the authors do not specifically mention which set of segmentations they used. Second, boundaries of cREs are fixed across cell types, which allows us to evaluate the cell type specificity of a cRE. While not all elements retain the same boundaries between cell types, we have observed that the vast majority do within 50 bp. In the future, we can further annotate cREs with estimated boundaries in each cell type using overlapping DHSs. Third, to annotate and investigate cREs, users can download the Registry from the ENCODE portal or use our web based tool

SCREEN. Unlike other web-based catalogs of regulatory elements, SCREEN is user friendly, especially for biologists without a computational background. Additionally, by integrating other data from the ENCODE Encyclopedia and the NHGRI-EBI GWAS catalog, users can easily characterize regions of interest. The Registry of cREs also bridges the gap between mouse and human for comparing gene regulation. Users can investigate cREs that are ortholgous between the species, which ultimately allows them to survey different types of data across new cell types.

Moving forward, we plan on expanding the Registry of cREs. During Phase IV of the ENCODE project, production labs will generate new DNase-seq and histone modification ChIP-seq datasets which we will incorporate into the registry. We are collaborating with data production labs to expand the Registry in two ways. One to generate datasets that will increase the number of cell types with four core epigenomic marks (DNase, H3K4me3, H3K27ac, and CTCF). We currently have 21 cell types with all four marks and by increasing this number we can further analyze cREs for tissue specificity and different regulatory roles across cell types. Two, generate datasets from underrepresented cell types that are not currently covered by the Registry. Using publicly available data processed by CISTROME, we plan on identifying cell types with a low numbers of overlapping histone modification peaks or DHSs and will prioritize these cell type for data generation.

We also aim to identify new classes of regulatory elements such as repressors. While labs have experimentally characterized repressive elements, there are currently no methods for computationally predicting them. One possible direction is to study REST ChIP-seq binding sites. As mentioned in Chapter I, REST is a repressive transcription factor that binds at RE-1 elements and prevents the transcription of neuronal genes in non-neuronal cell types[20] . The ENCODE consortium has generated REST ChIP-seq data for over 25 cell types. Using the Registry of cRE we can determine if REST binding sites overlap 1) our current set of cREs, or 2) rDHSs that are not currently classified as cREs. If the former, this suggests that: a) repressors are also enriched for H3K27ac signal, b) repressors act as enhancers in other cell types, or c) the majority of REST binds to enhancers. If the latter, we can determine which features (TF signal, histone modification signal, sequence motifs) are enriched at rDHSs with REST binding and using these features, identify cREs that are likely to silence gene expression. In addition to identifying new classes of cREs, we hope to develop a more sophisticated classification scheme for cREs. We currently categorize cREs in generalized groups, which make interpretation easy but may miss finer biological features. For example, we currently classify all proximal cREs with high H3K4me3 signal as PLS. However, we have observed that PLS cRE that directly overlap TSS have different features than those that do not. Therefore, we might be able to split these cREs into two distinctive classes. In cell types such as GM12878, K562 and H1-hESCs, we hope to integrate TF data and cluster cREs based on TF binding to observe if there are any natural classes of elements. These analyses may reveal sub-types of regulatory elements.

Evaluating Methods Prediction Enhancer-Gene Links

In order to accurately evaluate enhancer-gene linking methods we developed a benchmark of ELS-gene pairs base on chromatin and genetic interaction data. We used this benchmark to test correlation base methods and found they had low overall performance. While these methods can identify ELS-gene pairs, they have extremely high false negative and positive rates. These results suggest that previous work by ENCODE labs may not identify the correct target gene and should be used with caution. We then developed Random Forest models which had remarkable improvement over unsupervised approaches and can be applied across cell types. Finally, we demonstrated the practical applications of target gene predictions by identifying a novel GWAS gene associated with multiple sclerosis.

We felt that it was important to included different types of genomic interaction data in our benchmark because each experiment assays a different type of interaction. For example, POLII ChIA-PET links tend to be close together and occur within the larger domain of CTCF ChIA-PET links[67]. Eventually, we hope to train models that will be able to identify specific links; for example, using Shannon Entropy, we can predict whether an ELS cRE is more likely to be an a CTCF loop (ubiquitous activity) or POLII loop (cell type specific activity). In the future, we also hope to expand our benchmark. During phase IV of the ENCODE project, the Ruan and Aiden labs plan on generating new ChIA-PET and Hi-C datasets for cell types covered by the Registry of cREs. Therefore, we will incorporate these new links into our benchmark. Additionally, we plan on adding confidence levels to our benchmark. With our proposed classification scheme, most of our ELS-gene pairs would be considered a "bronze" standard (i.e. they are supported by one type of link). A subset of the pairs would be considered a "silver" standard if they are supported by both physical and genetic interactions (i.e. Hi-C/ChIA-PET and eQTLs). Finally, if some of these predictions are experimentally validated using genome editing techniques, we will label them as a "gold" standard. Expanding this benchmark will enable us to further refine our methods as well as continue to annotate the registry of cREs.

