Title
Systematic Identification and Analysis of Cell-state-associated cisregulatory Elements Using Statistical Approaches

Author
Yang, Yucheng

Publication Date
2017

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Systematic Identification and Analysis of Cell-state-associated *cis*-regulatory Elements Using Statistical Approaches

A thesis submitted in partial satisfaction

of the requirements for the degree Master of Science

in Statistics

by

Yucheng Yang

2017

ABSTRACT OF THE THESIS

Systematic Identification and Analysis of Cell-state-associated *cis*-regulatory Elements Using Statistical Approaches

by

Yucheng Yang

Master of Science in Statistics

University of California, Los Angeles, 2017

Assistant Professor Jingyi Li, Chair

Recent genome-wide studies have significantly advanced our understanding of the non-coding genome in higher eukaryotes. Here we developed a novel computational method to systematically identify cell-state-associated *cis*-regulatory elements for more than 300 cell and tissue types from human and mouse. Our method identified strong enrichment of associated enhancers with immune cells. We found that the *cis*-regulatory elements associated with more cell and tissue types exhibit certain genomic features, including longer length, higher conservation score and enrichment of CpG-islands. We identified enriched transcription factor (TF) motifs within the enhancers associated with each cell and tissue type. We also found that the single nucleotide polymorphisms (SNPs) identified by the Genome-Wide Association Study (GWAS) are particularly enriched in the cell-state-associated enhancers. Furthermore, we analyzed the association between human diseases and various cell and tissue types, and found that sclerosis diseases are associated with diverse immune-associated tissues and mature immune cells. Finally, we estimated enhancer-promoter signal correlations and identified enhancers exhibiting conserved correlations between human and mouse.

The thesis of Yucheng Yang is approved.

Ker Chau Li

Qing Zhou

Jason Ernst

Jingyi Li, Committee Chair

University of California, Los Angeles

2017

To my mother and father

Who have always encouraged and supported me

To explore the unknown

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1 Introduction

1.1 *cis*-regulatory elements

The vast majority (~98%) of the human genome do not code for proteins yet contain most of the disease- or phenotype-associated genetic variants. Several types of noncoding sequences, including *cis*-regulatory elements, are known to be functional in human genome. Genomic *cis*-regulatory elements, including promoters, enhancers, and insulators, exhibit dynamic activities across different cell states, and regulate spatial- and temporal-specific patterns of gene expression by recruiting sequence-specific TFs.

Genomic *cis*-regulatory elements, including promoters and enhancers, can be identified through specific epigenomic modification patterns (Figure 1-1). Nearly one decade ago, chromatin states in two human cell lines, HeLa cells and K562 cells, were mapped and used to predict ~50,000 candidate enhancers in human genome [1]. This study demonstrated that chromatin modifications at enhancers, in particular H3K4me1 and H3K27ac, are cell-type-specific and correlate with cell-type-specific gene expression throughout the genome, suggesting a potentially critical role for enhancers in lineage-specific gene regulation. Subsequent studies confirmed this result in additional cell types, and identified a large number of putative enhancers, typically between 10,000 and 150,000 per cell type. Later, researchers found that these enhancers could be further classified into "active enhancers" and "poised enhancers", which differ mainly in the presence or absence of H3K27ac mark [2]. Active enhancers are near expressed genes, while poised enhancers are next to inactive genes that can be turned on by external signals or stimulation, such as cell differentiation [2]. Now it is clear that promoters are marked by H3K4me3, active enhancers are marked by H3K27ac or H4K16ac, and poised enhancers are marked by

H3K4me1 alone or in combination with H3K27me3 [3]. More recently, researchers described super-enhancers as a class of regulatory regions with unusually strong enrichment for binding events of transcriptional coactivators [4,5]. Under the current definition, super-enhancers tend to span large genomic regions, with their median lengths generally an order of magnitude larger than those of normal enhancers (e.g. 8,667 bp versus 703 bp in mESCs) (Figure 1-2) [4,5].

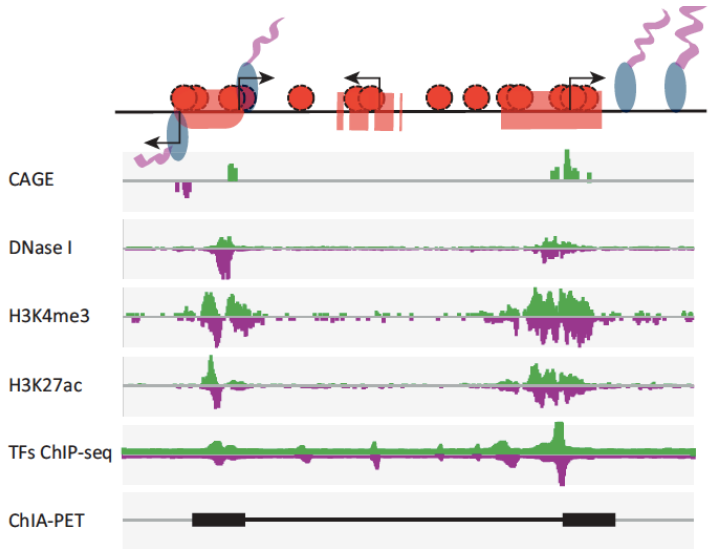


Figure 1-1: Overview of *cis*-regulation. The linear model with promoters (arrow), enhancers (carnation region), TFs (red ball), RNA Pol II (blue ellipse) and RNA transcripts (pink belt) on the genomic DNA (up panel). Examples of experimentally derived data providing evidence on transcriptional regulation (down panel). (Figure adapted from Mathelier *et al.* [6])

Many national and international epigenomic consortia have generated large data sets of epigenome maps across various cell and tissue types [7]. The researchers found that approximately 10% of the mammalian genomes are *cis*-regulatory elements [8], and about 5% of the epigenomes exhibit enhancer and promoter signatures, which are also enriched for evolutionarily conserved non-exonic elements [9,10]. It has been estimated that there are up to 1 million enhancer regions

with gene regulatory potential in mammalian genomes [11]. As researchers have annotated normal enhancers and super-enhancers in many mammalian cell types, they have realized that the activities of enhancers is highly cell-type-specific, which can determine the cell identity for that cell type [12-14]. In addition, enhancer regions have some other properties, including DNaseI hypersensitivity, combinatorial transcription factor (TF) binding, H3.3 and H2A.Z histone variant enrichment, bound RNA Pol II, and RNA production (i.e. enhancer RNAs) [12].

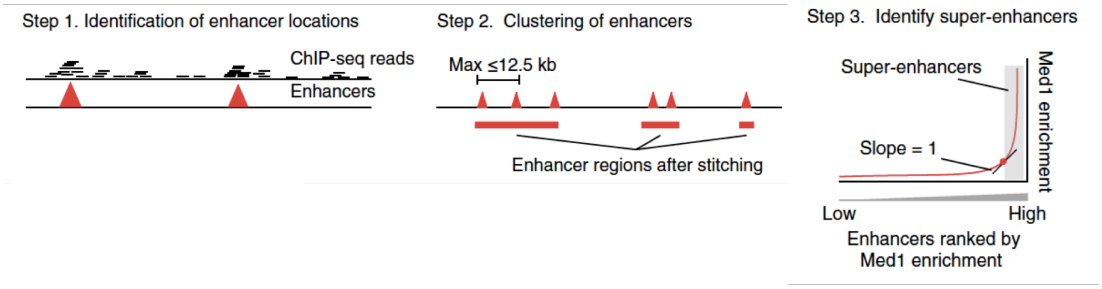


Figure 1-2: A three-step procedure of defining super-enhancers. (Figure adapted from Pott and Lieb [13])

Recent studies have revealed that active enhancers are transcribed, producing a class of noncoding RNAs called enhancer RNAs (eRNAs), which can control mRNA transcription (Figure 1-3) [15,16]. The eRNAs have been confirmed in many different cell and tissue types, suggesting a universal mechanism involved in regulating gene expression and enhancer. eRNAs have distinct genomic features with canonical lncRNAs [16]. First, although lncRNAs were defined based on the presence of H3K4me3 at their promoters, eRNAs can be produced without H3K4me3. Second, unlike the promoters of lncRNAs, eRNAs are bidirectionally transcribed. Third, although lncRNAs mostly undergo post-transcriptional maturation processes such as splicing and polyadenylation, eRNAs are rarely spliced or polyadenylated. Finally, the tissue

specificity of eRNAs is higher than lncRNAs, and the conservation of eRNAs is lower than lncRNAs.

Recent technologies, such as CAGE (cap analysis of gene expression), have been developed to quantify the *in vivo* activities of promoters and enhancers. CAGE captures the 5' ends of RNA molecules in a biological sample. In a previous study, FANTOM5 Consortium demonstrated that, instead of using histone modification maps from multiple ChIP-seq datasets, genomic signature from the CAGE technology can be used to identify *in vivo* promoters and enhancers and quantify their activities across hundreds of cell and tissue types in human and mouse, although at far lower sensitivity [17-19].

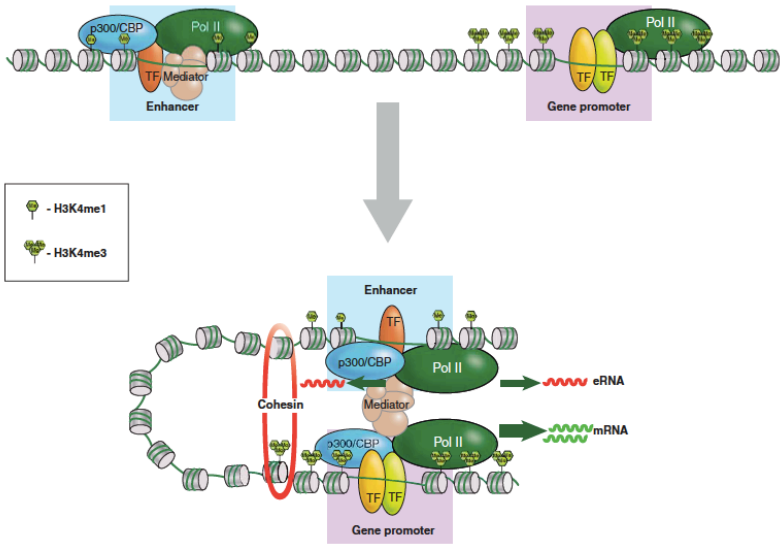


Figure 1-3: The synthesis and functions of eRNAs. (Figure adapted from Kim, Hemberg and Gray [15])

1.2 Enhancer-Promoter interactions

In the nucleus, the genome is organized and partitioned into functional compartments in the three-dimensional space [20]. Identifying the regulatory targets of enhancers is crucial for

understanding their biological functions in regulating cell differentiation, homeostasis and even disease development. One strategy is to identify the long-range looping interactions involving enhancer elements using a variety of chromosome conformation capture (3C)-based techniques [21]. Genome-wide applications of these techniques to define the chromatin interactomes of human and mouse cells confirmed that the genome is divided into active and inactive compartments [21]. These are further organized into sub-megabase-sized topologically associated domains (TADs) that correlate with genomic regions that constrain the spread of heterochromatin and are relatively conserved across cell types [22,23]. Although the genome-wide resolution of such studies remains somewhat limited, the resulting chromatin connectivity maps suggest that only approximately 7% of the looping interactions exist between adjacent genes, indicating that assignment based on linear proximity is error prone [24]. Indeed, many enhancers map large distances away from their targets, bypassing the nearest gene [24,25]. Long-range gene regulation by enhancers *in vivo* involves close spatial proximity between distal enhancers and their target gene promoters in the three-dimensional nuclear space, most likely involving a direct interaction, while the intervening sequences are looped out (Figure 1-4) [26].

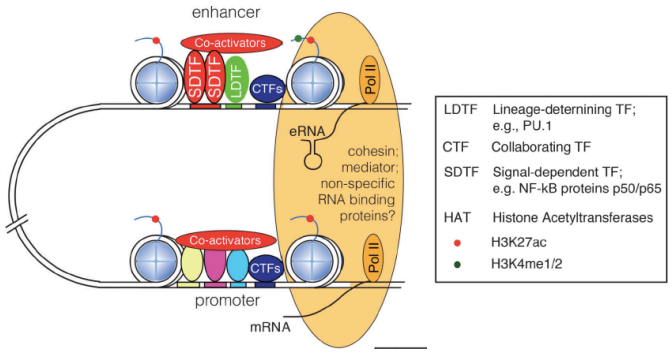


Figure 1-4: Structural interactions between enhancers and promoters. The connections (orange oval) involve cohesin and the mediator complex. (Figure adapted from Heinz *et al.* [27])

The principles encoded at the genomic sequence level underlying such three-dimensional organization and chromatin interaction are poorly understood. Although great efforts have been made (recently reviewed by Mora *et al*. [28]), our ability of discriminating the gene target of a distal regulatory element from other nearby transcribed genes is still limited. There exists some very recent work on predicting enhancer-promoter interactions based on multiple functional genomic features [29,30]. In Roy *et al*. [29], a method called RIPPLE was developed using a combination of random forests and group LASSO in a multi-task learning framework to predict enhancer-promoter interactions in multiple cell lines, using DNase-seq, histone marks, TF ChIP-seq, and RNA-seq data as input features. Whalen *et al*. developed TargetFinder based on boosted trees to predict enhancer-promoter interactions using DNase-seq, DNA methylation, TF ChIP-seq, histone marks, CAGE, and gene expression data [30]. Furthermore, considering that there are 10,000-150,000 enhancers in a typical cell type, one gene is anticipated as the regulatory target by multiple putative enhancers. In fact, computational predictions based on correlations between gene expression and activities of distal enhancers across panels of cell lines also led to the prediction that genes are regulated by multiple distal enhancers [8,31,32]. Deciphering the interaction networks between enhancers and promoters will greatly improve our understanding of gene expression regulation in development and disease.

1.3 Mutations and variants in the *cis*-regulatory elements

Systematic identification and interpretation of *cis*-regulatory elements is not only essential for understanding the mechanisms of human development, but also key to studying the phenotypic variations among human populations and the etiology of many human diseases [33]. Accumulating evidence indicates the importance of *cis*-regulatory element alteration associated with multiple

diseases (Figure 1-5). This is demonstrated by the human gene mutation database (HGMD), which includes more than 3000 disease-implicated mutations categorized as "regulatory" [34]. A meta-analysis of genome-wide association studies (GWAS) SNPs revealed enrichment of disease-associated sequence variants in putative *cis*-regulatory elements, providing insights into the pathogenesis of many common human diseases [35]. It is estimated that 93% of SNPs associated with human phenotypes by GWAS are located outside of protein coding regions, with most of which lying in *cis*-regulatory elements [36]. More complex human disease cases such as cohesinopathies, diabetes, and cancers are linked to variants located in *cis*-regulatory elements [37-40].

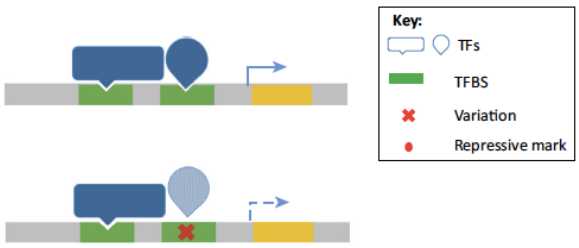


Figure 1-5: TF binding in a normal (top) and disease (down) condition. A genomic variant located within one of the TF binding sites disrupts the binding of the TF to the DNA sequence, thus affecting the expression of the target gene. (Figure adapted from Mathelier *et al.* [6])

Many computational tools have been developed to predict the impact of variants within *cis*-regulatory elements by integrating both experimentally-derived and sequence-based features. For example, RegulomeDB [41] and HaploReg [42] prioritizing genomic variants by computing a heuristic score from the number of regulatory features overlapping the variants. More sophisticated machine-learning approaches have been used to predict variants with pathogenic effects. For example, CADD [43] and DeepSEA [44] predict pathogenic variants using support vector machine and deep neural network approaches, respectively. Computational methods for genomic

variation annotation and gene prioritization have progressed fast [45]. By integrating data from multiple sources and advanced computational tools, these approaches will greatly contribute to biological discovery and translational medicine. Systematic identification of *cis*-regulatory elements and their interaction network will be a key foundation to this goal.

In this thesis, we aim to systematically identify and characterize cell-state-associated *cis*-regulatory elements in human and mouse. First, we developed a novel computational pipeline to identify cell-state-associated enhancers and promoters using CAGE datasets from FANTOM5 Consortium. We then analyzed the genomic features, biological functions and motif patterns for these cell-state-associated enhancers and promoters. In addition, we discovered cell-state-specific enrichment of the genomic variation for specific human disease. Finally, we inferred enhancer-promoter signal correlations and identified some enhancers with conserved correlations between human and mouse.

Chapter 2 Materials and methods

2.1 Group FANTOM5 samples into different cell states

We used 1,241 samples of CAGE peaks in human (892) and mouse (349), which provide genome-wide transcription start site (TSS) locations. The CAGE samples were generated by the FANTOM5 Consortium (http://fantom.gsc.riken.jp/5/). The FANTOM5 samples cover a wide range of cell and tissue types, including different cell lines, primary cells and *in vivo* tissues (http://fantom.gsc.riken.jp/5/sstar/Browse_samples). The replicate samples from the same cell or tissue type or cancer cell lines from the same cancer type were then categorized as the same group (i.e. cell state). The samples were grouped according to the FANTOM5 cell ontology (can be accessed via http://fantom.gsc.riken.jp/5/datafiles/latest/extra/Enhancers/) or cell line annotation information from Cellosaurus (http://web.expasy.org/cellosaurus). The cell states with only one sample were not used in our analysis. The remaining samples were categorized into 262 and 82 groups in human and mouse, respectively (Table 2-1).

Table 2-1 The statistics of grouping samples by cell and tissue types in human and mouse

Type	Human		Mouse	
	Number of groups	Number of samples	Number of groups	Number of samples
Cell line	54	206 (2-22 samples/group)	—	—
Primary cell	169	550 (2-13 samples/group)	43	121 (2-6 samples/group)
Tissue	39	136 (2-9 samples/group)	39	228 (2-24 samples/group)
Total	262	892	82	349

2.2 Activity data of *cis*-regulatory elements taken from FANTOM5 datasets

We considered three categories of *cis*-regulatory elements: (i) enhancers, (ii) promoters of protein-coding genes, and (iii) promoters of lncRNAs (Table 2-2). We used genome annotation

from GENCODE (https://www.gencodegenes.org) for the PCGs (protein-coding genes) and lncRNAs in human and mouse (Release 25 mapped to GRCh37 for human; Release 12 for mouse). The promoter regions were defined as the genomic intervals ranging from 500 bp upstream to 200 bp downstream of all transcription start sites for protein-coding genes and lncRNAs. The enhancer regions were obtained from FANTOM5 Consortium in which the enhancer regions were identified using CAGE peaks [17]. To avoid expression bias from gene regions, we removed the enhancer regions overlapping with exons from the FANTOM5 Consortium-defined dataset. Genomic coordinates of enhancer and promoter regions from FANTOM5 datasets that are mapped to mouse reference assembly mm9 were converted to mm10 using LiftOver utility [46].

Table 2-2 The statistics of the *cis*-regulatory elements in our analysis

cis-regulatory element	Human	Mouse
Enhancer	65,367	44,400
PCG promoter	93,970	59,630
lncRNA promoter	8,813	4,333

The activity scores (i.e. the TPM (tags per million) estimates from the CAGE samples) for the enhancer regions were obtained from the FANTOM5 Consortium [17]. We calculated the activity scores for the promoter regions using the CAGE peaks that are located within the promoter regions [19]. The scores of the peaks within the same promoter region were averaged.

2.3 Identification of cell-state-associated *cis*-regulatory elements

We developed the model using a scheme adapted from Li *et al* [14]. Our approach consists of several steps, which are described in depth in the following (Figure 2-1).

First of all, we collected CAGE activity data for each *cis*-regulatory element (*x*) in each sample. The groups with only one sample were removed. All the samples were organized into different groups (*m*) (*m*=262 and 82 in human and mouse, respectively). We calculated $\mu_{x,m}$, the mean CAGE score of the *cis*-regulatory element *x* in each group

$$\mu_{x,m} = \frac{1}{n}\sum_{i=1}^{n} C_{x,i}$$

Where $i = 1, 2, \ldots, n$ (the number of samples for each group), and C_i is the CAGE score of the *cis*-regulatory element *x* in sample i.

Step 1: We used the ANOVA to filter out *cis*-regulatory elements whose activity scores do not have significant variation across all cell and tissue types. ANOVA aims to test whether a *cis*-regulatory element (*x*) has the same group mean scores across all groups. The null hypothesis for element *x* can be expressed as

$$H_{0,x} : \mu_{x,1} = \mu_{x,2} = \cdots = \mu_{x,m}$$

We applied a threshold $\alpha_1 = 10^{-10}$ to the Bonferroni-corrected *p*-values, and selected element *x* as a candidate associated enhancer or promoter for the following analysis if the null hypothesis $H_{0,x}$ was rejected. Step 1 can increase the computational efficiency in Step 2 by reducing the number of candidate associated enhancers or promoters to be tested.

Step 2: We then applied the t-test to find cell-state-associated *cis*-regulatory elements for each cell and tissue type under a series of association thresholds. We performed pairwise one-tailed t-tests among the *m* groups to identify cell-state-associated enhancers or promoters for each group. Given two different groups *r* and *s*, the null hypothesis for element *x* is

$$H_{0,xrs} : \mu_{x,r} \leq \mu_{x,s}$$

We applied a threshold $\alpha_2 = 0.01$ to the resulting *p*-values, and further defined the element *x* as a cell-state-associated enhancer or promoter for group *r* if the null hypothesis $H_{0,xrs}$ was rejected for more than *t* percent (i.e. association threshold) of total (*m* - 1) tests, where group *r* is compared with the rest (*m* - 1) groups. We tried a series of association thresholds from 0.05 to 0.95 with a stepsize 0.05.

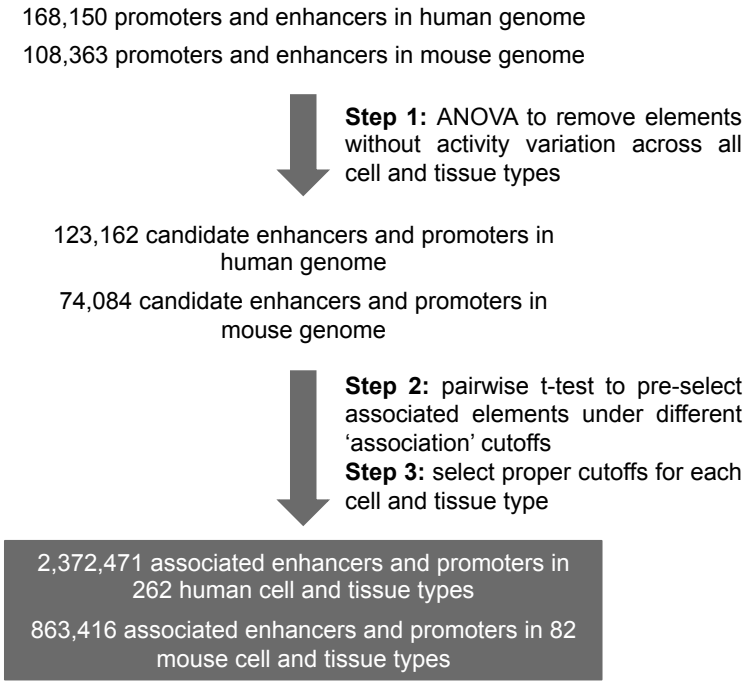


Figure 2-1: A framework of identifying cell state-associated *cis*-regulatory elements.

Step 3: Finally, we selected a reasonable association threshold *t* for each group *r* to determine its associated *cis*-regulatory elements. If we set *t* too large, the inclusion of too many irrelevant groups (i.e. cell and tissue types) will reduce the statistical power of the test in Step 3. If we set *t* too small, the inclusion of too few or too similar groups will make it difficult to distinguish which elements exhibit relatively higher activity in specific group. Thus, it will be necessary to decide an optimal association threshold for each group because selecting the proper

number of groups to compare with each group is important in our approach. Briefly, we tried a series of *t* values ranging from 0.05 to 0.95. We then defined the "association score", which is the product between the *t* value and the number of elements selected under the *t* value. The *t* value that achieves the largest "association score" was selected as the optimal association threshold.

2.4 Analysis of the associated *cis*-regulatory elements

For Gene Ontology analysis, we estimated the enrichment of the biological process terms for different cell states based on their associated PCG promoters in human and mouse. We calculated the significance of GO term enrichment in each cell state using a hypergeometric test. The top three most enriched GO terms in each cell state were displayed. The *p*-values were adjusted using the Bonferroni correction.

To investigate the genomic sequence features of the *cis*-regulatory elements, the enhancers and promoters were annotated with CpG-islands and conservation scores. CpG-island annotations [47] and phastCons vertebrate conservation scores [48] were downloaded from the UCSC Genome Browser. We overlapped each enhancer and promoter with the CpG-island annotations, and calculated an average conservation score for each element.

For the motif analysis, we selected the top 500 enhancers for each cell state in human and mouse. We then extracted sequences of these associated enhancers and searched for motifs using FIMO [49] with the following settings: zero or one occurrence per sequence (ZOOPS), a motif size range of 8-22 nt, and an E-value cutoff of 3. After identifying the *de novo* motifs, we used TOMTOM [50] to compare them to the JASPAR motif database [51], recording the top five matches for each cell state. The motifs were visualized using Ceqlogo in the MEME suite [52].

For the disease-associated SNP analysis, we estimated the enrichment of the sets of disease-associated SNPs for different human cell states based on their associated *cis*-regulatory elements. We calculated the significance of disease enrichment in each cell state using a hypergeometric test. The *p*-values were adjusted using the Bonferroni correction. The human trait/disease-associated SNPs were obtained from GWASdb v2 [53], which includes 250,984 SNPs associated with 1,831 phenotypes or diseases.

2.5 Inferring enhancer-promoter connections

We used two different similarity measures to infer connections between enhancers and promoters: (i) the Jaccard index and (ii) the Spearman correlation.

We first used the Jaccard index to infer the connection using the identified associated cell and tissue types of enhancers and promoters. The Jaccard index $J(x_1, x_2)$ measures the overlap of the associated cell and tissue types between an enhancer and a promoter as

$$J(X_1, X_2) = \frac{|X_1 \bigcap X_2|}{|X_1| + |X_2| - |X_1 \bigcap X_2|}$$

Where X_1 and X_2 is the associated cell and tissue types of the enhancer and the promoter, respectively, and $X_1 \cap X_2$ is the associated cell and tissue types shared between the enhancer and the promoter. We performed permutation for 10,000 times to estimate the significance for the observed Jaccard index. The enhancer-promoter pairs with *p*-value<0.001 were considered to be connected.