One interesting result from our analysis was the importantance of distance for predicting ELS-gene links. Simply ranking genes by distance has a higher AUPR than even the best performing correlation method and distance was consistently the most important feature in our Random Forest models. This heavy reliance on distance is partially due to how ELS-gene links are distributed. Some methods such as TargetFinder, use a distance matched negative set, but we feel this method is impractical. In practice, to predict the enhancer of a target gene one would test all genes within a specific distance boundary (e.g. 200 kb) not genes of matched distances. Therefore, while distance may dominate our models, it is a biologically relevant feature that should not be ignored. Additionally, this dependence on distance suggests that while some enhancers target genes at very far distances, the majority of interactions occur nearby.

Moving forward we plan on refining our Random Forest models, by including additional biological features such as CTCF binding sites, and patterns of gene expression. We also plan on comparing our models with the published models PReSTIGE, IM-PET, PETmodule, and TargetFinder. Of particular interest will be how each program performs on individual datasets. For example, TargetFinder, which is trained on Hi-C loops, claims to be able to better identify long range interactions than close range interactions. Therefore, TargetFinder may outperform the other algorithms for identifying Aiden Hi-C pairs but may have lower performance for POLII ChIA-PET pairs.

Ultimately, we hope to apply the best performing model to predict ELS-gene pairs across all cell types in the registry. Then we can identify differences in links between cell types and how these links change over embryonic development.

Annotating Genetic Variants Associated with Psychiatric Disorders

In chapter IV we demonstrated how we can use the Registry of cREs to annotated noncoding variants associated with schizophrenia (SCZ), bipolar disorder (BPD) and major depressive disorder (MDD). By analyzing cRE activity and gene expression, we determined that SCZ SNPs and CD SNPs were enriched for cREs and target genes active in brain tissue and neural development pathways. Interestingly, we also observed enrichments in immune related features such as cRE activity in blood cells and TFs involved in the immune system. These findings are also supported by the previous enrichment in B cell specific H3K27ac peaks reported by the PGC[105].

We developed several hypotheses for this observed enrichment. The first being that the immune system is interacting directly with the central nervous system and its dysregulation results in CNS changes. Second, these immune related TFs are active in glial cells that are not surveyed by ENCODE. The brain sub-regions contain a mixture of cell types and therefore the enrichments we observe could be due to these non-neuron cells. Third, these factors may have dual roles in the CNS and immune system. This hypothesis is not completely independent from our second hypothesis but suggests that TFs involved in the immune system also have roles in brain development. We plan on testing these hypotheses first by analyzing cRE activity and gene expression. For example, we can analyzed whether psych cREs are active in both brain and immune tissues or if they tissue specific. We also plan on integrating data from the psychENCODE consortium who have generated cell type specific (neuron +, neuron -) histone modification data from brain samples. Finally, we also plan to systematically analyzing TF expression between human and mouse datasets to determine whether there is a general enrichment for TF activity in the brain and immune system.

Overall our analysis demonstrated that psych SNPs across different disorders share common enrichments for cRE activity, gene expression, and TF motifs. The most strikely was that both SCZ and CD SNPs were enriched for TF SP4 motif sites. As we detailed in Chapter IV, SP4 has previously been linked with psychiatric disorders and has been shown to prune dendrites during brain development[147]. Moving forward, we plan to further characterize SP4 binding sites using the Registry of cREs and identify potential target genes to determine regulatory networks that may be disrupted by the dysregulation of SP4. We also plan on further characterizing SP4 expression to determine if it would be a viable therapeutic target. For example, if SP4 only targets psych risk genes during brain development, it would not be a reasonable target.

While we observed many similarities between enrichments across the four studies, we currently are unable to analyze differences between the disorders. For example, we have far more variants associated with SCZ than BPD. The lack of enrichment for BPD SNPs in brain regions does not mean that BPD risk factors are not active in the brain – just that the BPD GWAS did not report significant hits in these regions. This could be due to a number of factors most notably the size of the cohort. Therefore, in the future, we hope to develop a method for combining multiple GWAS for the same phenotype to increase statistical power. We cannot simply just concatenate results from multiple studies due differences in methodologies, but we could develop an ensemble voting scheme that would weight enrichments observed in different studies. Additionally, the NHGRI-EBI GWAS catalog continues to release summary statistics from GWAS. We could use these to further prioritize regions of interest that may not reach genome wide significance. Both of these approaches would allow us to substantially increase the number of variants we could analyze and possible reveal even more about the genetic risk factors of psychiatric disorders.

Additionally, we showcased two examples of cREs with distinct enhancer activity patterns across mouse embryonic development. We are currently collaborating with the Pinnacchio and Visel labs to experimentally validate these regions and investigate allele specific activity changes. These ELS cREs have important yet different biological implications. CACNA1C, is expressed in many different tissues, such as the heart, muscle, and brain. EH37E0250841, which overlaps a CD SNP, is a brain specific cRE that we predict regulates CACNA1C. Because of this brain tissue specificity, the SNP would presumably only effect the expression of CACNA1C in the brain, not the heart or muscle.

Unlike EH37E0250841, EH37E1112284, which overlaps a SCZ SNP, is only active during early stages of brain development. ASCL1, EH37E1112284's target gene, is involved in neuronal commitment and is highly expressed in brain at early embryonic time points. Interestingly, while psychiatric disorders manifest during early adulthood, we did not observe enrichment for SNPs in cREs active at later time points. Therefore, we hypothesize that these common genetic variants result in changes in brain structure and composition that increase an individual's risk for developing a psychiatric disorder.

These two examples also demonstrate the advantage of using orthologous mouse cREs to study cRE activity patterns across embryonic development. Without the mouse embryonic data, we would not be able to cleanly define temporal patterns of brain cRE activity. While we did have fetal brain DNase data for a number of fetal stages, these samples were not surveyed at uniform time points and were subject to a number of uncontrollable biases (i.e. genetic background, gestational conditions, gender, collection methods). The brain samples collected from the embryonic mice, however, were collected from mice with the same genetic background and similar gestational conditions using uniform protocols. In the future, we plan on extending our analysis to study SNPs associated with other human disease. By analyzing enrichments in temporal patterns, we can classify diseases based on when their genetic risk factors are active (i.e. development, adulthood) which will also give us more information about possible therapeutic targets.

We also plan on generalizing our analysis by annotating cREs allele specific binding of TFs and chromatin accessibility. We plan to apply our *in silico* genotyping method described in Chapter IV to genotype SNPs overlapping cREs across all 600 cell and tissue types covered by the Registry. For each cell type, we can test at heterozygous loci whether DNase, histone modification, or TF ChIP-seq reads favor one allele. This will be a valuable resource for interpreting genetic variants reported by GWAS and whole genome sequencing but also may explain underlying mechanisms of gene regulation. For example, one anecdotal observation from characterizing allelic imbalance at psych SNPs was that the direction of allelic imbalance changed depending on the cell type. We observed this for BPD SNP rs2861405 at ZNF490 and ZNF791. When we manually surveyed other SNPs that had allelic imbalance in both directions we observed that the many of them located at bidirectional promoters. This anecdotal evidence suggests that different TFs bind at promoters depending on cell type context, thus resulting in differences in chromatin accessibility between alleles. While this is a very small sampling, we aim to test if this phenomenon is more global by characterizing AS binding at cREs.

Final Comments

Overall our work has created a foundation for the future characterization and annotation of regulatory elements in the human genome. We developed a pipeline for curating and characterizing candidate regulatory regions in human and mouse which can be expanded as more epigenomic datasets are produced. We can functionally annotate these cREs by predicting the genes they regulate and our work demonstrated that several popular methods of linking enhancer and genes should not be used. Additionally, we can utilize these tools to interpret variants associated with disease. While we demonstrated the Registry's use for annotated common variants, these methods can be applied to annotate rare, personal, or cancerous mutations as well. This is especially pertinent to clinicians as researchers as whole genome sequencing become more widely used.

REFERENCES

1. Alexander, R. P., Fang, G., Rozowsky, J., Snyder, M. & Gerstein, M. B. Annotating non-coding regions of the genome. *Nat. Rev. Genet.* **11,** 559–571 (2010).
2. Lettice, L. A. *et al.* A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Human Molecular Genetics* **12,** 1725–1735 (2003).
3. Benko, S. *et al.* Highly conserved non-coding elements on either side of SOX9 associated with Pierre Robin sequence. *Nature Genetics* **41,** 359–364 (2009).
4. Smemo, S. *et al.* Regulatory variation in a TBX5 enhancer leads to isolated congenital heart disease. *Human Molecular Genetics* **21,** 3255–3263 (2012).
5. Miguel-Escalada, I., Pasquali, L. & Ferrer, J. Transcriptional enhancers: functional insights and role in human disease. *Curr. Opin. Genet. Dev.* **33,** 71–76 (2015).
6. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337,** 1190–1195 (2012).
7. Maston, G. A., Evans, S. K. & Green, M. R. Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet* **7,** 29–59 (2006).
8. Butler, J. E. F. & Kadonaga, J. T. The RNA polymerase II core promoter: a key component in the regulation of gene expression. *Genes Dev.* **16,** 2583–2592 (2002).
9. Yao, L., Berman, B. P. & Farnham, P. J. Demystifying the secret mission of enhancers: linking distal regulatory elements to target genes. *Crit. Rev. Biochem. Mol. Biol.* **50,** 550–573 (2015).
10. Visel, A., Rubin, E. M. & Pennacchio, L. A. Genomic views of distant-acting enhancers. *Nature* **461,** 199–205 (2009).
11. Banerji, J., Rusconi, S. & Schaffner, W. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27,** 299–308 (1981).
12. Moreau, P. *et al.* The SV40 72 base repair repeat has a striking effect on gene expression both in SV40 and other chimeric recombinants. *Nucleic Acids Res.* **9,** 6047–6068 (1981).
13. Banerji, J., Olson, L. & Schaffner, W. A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. *Cell* **33,** 729–740 (1983).
14. Hill, R. E. & Lettice, L. A. Alterations to the remote control of Shh gene expression cause congenital abnormalities. *Philosophical Transactions of the Royal Society B: Biological Sciences* **368,** 20120357–20120357

(2013).