In previous studies, the interactions between enhancers and promoters were generally inferred using cross-cell-type correlation (e.g., Pearson correlation and Spearman correlation) of their activity signals. Here we used the Spearman correlation to compare the CAGE signals

across all cell and tissue types between an enhancer and a promoter. The enhancer-promoter pairs with Spearman correlations larger than 0.6 were considered to be connected.

To identify conserved promoter-enhancer connections, we obtained orthologous families of protein-coding genes from TreeFam v9 [54], and conserved genomic regions in alignment between human and mouse from the UCSC Genome Browser [55]. We used the Markov clustering algorithm [56] to identify clusters of highly inter-connected conserved promoter-enhancer pairs.

Chapter 3 Identification of cell-state-associated *cis*-regulatory elements

3.1 Statistical approaches to identify cell-state-associated *cis*-regulatory elements

In our method, the first step (i.e., ANOVA procedure) aims to filter out the *cis*-regulatory elements whose CAGE signals do not have significant variation among all groups (i.e., tissue and cell types). After this step, the numbers of candidate *cis*-regulatory elements have greatly decreased relative to their total numbers, especially for enhancers (Figure 3-1). For example, in human, enhancer has a decrease rate at 57%, while PCG promoter and lncRNA promoter only have a decrease rate at 7% and 15%, respectively. These results suggest that a large fraction of the enhancers in the mammalian genomes show weak activity fluctuation across hundreds of cell and tissue types.

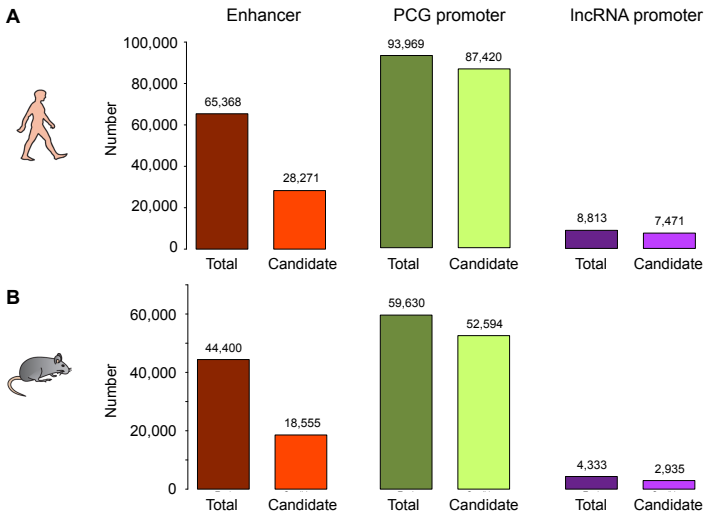


Figure 3-1: An ANOVA procedure reduces the number of candidate associated *cis*-regulatory elements in human (A) and mouse (B).

Then, we applied the t-test to find cell-state-associated *cis*-regulatory elements for each cell and tissue type. To identify an optimal association threshold "*t* percentage" for each cell and tissue type, we tried a series of *t* values ranging from 0.05 to 0.95. For example, the associated *cis*-regulatory elements selected at $t = 0.5$ for a tissue or cell type would have stronger activities in that type than at least 50% of the total cell and tissue types. Obviously, a larger *t* percentage threshold will lead to fewer cell-state-associated *cis*-regulatory elements. Thus, we defined the "association score", which is the product between the *t* percentage and the number of elements selected at the *t* percentage, to select the optimal association threshold, i.e., the "*t* percentage" that achieves the largest "association score".

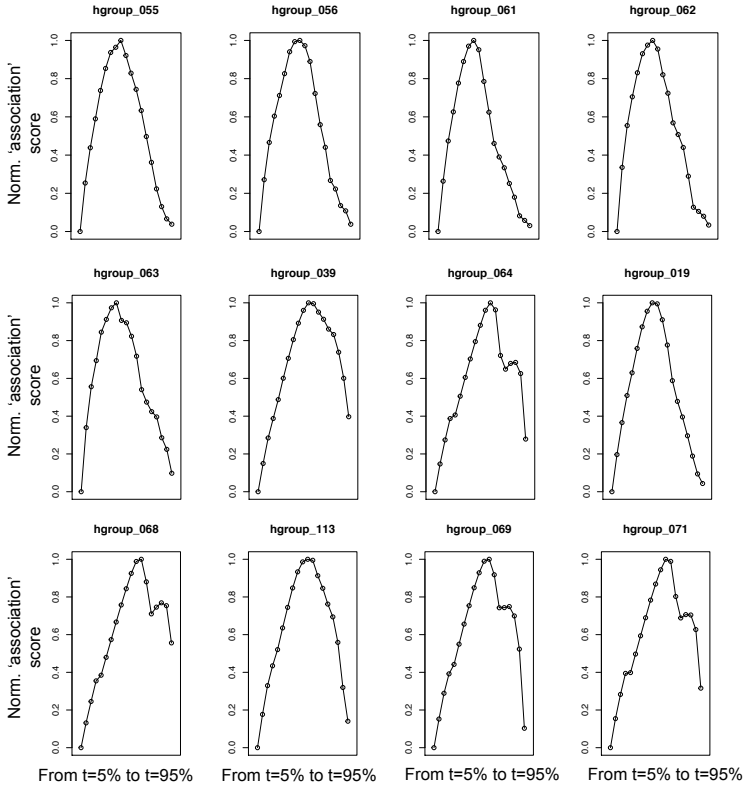


Figure 3-2: Examples of "association scores" as *t* percentages vary for multiple human cell and tissue types. The "association score" is the product of the *t* percentage (i.e., association threshold) and the number of elements under the *t* percentage.

We found that optimal association threshold (the *t* percentage corresponding to the peak of each association score curve) varies among different cell and tissue types (Figure 3-2).

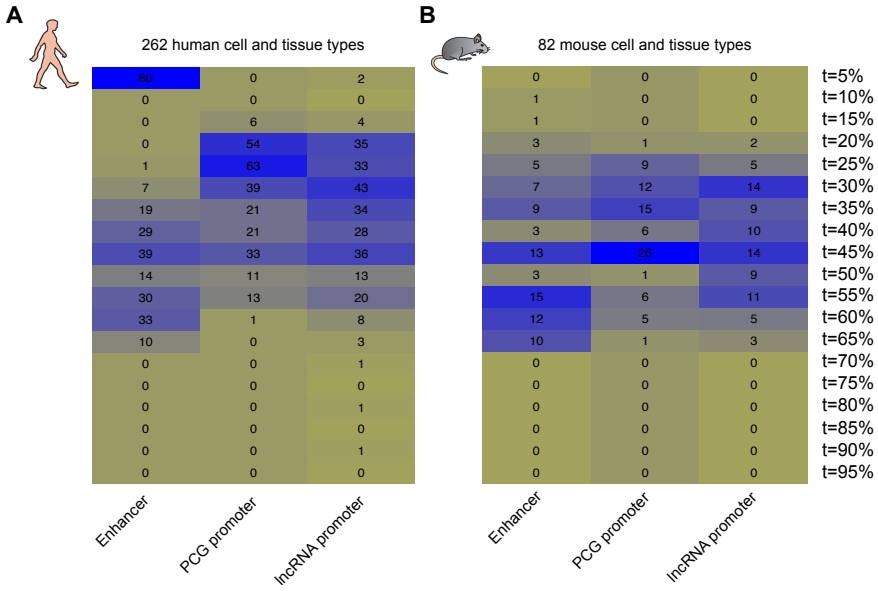


Figure 3-3: Selected *t* percentages (i.e., association thresholds) for different *cis*-regulatory elements in human (A) and mouse (B). The entries show the number of cell and tissue types corresponding to each selected *t* percentage. In other words, each column represents the distribution of the selected *t* percentages among all the cell and tissue types for each *cis*-regulatory element. Yellow and blue indicate smaller and greater numbers, respectively.

Using this method, we selected the optimal association threshold for each cell and tissue type, and obtained the cell-state-associated *cis*-regulatory elements for each cell and tissue type in human and mouse. Figure 3-3 shows that the distributions of the selected *t* percents (i.e., association thresholds) across all the cell and tissue types are similar for different *cis*-regulatory elements in human and mouse. In addition, we found that for a large fraction of human cell and

tissue types (80 out of 262), the association thresholds for enhancers were selected to be 5%, indicating that most enhancers in these cell and tissue types exhibit relatively lower activity.

3.2 Numbers of cell-state-associated *cis*-regulatory elements in human and mouse

First, we counted the number of associated *cis*-regulatory elements in all human and mouse samples (Figure 3-4). Compared to PCG and lncRNA promoters, the number of associated enhancers in most cell and tissue types is relatively small. These results are consistent with our previous observation that only a small subset of the total enhancers is active in a given cell type.

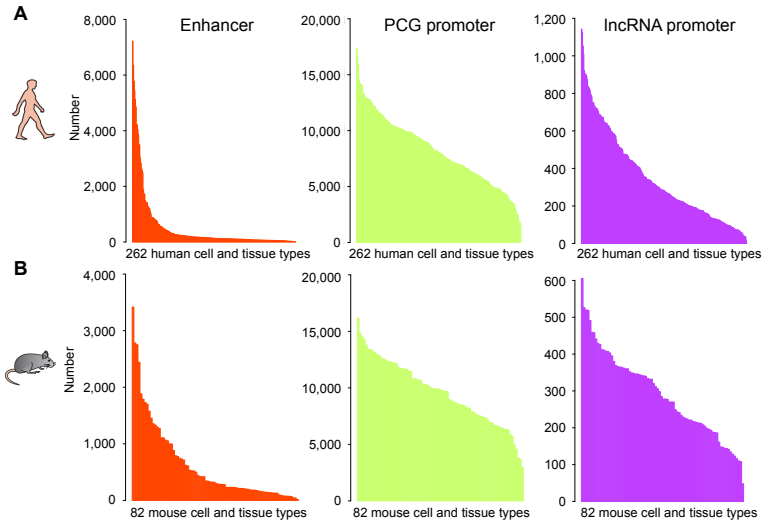


Figure 3-4: Numbers of associated *cis*-regulatory elements in 262 human (A) and 82 mouse (B) cell and tissue types.

Next, we asked which cell and tissue types have the most associated *cis*-regulatory elements. We identified top 20 cell and tissue types with the most associated *cis*-regulatory elements in human (Table 3-1). We found that these human cell and tissue types contain the most associated promoters: (i) neural cells, (ii) immune cells, (iii) stem/progenitor cells, (iv) muscle

cells/tissues, and (v) testis tissues. These observations are consistent with previous results from RNA-seq analysis [57]. Notably, the associated enhancers were extremely enriched in the immune cells, an observation also reported in a recent study [58]. This immune-specific enrichment of associated enhancers could be due to immune cells containing more enhancers and/or better sequencing coverage in the FANTOM5 datasets.

Table 3-1 Top 20 human cell and tissue types with the most associated *cis*-regulatory elements

Rank	Cell/tissue type with the most associated enhancers [number]	Cell/tissue type with the most associated PCG promoters [number]	Cell/tissue type with the most associated lncRNA promoters [number]
1	CD14+ monocyte, treated with Group A streptococci [7,222]	Astrocyte, cerebral cortex [17,325]	Testis [1,143]
2	CD14+ monocyte [6,358]	Ciliary epithelial cell [16,673]	Lymphoblastoid [1,127]
3	CD14+ monocyte, treated with Candida [5,781]	Lymphoblastoid [15,899]	Medulla oblongata [1,053]
4	CD14+ monocyte, treated with BCG [5,435]	Smooth muscle cell, colonic [14,595]	CD19+ B cell (pluriselect) [1,011]
5	CD14+ monocyte, treated with Trehalose dimycolate (TDM) [5,127]	Skeletal muscle satellite cell [14,353]	Natural killer cell [923]
6	CD14+ monocyte, treated with Salmonella [4,863]	Pineal gland [14,179]	CD34+ stem cell [905]
7	Basophils [4,222]	Schwannoma [14,128]	CD8+ T cell [904]
8	CD14+ monocyte, treated with B-glucan [4,211]	Whole blood (ribopure) [14,127]	Locus coeruleus [898]
9	CD14+ monocyte, treated with lipopolysaccharide [4,046]	Cervical adenocarcinoma [14,027]	CD4+CD25+CD45RA+ naive regulatory T cell expanded [890]
10	CD14+ monocyte, treated with Cryptococcus [3,883]	Endothelial cell, lymphatic [13,487]	CD4+CD25+CD45RA+ naive regulatory T cell [873]
11	CD14+ monocyte, treated with IFN + N-hexane [3,521]	Pituitary gland [13,386]	Pituitary gland [838]
12	Natural killer cell [3,483]	CD8+ T cell (pluriselect) [13,204]	CD4+CD25-CD45RA- memory conventional T cell [834]
13	CD14+CD16+ monocyte [3,027]	CD34+ stem cell [13,184]	Throat [825]
14	CD14+ monocyte, mock treated [2,844]	Medulla oblongata [13,028]	CD8+ T cell (pluriselect) [819]
15	CD14+CD16- monocyte [2,660]	Duodenum [12,944]	CD14+ monocyte [799]
16	Lymphoblastoid [2576]	Testis [12896]	CD4+CD25+CD45RA- memory regulatory T cell [790]
17	CD8+ T cell [2490]	Mesenchymal precursor cell, ovar [12878]	CD4+ T cell [783]
18	CD4+ T cell [1897]	Natural killer cell [12870]	Kidney [754]
19	Peripheral blood mononuclear cell [1732]	Mesenchymal precursor cell, adipose [12801]	CD4+CD25-CD45RA+ naive conventional T cell [750]
20	CD14-CD16+ monocyte [1722]	Myoblast [12785]	Pineal gland [749]

We next repeated these analyses on mouse cell and tissue types. We identified the top 10 mouse cell and tissue types with the most associated *cis*-regulatory elements (Table 3-2). We observed similar cell and tissue types associated with most PCG and lncRNA promoters in mouse and human. Interestingly, we again found that all the top cell and tissue types with the most associated enhancers are immune cells. In conclusion, the associated *cis*-regulatory elements show obvious cell-specific enrichment patterns, and more importantly, and the enrichment patterns are conserved between human and mouse.