15. Heintzman, N. D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459,** 108–112 (2009).
16. Ogbourne, S. & Antalis, T. M. Transcriptional control and the role of silencers in transcriptional regulation in eukaryotes. *Biochem. J.* **331 (Pt 1),** 1–14 (1998).
17. Bessis, A., Champtiaux, N., Chatelin, L. & Changeux, J. P. The neuron-restrictive silencer element: a dual enhancer/silencer crucial for patterned expression of a nicotinic receptor gene in the brain. *Proc. Natl. Acad. Sci. U.S.A.* **94,** 5906–5911 (1997).
18. Mori, N., Schoenherr, C., Vandenbergh, D. J. & Anderson, D. J. A common silencer element in the SCG10 and type II Na+ channel genes binds a factor present in nonneuronal cells but not in neuronal cells. *Neuron* **9,** 45–54 (1992).
19. Kraner, S. D., Chong, J. A., Tsay, H. J. & Mandel, G. Silencing the type II sodium channel gene: a model for neural-specific gene regulation. *Neuron* **9,** 37–44 (1992).
20. Zhao, Y. *et al.* Brain REST/NRSF Is Not Only a Silent Repressor but Also an Active Protector. *Mol Neurobiol* **54,** 541–550 (2017).
21. Banerjee, S., Smallwood, A., Lamond, S., Campbell, S. & Nargund, G. Igf2/H19 imprinting control region (ICR): an insulator or a position-dependent silencer? *ScientificWorldJournal* **1,** 218–224 (2001).
22. Du, M. *et al.* Insulator and silencer sequences in the imprinted region of human chromosome 11p15.5. *Human Molecular Genetics* **12,** 1927–1939 (2003).
23. Szabó, P. E., Tang, S.-H. E., Silva, F. J., Tsark, W. M. K. & Mann, J. R. Role of CTCF binding sites in the Igf2/H19 imprinting control region. *Mol. Cell. Biol.* **24,** 4791–4800 (2004).
24. ENCODE Project Consortium *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447,** 799–816 (2007).
25. ENCODE Project Consortium *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).
26. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518,** 317–330 (2015).
27. Gilchrist, D. A., Fargo, D. C. & Adelman, K. Using ChIP-chip and ChIP-seq to study the regulation of gene expression: genome-wide localization studies reveal widespread regulation of transcription elongation. *Methods* **48,** 398–408 (2009).
28. Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genetics* **39,** 311–318 (2007).

29. Visel, A. *et al.* ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457,** 854–858 (2009).
30. Ong, C.-T. & Corces, V. G. CTCF: an architectural protein bridging genome topology and function. *Nat. Rev. Genet.* **15,** 234–246 (2014).
31. Hesselberth, J. R. *et al.* Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat. Methods* **6,** 283–289 (2009).
32. Boyle, A. P. *et al.* High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome Res.* **21,** 456–464 (2011).
33. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10,** 1213–1218 (2013).
34. Mersfelder, E. L. & Parthun, M. R. The tale beyond the tail: histone core domain modifications and the regulation of chromatin structure. *Nucleic Acids Res.* **34,** 2653–2662 (2006).
35. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129,** 823–837 (2007).
36. He, Y. *et al.* Improved regulatory element prediction based on tissue-specific local epigenomic signatures. *Proc. Natl. Acad. Sci. U.S.A.* **114,** E1633–E1640 (2017).
37. Liu, F., Li, H., Ren, C., Bo, X. & Shu, W. PEDLA: predicting enhancers with a deep learning-based algorithmic framework. *Scientific Reports* **6,** 28517 (2016).
38. Rajagopal, N. *et al.* RFECS: A Random-Forest Based Algorithm for Enhancer Identification from Chromatin State. *PLoS Comput. Biol.* **9,** e1002968 (2013).
39. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473,** 43–49 (2011).
40. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9,** 215–216 (2012).
41. Hoffman, M. M. *et al.* Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat. Methods* **9,** 473–476 (2012).
42. Kim, T.-K. *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465,** 182–187 (2010).
43. Gebhard, C. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507,** 455–461 (2014).
44. Danko, C. G. *et al.* Identification of active transcriptional regulatory elements from GRO-seq data. *Nat. Methods* **12,** 433–438 (2015).
45. Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat.*

Biotechnol. **30,** 271–277 (2012).

46. Patwardhan, R. P. *et al.* Massively parallel functional dissection of mammalian enhancers in vivo. *Nat. Biotechnol.* **30,** 265–270 (2012).
47. Smith, R. P. *et al.* Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nature Genetics* **45,** 1021–1028 (2013).
48. Arnold, C. D. *et al.* Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq. *Science* **339,** 1074–1077 (2013).
49. Pennacchio, L. A. *et al.* In vivo enhancer analysis of human conserved non-coding sequences. *Nature* **444,** 499–502 (2006).
50. Dickel, D. E. *et al.* Genome-wide compendium and functional assessment of in vivo heart enhancers. *Nat Commun* **7,** 12923 (2016).
51. Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L. A. VISTA Enhancer Browser--a database of tissue-specific human enhancers. *Nucleic Acids Res.* **35,** D88–92 (2007).
52. Yao, L., Tak, Y. G., Berman, B. P. & Farnham, P. J. Functional annotation of colon cancer risk SNPs. *Nat Commun* **5,** 5114 (2014).
53. Lopes, R., Korkmaz, G. & Agami, R. Applying CRISPR-Cas9 tools to identify and characterize transcriptional enhancers. *Nat. Rev. Mol. Cell Biol.* **17,** 597–604 (2016).
54. Korkmaz, G. *et al.* Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9. *Nat. Biotechnol.* **34,** 192–198 (2016).
55. Rajagopal, N. *et al.* High-throughput mapping of regulatory DNA. *Nat. Biotechnol.* **34,** 167–174 (2016).
56. Diao, Y. *et al.* A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. *Nat. Methods* **14,** 629–635 (2017).
57. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295,** 1306–1311 (2002).
58. Tolhuis, B., Palstra, R. J., Splinter, E., Grosveld, F. & de Laat, W. Looping and interaction between hypersensitive sites in the active beta-globin locus. *Mol. Cell* **10,** 1453–1465 (2002).
59. Zhao, Z. *et al.* Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nature Genetics* **38,** 1341–1347 (2006).
60. Simonis, M. *et al.* Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nature Genetics* **38,** 1348–1354 (2006).
61. Dostie, J. *et al.* Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res.* **16,** 1299–1309 (2006).

62. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326,** 289–293 (2009).
63. Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159,** 1665–1680 (2014).
64. Rao, S. S. P. *et al. Cohesin loss eliminates all loop domains, leading to links among superenhancers and downregulation of nearby genes*. *BioRxiv*
65. Mifsud, B. *et al.* Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nature Genetics* **47,** 598–606 (2015).
66. Li, G. *et al.* Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* **148,** 84–98 (2012).
67. Tang, Z. *et al.* CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. *Cell* **163,** 1611–1627 (2015).
68. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489,** 75–82 (2012).
69. Sheffield, N. C. *et al.* Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions. *Genome Res.* **23,** 777–788 (2013).
70. Shen, Y. *et al.* A map of the cis-regulatory sequences in the mouse genome. *Nature* **488,** 116–120 (2012).
71. Corradin, O. *et al.* Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. *Genome Res.* **24,** 1–13 (2014).
72. He, B., Chen, C., Teng, L. & Tan, K. Global view of enhancer-promoter interactome in human cells. *Proc. Natl. Acad. Sci. U.S.A.* **111,** E2191–E2199 (2014).
73. Zhao, C., Li, X. & Hu, H. PETModule: a motif module based approach for enhancer target gene prediction. *Scientific Reports* **6,** 30043 (2016).
74. Whalen, S., Truty, R. M. & Pollard, K. S. Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nature Genetics* **48,** 488–496 (2016).
75. Jin, F. *et al.* A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503,** 290–294 (2013).
76. Association, A. P. *Diagnostic and Statistical Manual of Mental Disorders (DSM-5®)*. (American Psychiatric Pub, 2013).
77. Tandon, R., Nasrallah, H. A. & Keshavan, M. S. Schizophrenia,‘just the facts’ 5. Treatment and prevention past, present, and future. *Schizophr. Res.* (2010).
78. Garay, R. P., Llorca, P. M., Young, A. H. & Hameg, A. Bipolar disorder: recent clinical trials and emerging therapies for depressive episodes and maintenance treatment. *Drug discovery today* (2014).

79. Belmaker, R. H. & Agam, G. Major Depressive Disorder. *N Engl J Med* **358,** 55–68 (2008).

80. McGuffin, P. *et al.* The heritability of bipolar affective disorder and the genetic relationship to unipolar depression. *Arch. Gen. Psychiatry* **60,** 497–502 (2003).

81. McGuffin, P., Owen, M. & Farmer, A. Genetic basis of schizophrenia. *The Lancet* **346,** 678–682 (1995).

82. Sullivan, P. F., Neale, M. C. & Kendler, K. S. Genetic epidemiology of major depression: review and meta-analysis. *Am J Psychiatry* **157,** 1552–1562 (2000).

83. Lichtenstein, P. *et al.* Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study. *Lancet* **373,** 234–239 (2009).

84. Lee, S. H. *et al.* Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nature Genetics* **45,** 984–994 (2013).

85. Bush, W. S. & Moore, J. H. Chapter 11: Genome-wide association studies. *PLoS Comput. Biol.* **8,** e1002822 (2012).

86. Kim, Y., Zerwas, S., Trace, S. E. & Sullivan, P. F. Schizophrenia genetics: where next? *Schizophr Bull* **37,** 456–463 (2011).

87. Chen, D. T. *et al.* Genome-wide association study meta-analysis of European and Asian-ancestry samples identifies three novel loci associated with bipolar disorder. *Mol Psychiatry* **18,** 195–205 (2013).

88. Psychiatric GWAS Consortium Bipolar Disorder Working Group. Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nature Genetics* **43,** 977–983 (2011).

89. Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium. Genome-wide association study identifies five new schizophrenia loci. *Nature Genetics* **43,** 969–976 (2011).

90. Liu, Y. *et al.* Meta-analysis of genome-wide association data of bipolar disorder and major depressive disorder. *Mol Psychiatry* **16,** 2–4 (2011).

91. Athanasiu, L. *et al.* Gene variants associated with schizophrenia in a Norwegian genome-wide study are replicated in a large European cohort. *Journal of Psychiatric Research* **44,** 748–753 (2010).