Table 3-2 Top 10 mouse cell and tissue types with the most associated *cis*-regulatory elements

Rank	Cell/tissue type with the most associated enhancers [number]	Cell/tissue type with the most associated PCG promoters [number]	Cell/tissue type with the most associated lncRNA promoters [number]
1	Natural helper cell, naïve [3,418]	Lung, neonate [16,166]	Thymus, neonate [605]
2	CD4+CD25+ regulatory T cell [2,783]	Skin, neonate [14,852]	Cerebellum, embryo [526]
3	Stem cell (cKit+ Sca1- lineage-) [2,750]	Thymus, neonate [14,524]	Lung, neonate [520]
4	CD4+CD25-CD44- naive conventional T cell, PMA and ionomycin stimulation [2,441]	Heart, embryo [14,270]	Pituitary gland, embryo [519]
5	CD4+CD25-CD44- naive conventional T cell [1,883]	Testis, embryo [13,784]	Eyeball, neonate [491]
6	Common myeloid progenitor [1789]	Whole body, embryo [13422]	Stomach, embryo [459]
7	Thymus, neonate [1731]	Intestine, embryo [13390]	Kidney, neonate [458]
8	CD4+CD25+ regulatory T cell, antiCD3 CD28 stimulation [1702]	Cerebellum, embryo [13333]	Skin, neonate [441]
9	CD4+CD25+ regulatory T cell, PMA and ionomycin stimulation [1577]	Heart, neonate [13128]	Neuron, striatal [429]
10	MC1+Gr1+ myeloid-derived suppressor cell cancer [1454]	Whole body, neonate [12982]	Epididymis and seminiferous tubule, neonate [426]

3.3 Numbers of associated cell states for different *cis*-regulatory elements

We then analyzed the numbers of associated cell and tissue types for different *cis*-regulatory elements. First, we found that the specificity patterns of the *cis*-regulatory elements are

consistent between human and mouse: lncRNA promoters show higher cell type specificity than PCG promoters do, and enhancers exhibit the strongest cell type specificity (Figure 3-5). These results are confirmed by previous studies [17,18,27,59].

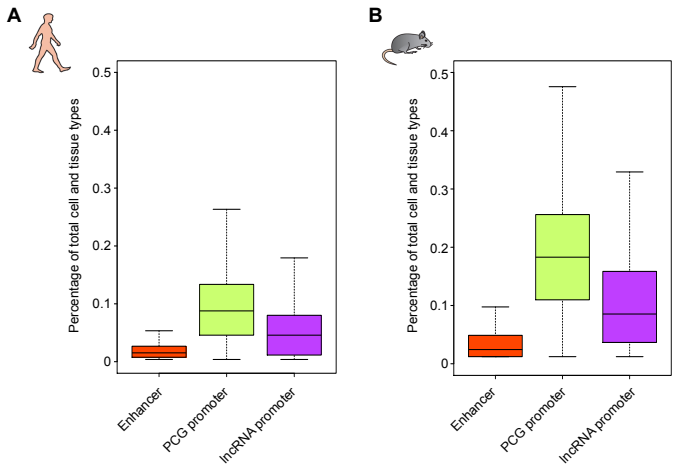


Figure 3-5: Numbers of associated cell and tissue types for different *cis*-regulatory elements in human (A) and mouse (B).

In addition, we found that, in both human and mouse, the enhancers associated with more cell and tissue types have greater lengths, indicating that longer enhancers may have higher potential in regulating cell-type-specific gene expression (Figure 3-6). Recently, it has been reported that super-enhancers, which are basically defined as large clusters of typical enhancers, can be occupied by multiple TFs, cofactors and chromatin regulators that are important in mediating cell differentiation states [4,5,13]. Although the associated enhancers we identified are not super-enhancers due to their much smaller sizes (median size ~200-400 bp) than those of super-enhancers (median size ~10,000 bp [5]), our results support the hypothesis that typical enhancers with longer lengths may be occupied by more regulators than shorter typical enhancers to fulfill their higher regulatory potential.

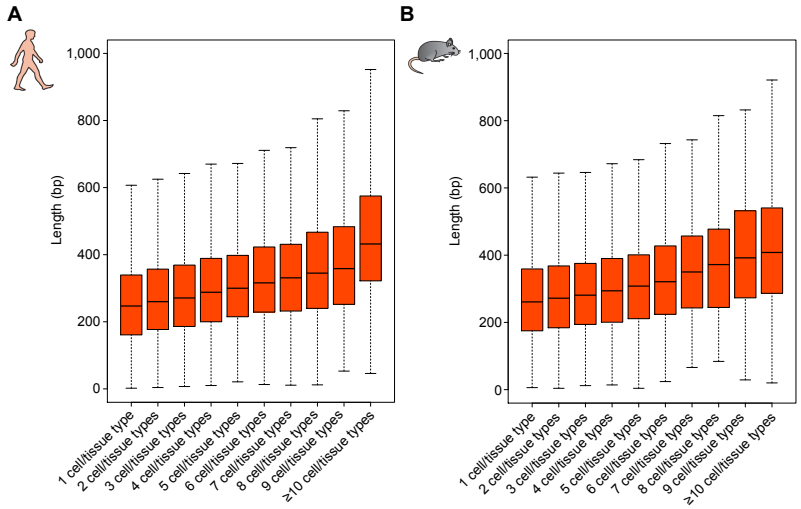


Figure 3-6: Relationship between the enhancer length and the number of associated cell and tissue types in human (A) and mouse (B).

We then asked whether the numbers of associated cell and tissue types revealed interesting functions of the genomic features they are associated with. We examined the CpG-islands and conservation scores of the associated *cis*-regulatory elements to characterize their relationship (Figure 3-7). We found that the *cis*-regulatory elements associated with more cell and tissue types exhibit higher conservation score and greater enrichment of CpG-islands. This trend exists for all the three categories of *cis*-regulatory elements in both human and mouse. Typically, CpG-islands are located in the promoter regions of protein-coding genes. The methylation state of CpG-islands within promoter regions is associated with the regulation of gene transcription in vertebrates [60,61]. More importantly, the *cis*-regulatory elements that are associated with more cell and tissue types exhibit stronger conservation and enrichment with CpG-islands, indicating that these functional *cis*-regulatory elements may be broadly hypomethylated across various cell states. Our results revealed that, similar to the CpG-islands within the promoter regions of

protein-coding genes, the CpG-islands within the enhancer regions and the lncRNA promoter regions are also important for gene expression regulation.

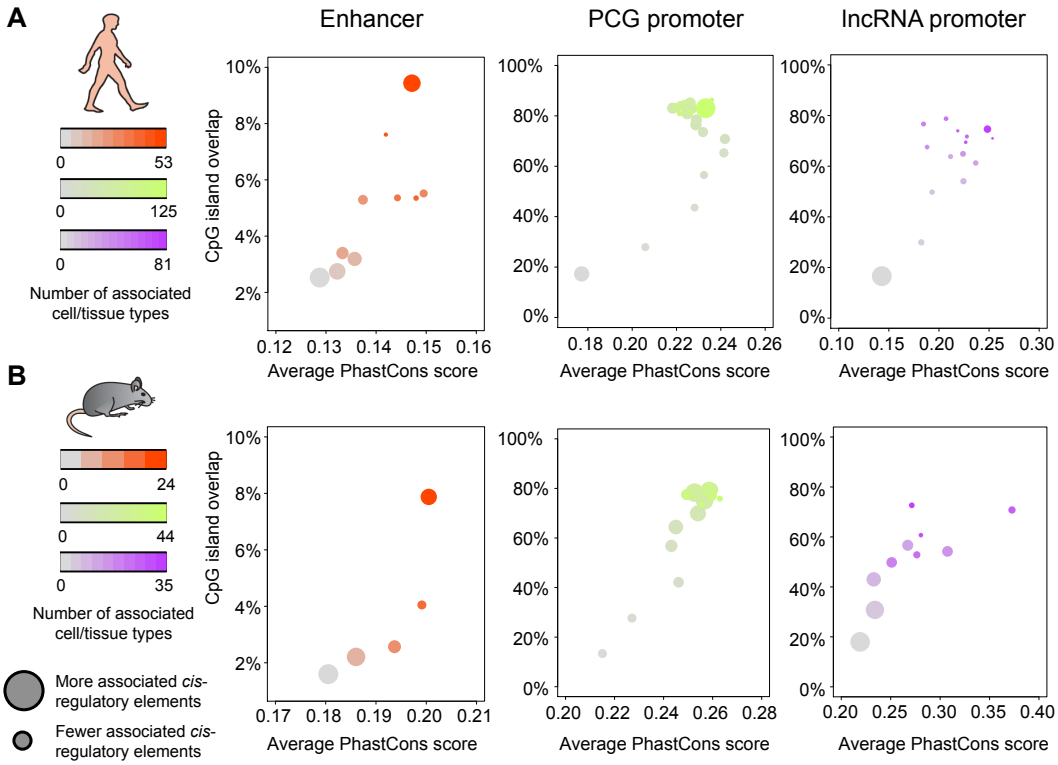


Figure 3-7: Relationship between the conservation score and the CpG island enrichment for different *cis*-regulatory elements in human (A) and mouse (B). The color of the circles indicates the number of associated cell and tissue types. The size of the circles represents how many *cis*-regulatory elements in this group.

Chapter 4 Biological functions of cell-state-associated *cis*-regulatory elements

4.1 Enriched biological functions of cell-state-associated PCG promoters

We first investigated the biological functions of the associated PCG promoters. We identified the GO terms enriched in the associated PCG promoters from various human tissue types (Figure 4-1). The results reveal that the associated PCG promoters are enriched with biological processes that largely define the identities of the respective cell states. For example, the neural tissues are enriched with nervous system development and synaptic signaling; the spleen is enriched with immune system processes and leukocyte activation; and testis is enriched with spermatogenesis and male gamete generation. These results are consistent with previous results from RNA-seq analysis [57] and confirm that the associated *cis*-regulatory elements identified by our approach are biologically meaningful.

We then repeated these analyses on mouse cell and tissue types. We confirmed that the enriched GO terms in the PCG promoters associated with mouse tissues are biologically meaningful (Figure 4-2). For example, the neural primary cells are enriched with neurogenesis and nervous system development; the hepatocyte is enriched with metabolic processes; immune and hematopoietic cells are enriched with immune response, immune process and various cellular metabolic processes [62,63]. Interestingly, mesenchymal stem cell shows similar functional enrichment to immune and hematopoietic cells [64,65].

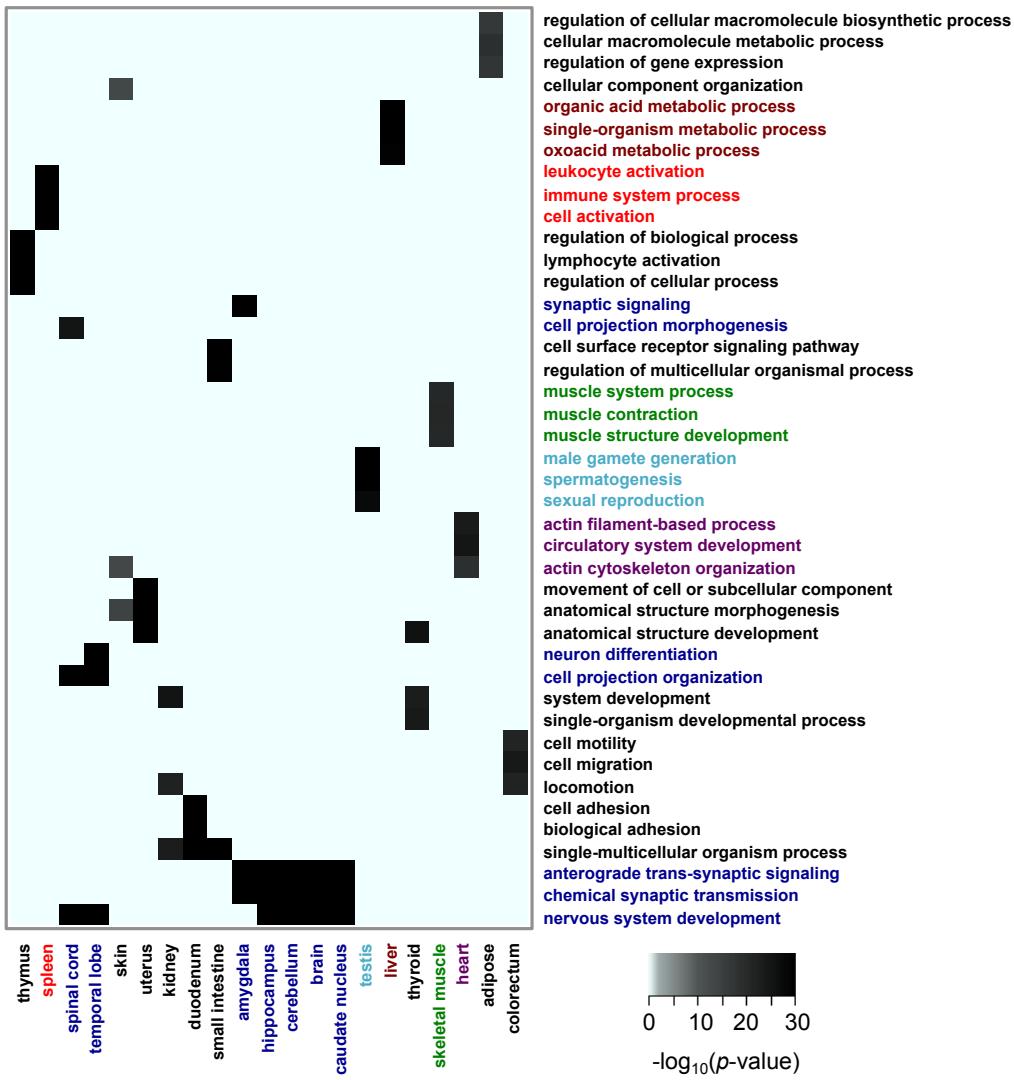


Figure 4-1: Enriched biological processes in the PCG promoters associated with 21 human tissues. Higher enrichment scores (defined as –log10 transformed Bonferroni-corrected *p*-values) are shown in darker colors. Tissue types and their corresponding biological processes are labeled with the same color.

Most of the mouse tissue datasets from FANTOM5 are from neonate and embryonic developmental stages. We found multiple tissue types that can be clearly separated by their developmental stages but not by their anatomical positions (Figure 4-3). The embryonic tissues

are specifically enriched with cellular component organization and biogenesis, indicating that the processes of organ assembly and arrangement are critical to the early development of various tissue types [66,67].

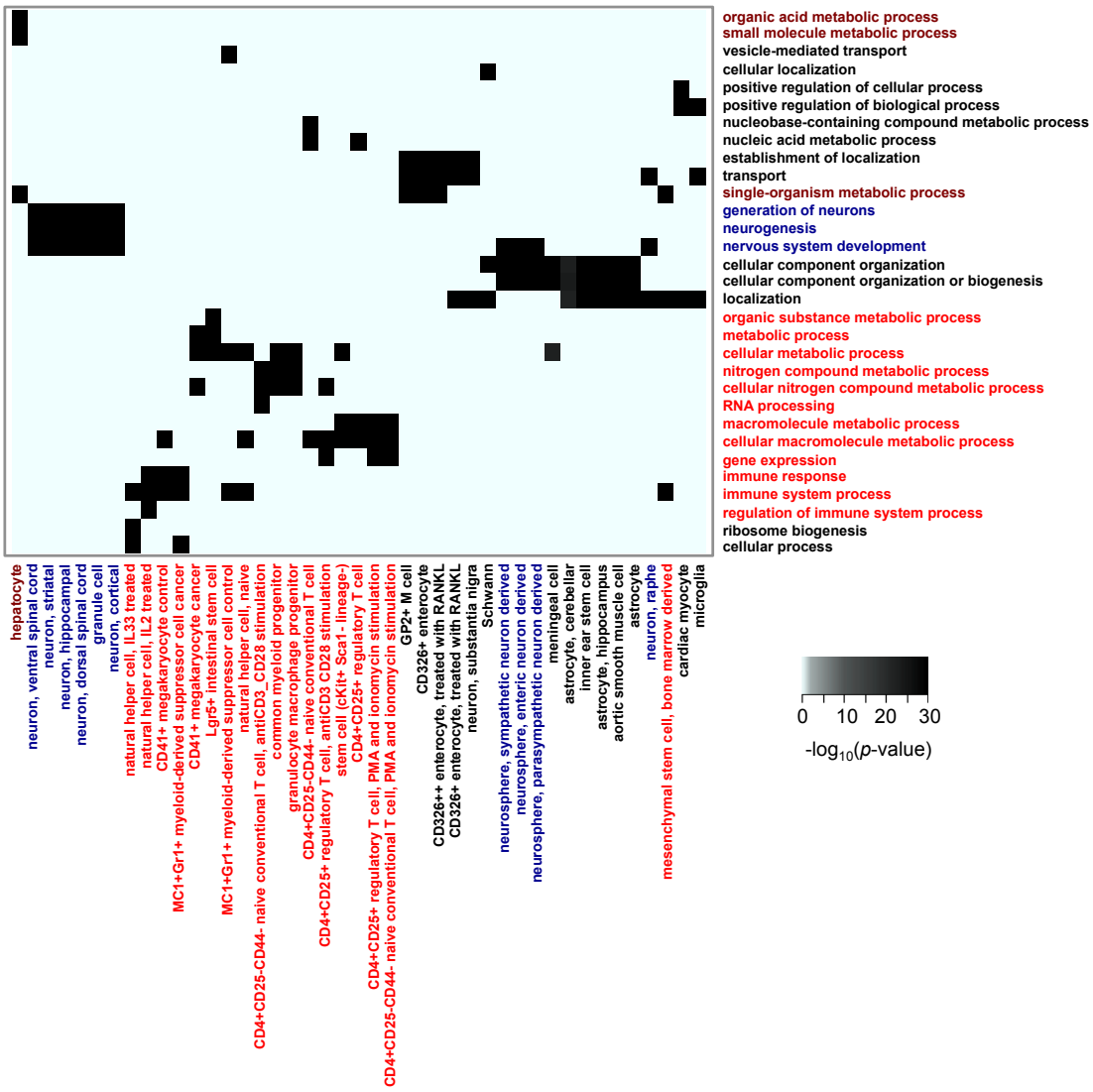


Figure 4-2: Enriched biological processes in the PCG promoters associated with 43 mouse primary cells. Higher enrichment scores (defined as –log10 transformed Bonferroni-corrected *p*-values) are shown in darker colors. Cell types and their corresponding biological processes are labeled with the same color.

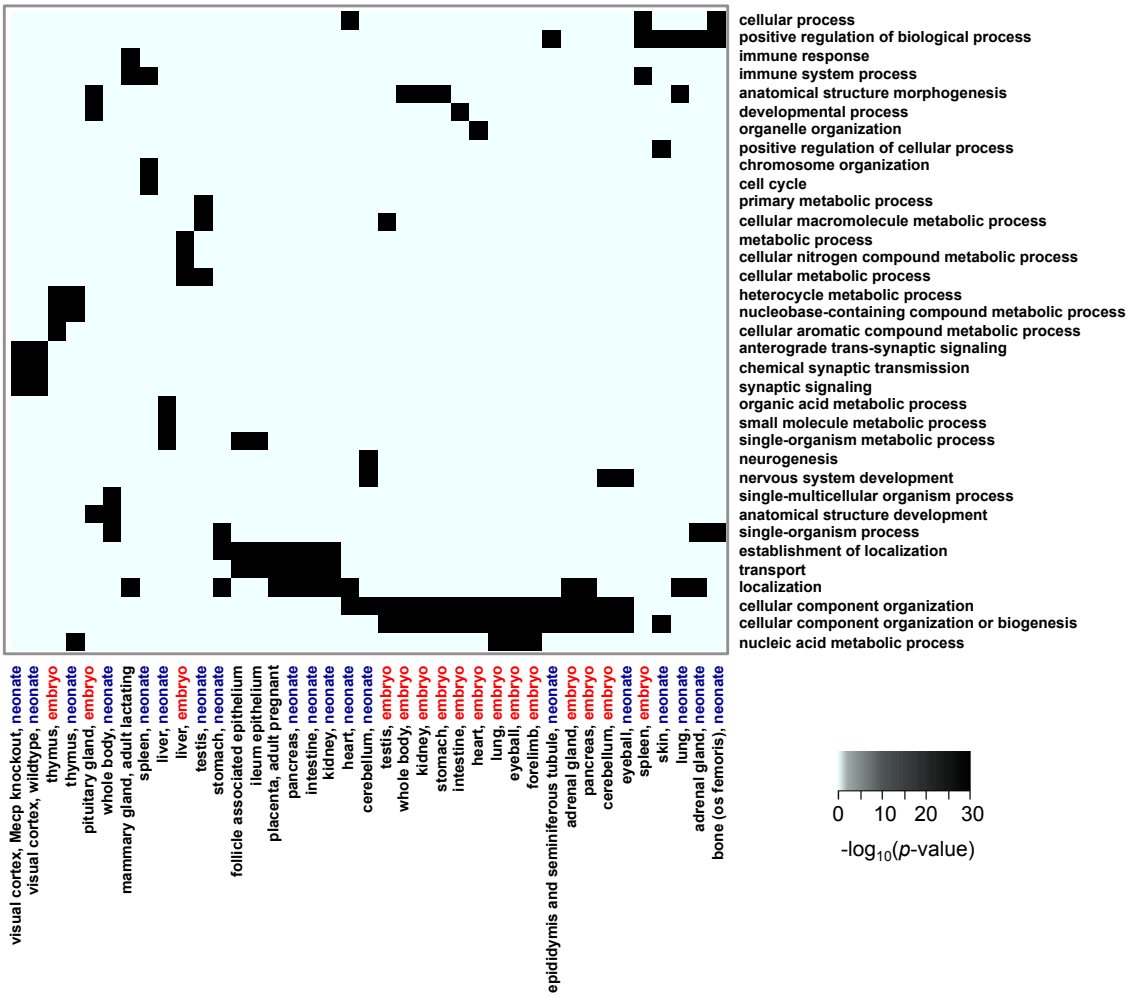


Figure 4-3: Enriched biological processes in the PCG promoters associated with 39 mouse tissues. Higher enrichment scores (defined as –log10 transformed Bonferroni-corrected *p*-values) are shown in darker colors. Tissue types from neonate and embryonic developmental stages are labeled using blue and red, respectively.

4.2 Motif discovery in cell-state-associated enhancers

Enhancer regions can be recognized and bound by TFs to establish cell-state-specific expression patterns, which are critical to developmental control and diseases [68,69]. Previous studies have identified several TFs that are important in regulating tissue development and

homeostasis. For the enhancers associated with each cell and tissue type, we systematically identified enriched *de novo* motifs using FIMO [49], and then matched them to putative TFs with known motif patterns using TOMTOM [50]. Here we show that the ZNF263, MEF2C, SP2 and NFYA motifs are predominantly enriched in the enhancers associated with neuron, heart, liver and testis, respectively (Figure 4-4). Most of these TFs (i.e., MEF2C, SP2 and NFYA) are well supported by previous studies [70-72]. Using this analysis, we could also predicted novel TFs that may play important roles in regulating cell differentiation states by binding to enhancer regions.

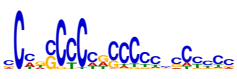
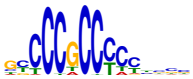
Cell/Tissue type	*De novo* discovered motif	Matched motif in database	Factor	E-value
Neuron			ZNF263	3.82e-07
Heart			MEF2C	4.81e-05
Liver			SP2	5.98e-07
Testis			NFYA	1.87e-06

Figure 4-4: Representative examples of *de novo* motif discovery results and significantly matched known motifs.

In previous analysis (Section 2 in Chapter 3), we discovered that the associated enhancers were top enriched in the immune cells. We then systematically predicted putative regulatory TFs in the enhancers associated with all the immune cell types in human (Figure 4-5). Nine TFs (ZNF263, IRF1, EGR1, SP2, SP1, FOXP1, ZNF740, RREB1, and EWSR1-FLI1) were identified in the associated enhancers from more than 20 immune cell types, and most of these TFs were

reported functional in immune system regulation [73-77]. We reasoned that these frequently identified TFs can function as co-factors in regulating gene expression in diverse cell types. For example, SP1 is a general factor with many other interacting partners [78].

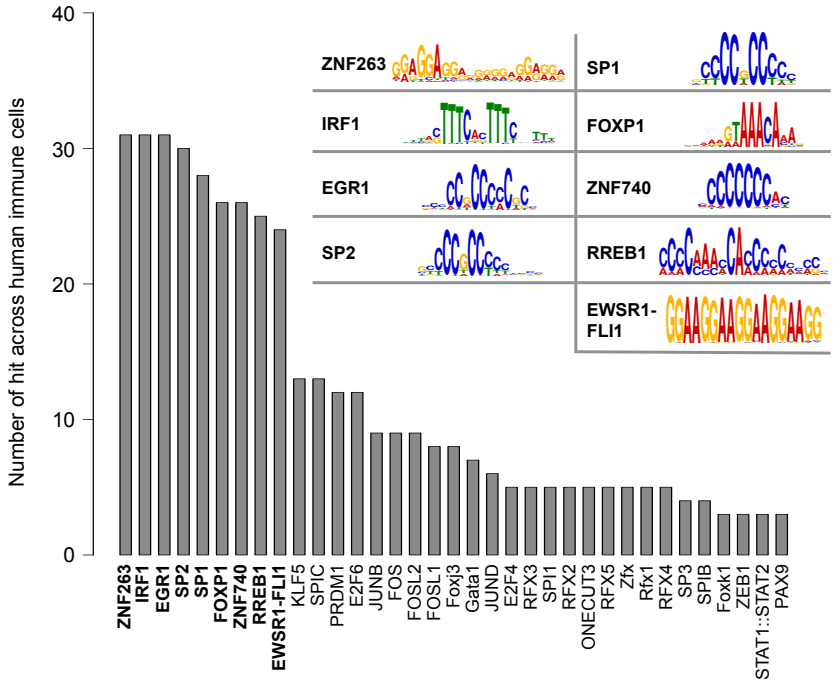


Figure 4-5: Motifs match with known TFs in the associated enhancers across human immune cell types. The nine TFs that were identified in more than 20 immune cell types are highlighted. Known motifs of these TFs are also shown.