92. Ferreira, M. A. R. *et al.* Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. *Nature Genetics* **40,** 1056–1058 (2008).

93. Cross-Disorder Group of the Psychiatric Genomics Consortium. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet* **381,** 1371–1379 (2013).

94. Bergen, S. E. *et al.* 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,** 880–886 (2012).

95. International Schizophrenia Consortium *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460,** 748–752 (2009).
96. Stefansson, H. *et al.* Common variants conferring risk of schizophrenia. *Nature* **460,** 744–747 (2009).
97. Irish Schizophrenia Genomics Consortium and the Wellcome Trust Case Control Consortium 2. Genome-wide association study implicates HLA-C*01:02 as a risk factor at the major histocompatibility complex locus in schizophrenia. *Biological Psychiatry* **72,** 620–628 (2012).
98. Ruderfer, D. M. *et al.* Polygenic dissection of diagnosis and clinical dimensions of bipolar disorder and schizophrenia. *Mol Psychiatry* **19,** 1017–1024 (2014).
99. Sleiman, P. *et al.* GWAS meta analysis identifies TSNARE1 as a novel Schizophrenia / Bipolar susceptibility locus. *Scientific Reports* **3,** 3075 (2013).
100. Schaub, M. A., Boyle, A. P., Kundaje, A., Batzoglou, S. & Snyder, M. Linking disease associations with regulatory information in the human genome. *Genome Res.* **22,** 1748–1759 (2012).
101. Hayes, J. E. *et al.* Tissue-Specific Enrichment of Lymphoma Risk Loci in Regulatory Elements. *PLoS ONE* **10,** e0139360 (2015).
102. Maurano, M. T. *et al.* Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo. *Nature Genetics* (2015). doi:10.1038/ng.3432
103. Büdingen, von, H.-C., Bar-Or, A. & Zamvil, S. S. B cells in multiple sclerosis: connecting the dots. *Curr. Opin. Immunol.* **23,** 713–720 (2011).
104. Hauser, S. L. *et al.* Ocrelizumab versus Interferon Beta-1a in Relapsing Multiple Sclerosis. *N Engl J Med* **376,** 221–234 (2017).
105. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511,** 421–427 (2014).
106. Ward, L. D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. **40,** D930–4 (2012).
107. Boyle, A. P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* **22,** 1790–1797 (2012).
108. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nature Genetics* **45,** 580–585 (2013).
109. McVicker, G. *et al.* Identification of genetic variants that affect histone modifications in human cells. *Science* **342,** 747–749 (2013).

110. Grubert, F. *et al.* Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions. *Cell* **162,** 1051–1065 (2015).

111. Degner, J. F. *et al.* DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482,** 390–394 (2012).

112. Gibbs, R. A. *et al.* The International HapMap Project. *Nature* **426,** 789–796 (2003).

113. Aldridge, S. *et al.* 1000 Genomes project. *Nat. Biotechnol.* **26,** 256–256 (2008).

114. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501,** 506–511 (2013).

115. Reddy, T. E. *et al.* Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. *Genome Res.* **22,** 860–869 (2012).

116. Cavalli, M. *et al.* Allele-specific transcription factor binding to common and rare variants associated with disease and gene expression. *Hum. Genet.* **135,** 485–497 (2016).

117. Panousis, N. I., Gutierrez-Arcelus, M., Dermitzakis, E. T. & Lappalainen, T. Allelic mapping bias in RNA-sequencing is not a major confounder in eQTL studies. *Genome Biol.* **15,** 467 (2014).

118. Degner, J. F. *et al.* Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* **25,** 3207–3212 (2009).

119. Borel, C. *et al.* Biased allelic expression in human primary fibroblast single cells. *Am. J. Hum. Genet.* **96,** 70–80 (2015).

120. Waszak, S. M. *et al.* Identification and removal of low-complexity sites in allele-specific analysis of ChIP-seq data. *Bioinformatics* **30,** 165–171 (2014).

121. Satya, R. V., Zavaljevski, N. & Reifman, J. A new strategy to reduce allelic bias in RNA-Seq readmapping. **40,** e127 (2012).

122. Wu, T. D. & Nacu, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26,** 873–881 (2010).

123. Vockley, C. M. *et al.* Massively parallel quantification of the regulatory effects of noncoding genetic variation in a human cohort. *Genome Res.* **25,** 1206–1214 (2015).

124. Paquet, D. *et al.* Efficient introduction of specific homozygous and heterozygous mutations using CRISPR/Cas9. *Nature* **533,** 125–129 (2016).

125. Kim, K. *et al.* Highly efficient RNA-guided base editing in mouse embryos. *Nat. Biotechnol.* **35,** 435–437 (2017).

126. Dao, L. T. M. *et al.* Genome-wide characterization of mammalian promoters with distal enhancer functions. *Nature Genetics* **49,** 1073–1081 (2017).

127. Mei, S. *et al.* Cistrome Data Browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse. *Nucleic Acids Res.* **45,** D658–D662 (2017).

128. Yáñez-Cuna, J. O. *et al.* Dissection of thousands of cell type-specific enhancers identifies dinucleotide repeat motifs as general enhancer features. (2014).