4.3 Dysregulation of *cis*-regulatory elements in human disease

In recent years, GWAS have greatly extended our knowledge of genetic loci associated with human disease risks and other phenotypes. GWAS have discovered thousands of genetic loci that contribute susceptibility to various diseases [79-81]. However, the understanding and interpretation of disease mechanisms was limited because most of the GWAS SNPs are either intronic or intergenic, and do not affect protein-coding sequence [82-84]. We found that 33% of the disease-

associated SNPs located in the associated *cis*-regulatory elements were from enhancers, a large increase from 24%, the percentage of the disease-associated SNPs in all *cis*-regulatory elements (Figure 4-6). These results suggest that, compared to PCG and lncRNA promoters, GWAS SNPs are particularly enriched in the cell-state-associated enhancers.

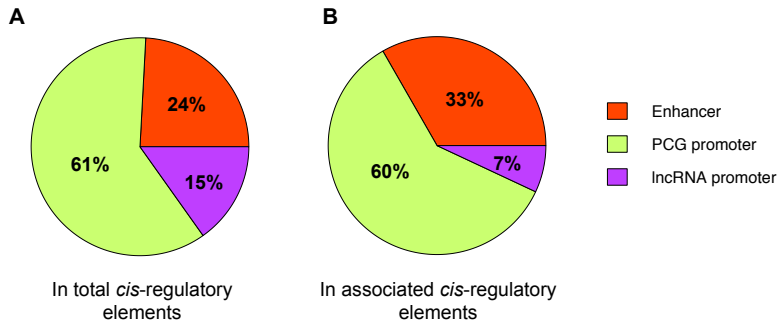


Figure 4-6: The distribution of GWAS SNPs in all *cis*-regulatory elements (A) and cell-state-associated *cis*-regulatory elements (B).

We then used the associated *cis*-regulatory elements for discovering disease-relevant cell and tissue types. We estimated the enrichment of trait-relevant variants from the GWASdb v2 [53]. We confirmed that the cell and tissue types with the strongest enrichment for a given disease were generally biologically meaningful (Figure 4-7). Notably, a large set of sclerosis diseases were predicted relevant to six tissues, including spleen, thymus, lung, kidney, thyroid and amygdala, all of which are related to immune activity [85-87]. These results suggest that the germ-line mutations of these autoimmune diseases may lead to broad tissue-specific pathology. Interestingly, we noticed that biliary cirrhosis was relevant to many tissues, some of which were surprisingly brain tissues. Recent studies reported brain abnormalities in primary biliary cirrhosis and biliary cholangitis [88,89].

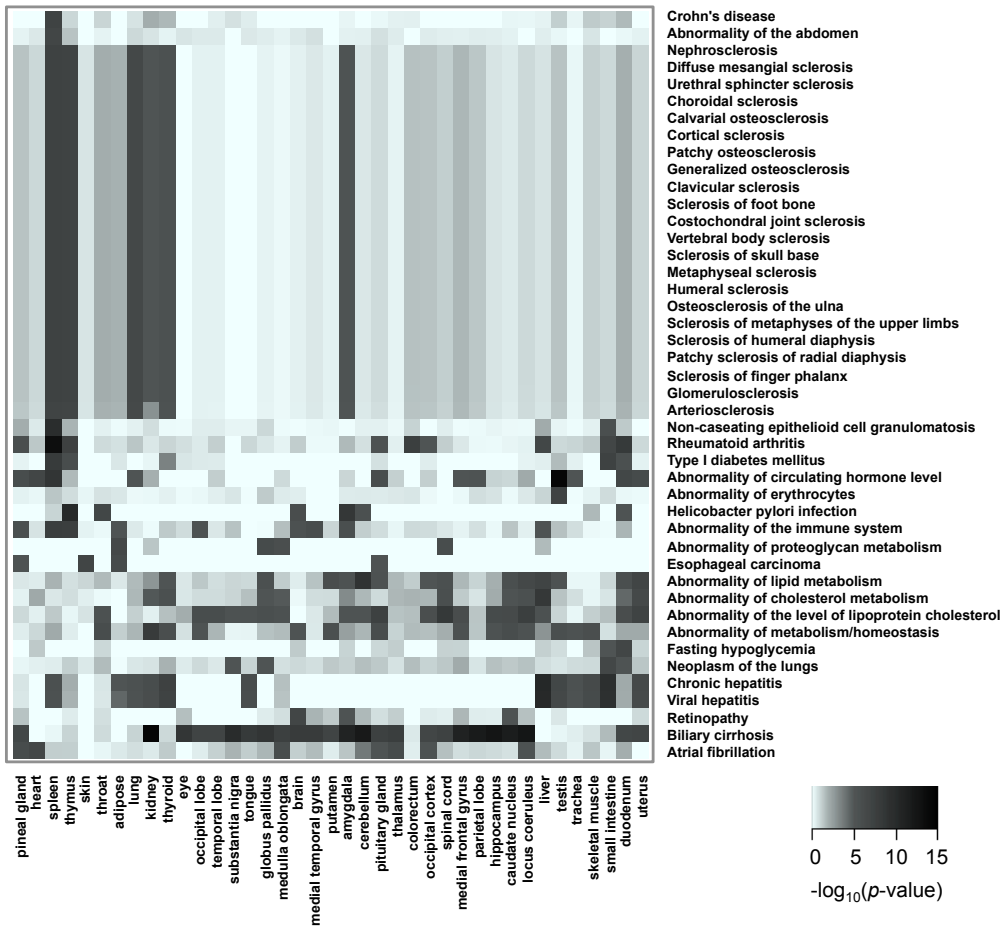


Figure 4-7: GWAS SNP enrichment in *cis*-regulatory elements associated with 39 human tissues. Higher enrichment scores (defined as –log10 transformed Bonferroni-corrected *p*-values) are shown in darker colors.

We further systematically analyzed the association between human diseases and immune cells (Figure 4-8). The enrichments for immune cells were generally biologically relevant to human diseases. For example, we found that Crohn's disease, type I diabetes mellitus and rheumatoid arthritis SNPs were enriched in diverse immune cells. In addition, we confirmed that various sclerosis diseases were associated with diverse mature immune cells, consistent with our previous result.

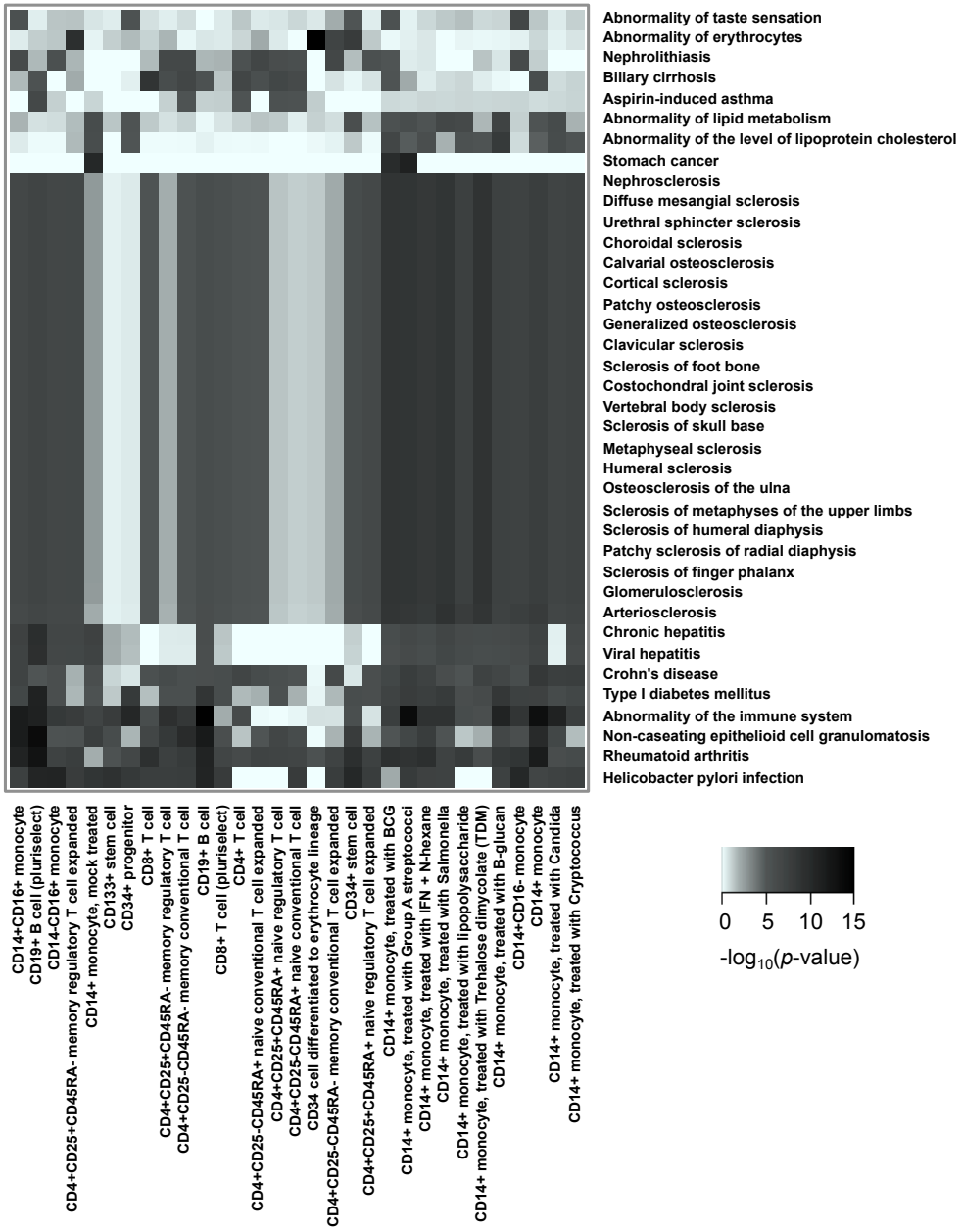


Figure 4-8: GWAS SNP enrichment in *cis*-regulatory elements associated with 31 human immune cells. Higher enrichment scores (defined as –log10 transformed Bonferroni-corrected *p*-values) are shown in darker colors.

4.4 Predicting enhancer-promoter signal dependency

We reasoned that if the cell-state-associated or activity signal of an enhancer across cell and tissue types matches the pattern of a promoter across cell and tissue types, this observation can provide evidence that the gene is a potential regulatory target of that enhancer. Therefore, we used two different but complementary similarity measures, Jaccard index and Spearman correlation coefficient, to infer enhancer-promoter signal dependency in human and mouse (Figure 4-9). Jaccard index measures the overlap of the associated cell and tissue types; while Spearman correlation coefficient evaluates the monotonicity of the CAGE signal across all cell and tissue types.

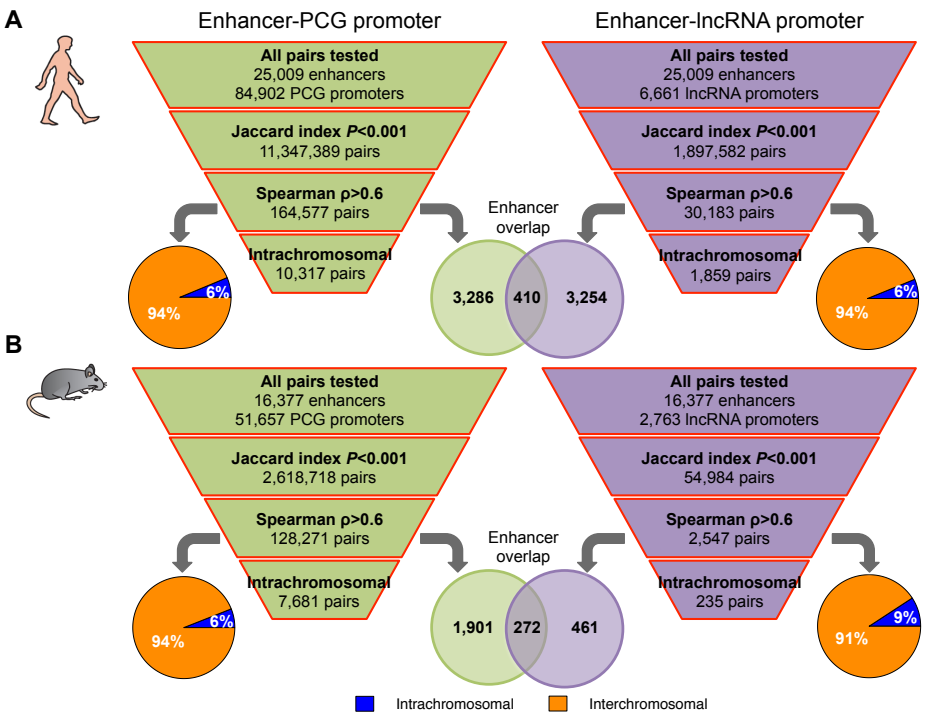


Figure 4-9: A framework of identifying enhancer-promoter signal dependency in human (A) and mouse (B). Putative enhancer-promoter pairs were captured by Jaccard index and Spearman correlation coefficient and filtered by chromosomal location.