129. Phanstiel, D. H., Boyle, A. P., Heidari, N. & Snyder, M. P. Mango: a bias-correcting ChIA-PET analysis pipeline. *Bioinformatics* **31,** 3092–3098 (2015).

130. Breiman, L. Random forests. *Mach Learn* **45,** 5–32 (2001).

131. Goldenberg, M. M. Multiple sclerosis review. *P T* **37,** 175–184 (2012).

132. Patsopoulos, N. A. *et al.* Genome-wide meta-analysis identifies novel multiple sclerosis susceptibility loci. *Ann. Neurol.* **70,** 897–912 (2011).

133. De Jager, P. L. *et al.* Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci. *Nature Genetics* **41,** 776–782 (2009).

134. Gallo, P. *et al.* On the role of interleukin-2 (IL-2) in multiple sclerosis (MS). IL-2-mediated endothelial cell activation. *Ital J Neurol Sci* **13,** 65–68 (1992).

135. Vaknin-Dembinsky, A., Balashov, K. & Weiner, H. L. 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,** 7768–7774 (2006).

136. Lee, D. *et al.* A method to predict the impact of regulatory variants from DNA sequence. *Nature Genetics* **47,** 955–961 (2015).

137. Fewings, N. L. *et al.* The autoimmune risk gene ZMIZ1 is a vitamin D responsive marker of a molecular phenotype of multiple sclerosis. *J. Autoimmun.* **78,** 57–69 (2017).

138. Forte, M. *et al.* Cyclophilin D inactivation protects axons in experimental autoimmune encephalomyelitis, an animal model of multiple sclerosis. *Proc. Natl. Acad. Sci. U.S.A.* **104,** 7558–7563 (2007).

139. Warne, J. *et al.* Selective Inhibition of the Mitochondrial Permeability Transition Pore Protects against Neurodegeneration in Experimental Multiple Sclerosis. *J. Biol. Chem.* **291,** 4356–4373 (2016).

140. McGuffin, P. Twin Concordance for Operationally Defined Schizophrenia. *Arch. Gen. Psychiatry* **41,** 541 (1984).

141. Roussos, P. *et al.* A role for noncoding variation in schizophrenia. *Cell Rep* **9,** 1417–1429 (2014).

142. Mi, H., Muruganujan, A., Casagrande, J. T. & Thomas, P. D. Large-scale gene function analysis with the PANTHER classification system. *Nat Protoc* **8,** 1551–1566 (2013).

143. Lin, G. N. *et al.* Spatiotemporal 16p11.2 protein network implicates

cortical late mid-fetal brain development and KCTD13-Cul3-RhoA pathway in psychiatric diseases. *Neuron* **85,** 742–754 (2015).

144. Degenhardt, F. *et al.* Identification of rare variants in KCTD13 at the schizophrenia risk locus 16p11.2. *Psychiatr. Genet.* **26,** 293–296 (2016).

145. Weirauch, M. T. *et al.* Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158,** 1431–1443 (2014).

146. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27,** 1017–1018 (2011).

147. Ramos, B., Gaudillière, B., Bonni, A. & Gill, G. Transcription factor Sp4 regulates dendritic patterning during cerebellar maturation. *Proc. Natl. Acad. Sci. U.S.A.* **104,** 9882–9887 (2007).

148. Sun, X. *et al.* Transcription factor Sp4 regulates expression of nervous wreck 2 to control NMDAR1 levels and dendrite patterning. *Dev Neurobiol* (2014). doi:10.1002/dneu.22212

149. Saia, G., Lalonde, J., Sun, X., Ramos, B. & Gill, G. Phosphorylation of the transcription factor Sp4 is reduced by NMDA receptor signaling. *J. Neurochem.* **129,** 743–752 (2014).

150. Goes, F. S. *et al.* Genome-wide association study of schizophrenia in Ashkenazi Jews. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **168,** 649–659 (2015).

151. Shi, J. *et al.* Genome-wide association study of recurrent early-onset major depressive disorder. *Mol Psychiatry* **16,** 193–201 (2011).

152. Shyn, S. I. *et al.* Novel loci for major depression identified by genome-wide association study of Sequenced Treatment Alternatives to Relieve Depression and meta-analysis of three studies. *Mol Psychiatry* **16,** 202–215 (2011).

153. Landgren, H. & Carlsson, P. FoxJ3, a novel mammalian forkhead gene expressed in neuroectoderm, neural crest, and myotome. *Dev. Dyn.* **231,** 396–401 (2004).

154. Hu, J. S. *et al.* Coup-TF1 and Coup-TF2 control subtype and laminar identity of MGE-derived neocortical interneurons. *Development* **144,** 2837–2851 (2017).

155. Peukert, D., Weber, S., Lumsden, A. & Scholpp, S. Lhx2 and Lhx9 determine neuronal differentiation and compartition in the caudal forebrain by regulating Wnt signaling. *PLoS Biol.* **9,** e1001218 (2011).

156. Bertuzzi, S. *et al.* Characterization of Lhx9, a novel LIM/homeobox gene expressed by the pioneer neurons in the mouse cerebral cortex. *Mech. Dev.* **81,** 193–198 (1999).