Among all possible enhancer-promoter pairs, about 16 million pairs overlapped significantly in terms of their associated cell and tissue types (p-value<0.001). Furthermore, we identified 325,578 pairs (~2%) of them exhibiting strong signal correlation (Spearman correlation coefficient ρ>0.6) across all cell and tissue types, and only ~5% of the 325,578 pairs were from the same chromosome. This observation highlights the existence of putative interchromosomal interactions between enhancers and promoters, which are largely ignored in current studies [90,91]. In addition, we noticed that only a small fraction of enhancers were shared by PCG and lncRNA promoters, indicating different regulatory architecture of eRNA-producing enhancers for protein-coding genes and lncRNAs. Next, we investigated the distance between the enhancers and promoters for the correlation pairs we identified (Figure 4-10). We found that eRNA-producing enhancers are preferentially engaged in an interaction with the proximal promoters, consistent with a recent study [24].

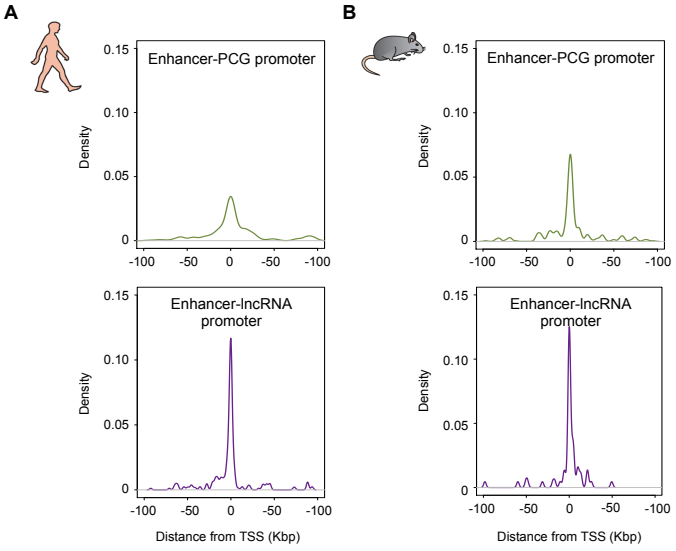


Figure 4-10: Distances between inferred enhancer-promoter pairs in human (A) and mouse (B).

Finally, we identified evolutionally conserved enhancer-promoter pairs between human and mouse, which are defined as those enhancers from conserved genomic regions exhibiting significant dependency with promoters of homologous genes in human and mouse. In total, 18 enhancers were identified in conserved signal correlation (Table 4-1). These conserved enhancer-promoter correlation pairs could form three clusters, in which two enhancers (Enhancer E1 and E2) exhibit higher connectivity degree (Figure 4-11). The eRNA transcripts from these enhancers may have potential regulatory functions on their dependent protein-coding genes, which can be validated by further experiments.

Table 4-1 Enhancers with evolutionally conserved signal dependency

Enhancer ID	Conserved enhancer in human genome	Conserved enhancer in mouse genome	Number of dependent protein-coding genes
E1	chr20:57738628-57739096	chr2:174613995-174614499	117
E2	chr17:43303050-43303852	chr11:103175003-103175320	44
E3	chr9:117147567-117148178	chr4:63400996-63401327	8
E4	chr2:43401535-43402035	chr17:84145888-84146207	5
E5	chr6:35279423-35279720	chr17:28218336-28218767	3
E6	chr11:14600110-14600431	chr7:114317396-114318106	3
E7	chr14:88472465-88473193	chr12:98269655-98269954	2
E8	chr2:137084668-137085432	chr1:128787173-128787581	1
E9	chr2:158273487-158273978	chr2:58135097-58135380	1
E10	chr6:37017809-37018367	chr17:29395241-29395748	1
E11	chr7:50350065-50350197	chr11:11692574-11693120	1
E12	chr7:150265859-150266137	chr6:48685652-48686118	1
E13	chr14:81685513-81686063	chr12:91588218-91588758	1
E14	chr15:66111254-66111589	chr9:64789302-64789821	1
E15	chr20:4792276-4792776	chr2:132019271-132019811	1
E16	chr20:34356129-34356371	chr2:156190481-156190740	1
E17	chr21:15854421-15854886	chr16:75855510-75855821	1
E18	chrX:78363432-78363724	chrX:107217728-107217877	1

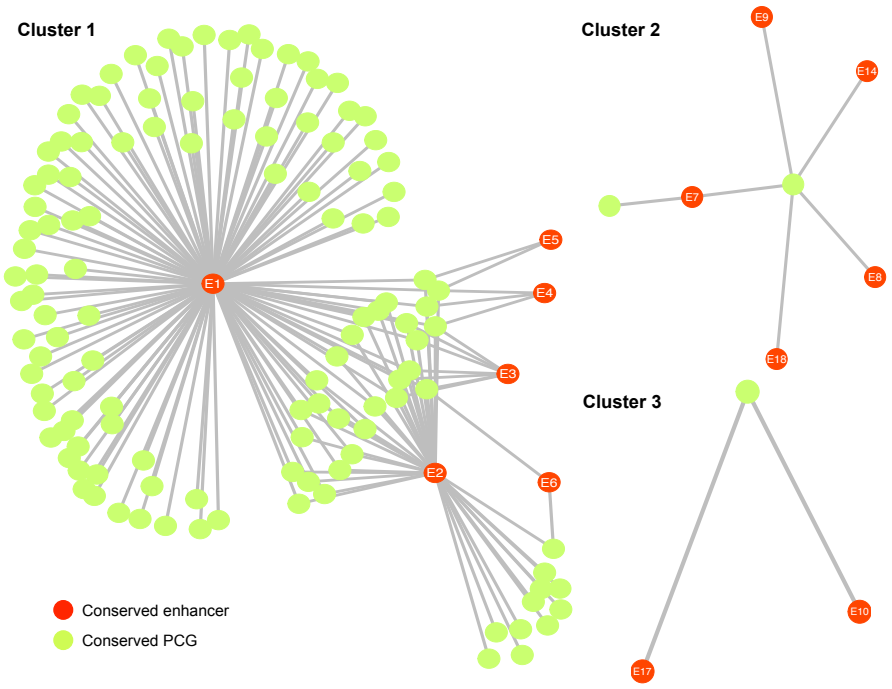


Figure 4-11: Three clusters of conserved enhancer-PCG promoter pairs. The networks were visualized using ggplot2 [92].

Chapter 5 Conclusions and Discussion

5.1 Conclusions

A long-standing question in genomics is to identify the functional noncoding regions and further understand their biological functions in mammalian genomes. Here we applied a novel computational approach to FANTOM5 data to systematically identify cell-state-associated *cis*-regulatory elements for more than 300 cell and tissue types from human and mouse. We first applied a modified version of t-test to find associated *cis*-regulatory elements for each cell and tissue type. Our method identified strong immune cell-specific enrichment for their associated enhancers. We found that the enhancers associated with more cell and tissue types were longer in length. Furthermore, all the three types of *cis*-regulatory elements that are associated with more cell and tissue types exhibit higher conservation scores and greater enrichment of CpG-islands.

The enriched biological functions of the associated PCG promoters confirmed previously knowledge. In addition, we identified enriched TF motifs for the associated enhancers, providing insights into their regulatory circuits. Furthermore, we found that GWAS SNPs are particularly enriched in the cell-state-associated enhancers, and analyzed the association between human diseases and tissue types. Various sclerosis diseases were associated with diverse immune-associated tissues and mature immune cells. Finally, we inferred enhancer-promoter signal dependency and identified multiple enhancers with conserved putative relationships with promoters between human and mouse.

To the best of our knowledge, this is the first work that comprehensively identifies *cis*-regulatory elements associated with various cell differentiation states in human and mouse. We

anticipate that these *cis*-regulatory elements are valuable candidates for further experimental studies.

5.2 Future directions

Previous studies suggest that eRNA transcription is a regulated process and not transcriptional noise [93,94]. The eRNA-producing enhancers are actively engaged in promoting the expression of their target genes, which are generally located near the enhancers [24,93]. Although determining enhancer targets is difficult, some computational methods were developed to reconstruct enhancer-target networks [29,30,32,95-99]. These methods use experimental datasets of epigenomes and TF binding to infer enhancer-target networks in a cell type-specific manner. However, in our current work, the predicted enhancer-promoter correlation is not cell type-specific. Thus, one main goal for future research is to expend the current model to enable prediction of enhancer-promoter correlation in a cell type-specific manner. In addition, because one gene can be targeted by multiple enhancers, considering each enhancer independently could miss some important enhancer-promoter interactions.

The systematically reconstructed enhancer-promoter interactions can be used to study gene expression regulation in both normal and disease states on a large scale. In our current work, we performed a preliminary analysis on the association between genetic diseases and *cis*-regulatory elements. Next step, we may improve this analysis in the context of enhancer-promoter interactions to identify genes potentially affected by perturbed enhancers [100]. Currently, most cancer genomic studies focused on identifying cancer genes based on frequently somatic mutations and indels or differential gene expression for protein-coding regions in cancer. Aberrant *cis*-regulatory elements in cancer are poorly characterized and understood [40,101-107].

Ongoing efforts will greatly advance our understanding of genomic mutations in these noncoding *cis*-regulatory elements.

References

1 Heintzman, N. D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108-112, doi:10.1038/nature07829 (2009).

2 Rada-Iglesias, A. *et al.* A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**, 279-283, doi:10.1038/nature09692 (2011).

3 Rivera, C. M. & Ren, B. Mapping human epigenomes. *Cell* **155**, 39-55, doi:10.1016/j.cell.2013.09.011 (2013).

4 Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934-947, doi:10.1016/j.cell.2013.09.053 (2013).

5 Whyte, W. A. *et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307-319, doi:10.1016/j.cell.2013.03.035 (2013).

6 Mathelier, A., Shi, W. & Wasserman, W. W. Identification of altered cis-regulatory elements in human disease. *Trends Genet* **31**, 67-76, doi:10.1016/j.tig.2014.12.003 (2015).

7 Stunnenberg, H. G., International Human Epigenome Consortium & Hirst, M. The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. *Cell* **167**, 1145-1149, doi:10.1016/j.cell.2016.11.007 (2016).

8 Shen, Y. *et al.* A map of the cis-regulatory sequences in the mouse genome. *Nature* **488**, 116-120, doi:10.1038/nature11243 (2012).

9 Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330, doi:10.1038/nature14248 (2015).

10 Davydov, E. V. *et al.* Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol* **6**, e1001025, doi:10.1371/journal.pcbi.1001025 (2010).

11 Encode Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74, doi:10.1038/nature11247 (2012).

12 Buecker, C. & Wysocka, J. Enhancers as information integration hubs in development: lessons from genomics. *Trends Genet* **28**, 276-284, doi:10.1016/j.tig.2012.02.008 (2012).

13 Pott, S. & Lieb, J. D. What are super-enhancers? *Nat Genet* **47**, 8-12, doi:10.1038/ng.3167 (2015).

14 Li, W. V., Razaee, Z. S. & Li, J. J. Epigenome overlap measure (EPOM) for comparing tissue/cell types based on chromatin states. *BMC Genomics* **17 Suppl 1**, 10, doi:10.1186/s12864-015-2303-9 (2016).

15 Kim, T. K., Hemberg, M. & Gray, J. M. Enhancer RNAs: a class of long noncoding RNAs synthesized at enhancers. *Cold Spring Harb Perspect Biol* **7**, a018622, doi:10.1101/cshperspect.a018622 (2015).

16 Li, W., Notani, D. & Rosenfeld, M. G. Enhancers as non-coding RNA transcription units: recent insights and future perspectives. *Nat Rev Genet* **17**, 207-223, doi:10.1038/nrg.2016.4 (2016).

17 Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455-461, doi:10.1038/nature12787 (2014).

18 Hon, C. C. *et al.* An atlas of human long non-coding RNAs with accurate 5' ends. *Nature* **543**, 199-204, doi:10.1038/nature21374 (2017).

19 The FANTOM Consortium and the RIKEN PMI and CLST (DGT). A promoter-level mammalian expression atlas. *Nature* **507**, 462-470, doi:10.1038/nature13182 (2014).

20 Bonev, B. & Cavalli, G. Organization and function of the 3D genome. *Nat Rev Genet* **17**, 661-678, doi:10.1038/nrg.2016.112 (2016).

21 Dekker, J., Marti-Renom, M. A. & Mirny, L. A. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat Rev Genet* **14**, 390-403, doi:10.1038/nrg3454 (2013).

22 Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-380, doi:10.1038/nature11082 (2012).

23 Schmitt, A. D. *et al.* A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. *Cell Rep* **17**, 2042-2059, doi:10.1016/j.celrep.2016.10.061 (2016).

24 Sanyal, A., Lajoie, B. R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* **489**, 109-113, doi:10.1038/nature11279 (2012).

25 Mifsud, B. *et al.* Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat Genet* **47**, 598-606, doi:10.1038/ng.3286 (2015).

26 Deng, W. *et al.* Reactivation of developmentally silenced globin genes by forced chromatin looping. *Cell* **158**, 849-860, doi:10.1016/j.cell.2014.05.050 (2014).

27 Heinz, S., Romanoski, C. E., Benner, C. & Glass, C. K. The selection and function of cell type-specific enhancers. *Nat Rev Mol Cell Biol* **16**, 144-154, doi:10.1038/nrm3949 (2015).

28 Mora, A., Sandve, G. K., Gabrielsen, O. S. & Eskeland, R. In the loop: promoter-enhancer interactions and bioinformatics. *Brief Bioinform* **17**, 980-995, doi:10.1093/bib/bbv097 (2016).

29 Roy, S. *et al.* A predictive modeling approach for cell line-specific long-range regulatory interactions. *Nucleic Acids Res* **43**, 8694-8712, doi:10.1093/nar/gkv865 (2015).

30 Whalen, S., Truty, R. M. & Pollard, K. S. Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nat Genet* **48**, 488-496, doi:10.1038/ng.3539 (2016).

31 Gerstein, M. B. *et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**, 91-100, doi:10.1038/nature11245 (2012).

32 Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75-82, doi:10.1038/nature11232 (2012).

33 Pennacchio, L. A., Bickmore, W., Dean, A., Nobrega, M. A. & Bejerano, G. Enhancers: five essential questions. *Nat Rev Genet* **14**, 288-295, doi:10.1038/nrg3458 (2013).

34 Stenson, P. D. *et al.* The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum Genet* **133**, 1-9, doi:10.1007/s00439-013-1358-4 (2014).

35 Visel, A., Rubin, E. M. & Pennacchio, L. A. Genomic views of distant-acting enhancers. *Nature* **461**, 199-205, doi:10.1038/nature08451 (2009).

36 Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190-1195, doi:10.1126/science.1222794 (2012).

37 Lee, T. I. & Young, R. A. Transcriptional regulation and its misregulation in disease. *Cell* **152**, 1237-1251, doi:10.1016/j.cell.2013.02.014 (2013).

38 Sur, I., Tuupanen, S., Whitington, T., Aaltonen, L. A. & Taipale, J. Lessons from functional analysis of genome-wide association studies. *Cancer Res* **73**, 4180-4184, doi:10.1158/0008-5472.CAN-13-0789 (2013).

39 Ward, L. D. & Kellis, M. Interpreting noncoding genetic variation in complex traits and human disease. *Nat Biotechnol* **30**, 1095-1106, doi:10.1038/nbt.2422 (2012).

40 Khurana, E. *et al.* Role of non-coding sequence variants in cancer. *Nat Rev Genet* **17**, 93-108, doi:10.1038/nrg.2015.17 (2016).

41 Boyle, A. P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res* **22**, 1790-1797, doi:10.1101/gr.137323.112 (2012).

42 Ward, L. D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res* **40**, D930-934, doi:10.1093/nar/gkr917 (2012).

43 Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310-315, doi:10.1038/ng.2892 (2014).

44 Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Methods* **12**, 931-934, doi:10.1038/nmeth.3547 (2015).

45 Moreau, Y. & Tranchevent, L. C. Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nat Rev Genet* **13**, 523-536, doi:10.1038/nrg3253 (2012).

46 Kuhn, R. M., Haussler, D. & Kent, W. J. The UCSC genome browser and associated tools. *Brief Bioinform* **14**, 144-161, doi:10.1093/bib/bbs038 (2013).

47 Bock, C., Walter, J., Paulsen, M. & Lengauer, T. CpG island mapping by epigenome prediction. *PLoS Comput Biol* **3**, e110, doi:10.1371/journal.pcbi.0030110 (2007).

48 Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**, 1034-1050, doi:10.1101/gr.3715005 (2005).

49 Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017-1018, doi:10.1093/bioinformatics/btr064 (2011).

50 Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L. & Noble, W. S. Quantifying similarity between motifs. *Genome Biol* **8**, R24, doi:10.1186/gb-2007-8-2-r24 (2007).

51 Mathelier, A. *et al.* JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* **44**, D110-115, doi:10.1093/nar/gkv1176 (2016).

52 Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* **37**, W202-208, doi:10.1093/nar/gkp335 (2009).

53 Li, M. J. *et al.* GWASdb v2: an update database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Res* **44**, D869-876, doi:10.1093/nar/gkv1317 (2016).

54 Schreiber, F., Patricio, M., Muffato, M., Pignatelli, M. & Bateman, A. TreeFam v9: a new website, more species and orthology-on-the-fly. *Nucleic Acids Res* **42**, D922-925, doi:10.1093/nar/gkt1055 (2014).

55 Tyner, C. *et al.* The UCSC Genome Browser database: 2017 update. *Nucleic Acids Res* **45**, D626-D634, doi:10.1093/nar/gkw1134 (2017).

56 van Dongen, S. & Abreu-Goodger, C. Using MCL to extract clusters from networks. *Methods Mol Biol* **804**, 281-295, doi:10.1007/978-1-61779-361-5_15 (2012).

57 Yang, Y., Yang, Y. T., Yuan, J., Lu, Z. J. & Li, J. J. Large-scale mapping of mammalian transcriptomes identifies conserved genes associated with different cell states. *Nucleic Acids Res* **45**, 1657-1672, doi:10.1093/nar/gkw1256 (2017).

58 Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* **47**, 1228-1235, doi:10.1038/ng.3404 (2015).

59 Cabili, M. N. *et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev* **25**, 1915-1927, doi:10.1101/gad.17446611 (2011).

60 Deaton, A. M. & Bird, A. CpG islands and the regulation of transcription. *Genes Dev* **25**, 1010-1022, doi:10.1101/gad.2037511 (2011).

61 Weber, M. *et al.* Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat Genet* **39**, 457-466, doi:10.1038/ng1990 (2007).

62 Ganeshan, K. & Chawla, A. Metabolic regulation of immune responses. *Annu Rev Immunol* **32**, 609-634, doi:10.1146/annurev-immunol-032713-120236 (2014).

63 Pearce, E. L. & Pearce, E. J. Metabolic pathways in immune cell activation and quiescence. *Immunity* **38**, 633-643, doi:10.1016/j.immuni.2013.04.005 (2013).

64 Ankrum, J. A., Ong, J. F. & Karp, J. M. Mesenchymal stem cells: immune evasive, not immune privileged. *Nat Biotechnol* **32**, 252-260, doi:10.1038/nbt.2816 (2014).

65 Sotiropoulou, P. A. & Papamichail, M. Immune properties of mesenchymal stem cells. *Methods Mol Biol* **407**, 225-243, doi:10.1007/978-1-59745-536-7_16 (2007).

66 Arnold, S. J. & Robertson, E. J. Making a commitment: cell lineage allocation and axis patterning in the early mouse embryo. *Nat Rev Mol Cell Biol* **10**, 91-103, doi:10.1038/nrm2618 (2009).

67 Bedzhov, I., Graham, S. J., Leung, C. Y. & Zernicka-Goetz, M. Developmental plasticity, cell fate specification and morphogenesis in the early mouse embryo. *Philos Trans R Soc Lond B Biol Sci* **369**, doi:10.1098/rstb.2013.0538 (2014).

68 Sur, I. & Taipale, J. The role of enhancers in cancer. *Nat Rev Cancer* **16**, 483-493, doi:10.1038/nrc.2016.62 (2016).

69 Spitz, F. & Furlong, E. E. Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet* **13**, 613-626, doi:10.1038/nrg3207 (2012).

70 Iida, K. *et al.* Expression of MEF2 genes during human cardiac development. *Tohoku J Exp Med* **187**, 15-23 (1999).

71 Snykers, S. *et al.* Role of epigenetics in liver-specific gene transcription, hepatocyte differentiation and stem cell reprogrammation. *J Hepatol* **51**, 187-211, doi:10.1016/j.jhep.2009.03.009 (2009).

72 Hou, Y. *et al.* Nuclear factor-Y (NF-Y) regulates transcription of mouse Dmrt7 gene by binding to tandem CCAAT boxes in its proximal promoter. *Int J Biol Sci* **6**, 655-664 (2010).

73 Zhao, G. N., Jiang, D. S. & Li, H. Interferon regulatory factors: at the crossroads of immunity, metabolism, and disease. *Biochim Biophys Acta* **1852**, 365-378, doi:10.1016/j.bbadis.2014.04.030 (2015).

74 Gururajan, M. *et al.* Early growth response genes regulate B cell development, proliferation, and immune response. *J Immunol* **181**, 4590-4602 (2008).

75 Hu, H. *et al.* Foxp1 is an essential transcriptional regulator of B cell development. *Nat Immunol* **7**, 819-826, doi:10.1038/ni1358 (2006).

76 Dupuis-Maurin, V. *et al.* Overexpression of the transcription factor Sp1 activates the OAS-RNAse L-RIG-I pathway. *PLoS One* **10**, e0118551, doi:10.1371/journal.pone.0118551 (2015).

77 Flajollet, S., Poras, I., Carosella, E. D. & Moreau, P. RREB-1 is a transcriptional repressor of HLA-G. *J Immunol* **183**, 6948-6959, doi:10.4049/jimmunol.0902053 (2009).

78 Kaczynski, J., Cook, T. & Urrutia, R. Sp1- and Kruppel-like transcription factors. *Genome Biol* **4**, 206 (2003).

79 Marjoram, P., Zubair, A. & Nuzhdin, S. V. Post-GWAS: where next? More samples, more SNPs or more biology? *Heredity (Edinb)* **112**, 79-88, doi:10.1038/hdy.2013.52 (2014).

80 Gibson, G. Rare and common variants: twenty arguments. *Nat Rev Genet* **13**, 135-145, doi:10.1038/nrg3118 (2012).

81 Bush, W. S. & Moore, J. H. Chapter 11: Genome-wide association studies. *PLoS Comput Biol* **8**, e1002822, doi:10.1371/journal.pcbi.1002822 (2012).

82 Elkon, R. & Agami, R. Characterization of noncoding regulatory DNA in the human genome. *Nat Biotechnol* **35**, 732-746, doi:10.1038/nbt.3863 (2017).

83 Edwards, S. L., Beesley, J., French, J. D. & Dunning, A. M. Beyond GWASs: illuminating the dark road from association to function. *Am J Hum Genet* **93**, 779-797, doi:10.1016/j.ajhg.2013.10.012 (2013).

84 MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res* **45**, D896-D901, doi:10.1093/nar/gkw1133 (2017).

85 Chen, K. & Kolls, J. K. T cell-mediated host immune defenses in the lung. *Annu Rev Immunol* **31**, 605-633, doi:10.1146/annurev-immunol-032712-100019 (2013).

86 Nistico, G., Caroleo, M. C., Arbitrio, M. & Pulvirenti, L. Dopamine D1 receptors in the amygdala enhance the immune response in the rat. *Ann N Y Acad Sci* **741**, 316-323 (1994).

87 Kurts, C., Panzer, U., Anders, H. J. & Rees, A. J. The immune system and kidney disease: basic concepts and clinical implications. *Nat Rev Immunol* **13**, 738-753, doi:10.1038/nri3523 (2013).

88 Newton, J. L. *et al.* Cognitive impairment in primary biliary cirrhosis: symptom impact and potential etiology. *Hepatology* **48**, 541-549, doi:10.1002/hep.22371 (2008).

89 Grover, V. P. *et al.* Early primary biliary cholangitis is characterised by brain abnormalities on cerebral magnetic resonance imaging. *Aliment Pharmacol Ther* **44**, 936-945, doi:10.1111/apt.13797 (2016).

90 Williams, A., Spilianakis, C. G. & Flavell, R. A. Interchromosomal association and gene regulation in trans. *Trends Genet* **26**, 188-197, doi:10.1016/j.tig.2010.01.007 (2010).

91 Spilianakis, C. G., Lalioti, M. D., Town, T., Lee, G. R. & Flavell, R. A. Interchromosomal associations between alternatively expressed loci. *Nature* **435**, 637-645, doi:10.1038/nature03574 (2005).

92 Wickham, H. *Ggplot2 : elegant graphics for data analysis*. (Springer, 2009).

93 Kim, T. K. *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**, 182-187, doi:10.1038/nature09033 (2010).

94 Lai, F. *et al.* Activating RNAs associate with Mediator to enhance chromatin architecture and transcription. *Nature* **494**, 497-501, doi:10.1038/nature11884 (2013).

95 Cao, Q. *et al.* Reconstruction of enhancer-target networks in 935 samples of human primary cells, tissues and cell lines. *Nat Genet* **49**, 1428-1436, doi:10.1038/ng.3950 (2017).

96 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, doi:10.1101/gr.164079.113 (2014).

97 Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43-49, doi:10.1038/nature09906 (2011).

98 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**, E2191-2199, doi:10.1073/pnas.1320308111 (2014).

99 Zhu, Y. *et al.* Constructing 3D interaction maps from 1D epigenomes. *Nat Commun* **7**, 10812, doi:10.1038/ncomms10812 (2016).

100 Mumbach, M. R. *et al.* Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. *Nat Genet*, doi:10.1038/ng.3963 (2017).

101 Vinagre, J. *et al.* Frequency of TERT promoter mutations in human cancers. *Nat Commun* **4**, 2185, doi:10.1038/ncomms3185 (2013).

102 Kim, K. *et al.* Chromatin structure-based prediction of recurrent noncoding mutations in cancer. *Nat Genet* **48**, 1321-1326, doi:10.1038/ng.3682 (2016).

103 Araya, C. L. *et al.* Identification of significantly mutated regions across cancer types highlights a rich landscape of functional molecular alterations. *Nat Genet* **48**, 117-125, doi:10.1038/ng.3471 (2016).

104 Juul, M. *et al.* Non-coding cancer driver candidates identified with a sample- and position-specific model of the somatic mutation rate. *Elife* **6**, doi:10.7554/eLife.21778 (2017).

105 Kumar, S. & Gerstein, M. Cancer genomics: Less is more in the hunt for driver mutations. *Nature* **547**, 40-41, doi:10.1038/nature23085 (2017).

106 Rheinbay, E. *et al.* Recurrent and functional regulatory mutations in breast cancer. *Nature* **547**, 55-60, doi:10.1038/nature22992 (2017).

107 Zehir, A. *et al.* Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nat Med* **23**, 703-713, doi:10.1038/nm.4333 (2017).