157. Aberg, K. A. *et al.* A comprehensive family-based replication study of schizophrenia genes. *JAMA Psychiatry* **70,** 573–581 (2013).

158. Ripke, S. *et al.* Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nature Genetics* **45,** 1150–1159 (2013).

159. Autism Spectrum Disorders Working Group of The Psychiatric Genomics Consortium. Meta-analysis of GWAS of over 16,000 individuals with autism spectrum disorder highlights a novel locus at 10q24.32 and a significant overlap with schizophrenia. *Mol Autism* **8,** 21 (2017).

160. Hurlin, P. J., Steingrìmsson, E., Copeland, N. G., Jenkins, N. A. & Eisenman, R. N. Mga, a dual-specificity transcription factor that interacts with Max and contains a T-domain DNA-binding motif. *EMBO J* **18,** 7019–7028 (1999).

161. van Riel, B. & Rosenbauer, F. Epigenetic control of hematopoiesis: the PU.1 chromatin connection. *Biol. Chem.* **395,** 1265–1274 (2014).

162. Flajollet, S., Poras, I., Carosella, E. D. & Moreau, P. RREB-1 is a transcriptional repressor of HLA-G. *J. Immunol.* **183,** 6948–6959 (2009).

163. Argiropoulos, B., Yung, E. & Humphries, R. K. Unraveling the crucial roles of Meis1 in leukemogenesis and normal hematopoiesis. *Genes Dev.* **21,** 2845–2849 (2007).

164. Jojic, V. *et al.* Identification of transcriptional regulators in the mouse immune system. *Nat. Immunol.* **14,** 633–643 (2013).

165. Theodorou, E. *et al.* A high throughput embryonic stem cell screen identifies Oct-2 as a bifunctional regulator of neuronal differentiation. *Genes Dev.* **23,** 575–588 (2009).

166. Lesch, K.-P. *et al.* Molecular genetics of adult ADHD: converging evidence from genome-wide association and extended pedigree linkage studies. *J Neural Transm* **115,** 1573–1585 (2008).

167. Kominami, R. Role of the transcription factor Bcl11b in development and lymphomagenesis. *Proc. Jpn. Acad., Ser. B, Phys. Biol. Sci.* **88,** 72–87 (2012).

168. Yoshida, T. & Georgopoulos, K. Ikaros fingers on lymphocyte differentiation. *Int. J. Hematol.* **100,** 220–229 (2014).

169. Zhou, X. *et al.* Transcription Factor SP4 Is a Susceptibility Gene for Bipolar Disorder. *PLoS ONE* **4,** e5196 (2009).

170. Pinacho, R. *et al.* Increased SP4 and SP1 transcription factor expression in the postmortem hippocampus of chronic schizophrenia. *Journal of Psychiatric Research* **58,** 189–196 (2014).

171. Pinacho, R. *et al.* The transcription factor SP4 is reduced in postmortem cerebellum of bipolar disorder subjects: control by depolarization and lithium. *Bipolar Disorders* **13,** 474–485 (2011).

172. Zhou, X. *et al.* Reduced expression of the Sp4 gene in mice causes deficits in sensorimotor gating and memory associated with hippocampal vacuolization. *Mol Psychiatry* **10,** 393–406 (2004).

173. Zhou, X., Qyang, Y., Kelsoe, J. R., Masliah, E. & Geyer, M. A. Impaired postnatal development of hippocampal dentate gyrus in Sp4 null mutant mice. *Genes Brain Behav* **6,** 269–276 (2007).

174. Smith, R. S. The macrophage theory of depression. *Medical Hypotheses* (1991).
175. Smith, R. S. A comprehensive macrophage-T-lymphocyte theory of schizophrenia. *Medical Hypotheses* **39,** 248–257 (1992).
176. Smith, R. S. & Maes, M. The macrophage-T-lymphocyte theory of schizophrenia: Additional evidence. *Medical Hypotheses* **45,** 135–141 (1995).
177. Schwarz, M. J., Müller, N., Riedel, M. & Ackenheil, M. The Th2-hypothesis of schizophrenia: a strategy to identify a subgroup of schizophrenia caused by immune mechanisms. *Medical Hypotheses* **56,** 483–486 (2001).
178. Schwarz, M. J., Chiang, S., Müller, N. & Ackenheil, M. T-helper-1 and T-helper-2 responses in psychiatric disorders. *Brain, Behavior, and Immunity* **15,** 340–370 (2001).
179. Miller, B. J., Buckley, P., Seabolt, W., Mellor, A. & Kirkpatrick, B. Meta-Analysis of Cytokine Alterations in Schizophrenia: Clinical Status and Antipsychotic Effects. *Biological Psychiatry* **70,** 663–671 (2011).
180. Barbosa, I. G., Bauer, M. E., Machado-Vieira, R. & Teixeira, A. L. Cytokines in Bipolar Disorder: Paving the Way for Neuroprogression. *Neural Plast.* **2014,** 360481 (2014).
181. Filiano, A. J. *et al.* Unexpected role of interferon-γ in regulating neuronal connectivity and social behaviour. *Nature* **535,** 425–429 (2016).
182. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U.S.A.* **106,** 9362–9367 (2009).
183. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. **42,** D1001–6 (2014).