# IDENTIFICATION OF LONG-RANGE REGULATORY ELEMENTS IN THE HUMAN GENOME

Yih-Chii Hwang

A DISSERTATION

in

Genomics and Computational Biology

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2015

Supervisor of Dissertation:

_____

Li-San Wang, Ph. D.
Associate Professor of Pathology
and Laboratory Medicine

Co-Supervisor of Dissertation:

_____

Brian D. Gregory, Ph. D.
Assistant Professor of Biology

Graduate Group Chairperson:

_____

Li-San Wang, Ph. D.,
Associate Professor of Pathology and Laboratory Medicine

Dissertation Committee:

Doris Wagner, Ph.D., Professor of Biology

Uwe Ohler, Ph.D., Professor of Biology, Max Delbrueck Center, Berlin

Nancy Zhang, Ph.D., Associate Professor of Statistics

Gerard D. Schellenberg, Ph.D., Professor of Pathology and Laboratory Medicine

IDENTIFICATION OF LONG-RANGE REGULATORY ELEMENTS IN THE HUMAN GENOME

# ACKNOWLEDGMENT

First, I would like to extend my deepest gratitude to both of my advisors, Li-San Wang and Brian D. Gregory for their support and supervision. Furthermore, having the guidance from both of them has broadened my horizons in the field of science, transformed me into a more mature person, and made my past six and a half years such a memorable experience.

I would like to thank my thesis committee members: Doris Wagner, Uwe Ohler, Gerard D. Schellenberg, Nancy Zhang, for their guidance and encouragement throughout all the years.

I am grateful for the lab members from both the Wang and the Gregory labs. They are not only good colleagues but also good friends. In particular, I would like to thank Yuk Yee (Fanny) Leung, Pavel Kuksa, Chiao-Feng Lin, Kajia Cao, Fan Li, Qi Zheng, Matthew Willmann, Ian Silverman, Nathan Berkowitz, Lee Vandivier, Alexandre Amlie-Wolf, and Otto Valladares for their camaraderie and invaluable scientific conversations.

I would like to acknowledge my GCB fellows, especially Ellen Tsai, Jun Chen, Scott Sherrill-Mix, Ying Chen, Hannah Dueck, Sarah Middleton, Zhang (Eric) Chen, and Yuchao Jiang for the classes we attended together and their sharing this precious graduate school experience at Penn with me. I would like to thank the staffs at GCB, Hannah Chervitz, Tiffany Barlow, and Maureen Kirsch for their management of this program. . A special thank is given to Rebecca Cweibel and Postdoctoral Editors Association for helping me on editing my thesis writings.

My spectacular appreciation goes to my boyfriend, Hsin-Ta Wu, for always being my best friend and his patience with me through the ups and downs in my life.

Last but not the least, I owed a great debt of gratitude to my family: my mother, father, and my brother, for supporting me spiritually throughout this PhD journey and my life in general.

# ABSTRACT

IDENTIFICATION OF LONG-RANGE REGULATORY ELEMENTS IN THE HUMAN GENOME

Yih-Chii Hwang

Li-San Wang

Brian D. Gregory

Genome-wide association studies have shown that the majority of disease-associated genetic variants lie within non-coding regions of the human genome. Subsequently, a challenge following these discoveries is to identify how these variants modulate the risk of disease. Enhancers are non-coding regulatory elements that can be bound by proteins to activate the expression of a gene that may be linearly distant. Experimentally probing all possible enhancer–target gene pairs can be laborious. Hi-C, a technique developed by Job Dekker's group in 2009, combines high-throughput sequencing with chromosome conformation capture to detect DNA interactions genome-wide and thereby reveals the three-dimensional architecture of chromatin in the nucleus. However, the utility of the datasets produced by this technique for discovering long-range regulatory interactions is largely unexplored.

In this thesis, we develop novel approaches to identify DNA-interacting units and their interactions in Hi-C datasets with the goal of uncovering all enhancer–target gene interactions.

We began by identifying significantly interacting regions in these datasets, subsequently focusing on candidate enhancer–gene pairs. We found that the identified putative enhancers are enriched for p300 binding activity, while their target promoters are likely to be cell-type-specific. Furthermore, we revealed that enhancers and target genes often interact in many-to-many

relationships and the majority of enhancer–target gene interactions are intra-chromosomal and within 1 Mb of each other.

Next, we refined our analytical approach to identify physically-interacting DNA regions at ~1 kb resolution and better define the boundaries of likely enhancer elements. By searching for over-represented sequences (motifs) in these putative promoter-interacting enhancers, we were then able to identify bound transcription factors. This newer approach provides the potential to identify protein complexes involved in enhancer–promoter interactions, which can be verified in future experiments.

We implemented a high-throughput identification pipeline for promoter-interacting enhancer elements (HIPPIE) using both of the above described approaches. HIPPIE can be run efficiently on typical Linux servers and grid computing environments and is available as open-source software. In summary, our findings demonstrate the potential utility of Hi-C technologies for elucidating the mechanisms by which long-range enhancers regulate gene expression and ultimately result in human disease phenotypes.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF ILLUSTRATIONS

# Chapter 1 : Introduction

## 1.1 Gene regulation in eukaryotic genomes

### 1.1.1 Transcriptional regulation of gene expression

In 1956, Francis Crick coined the idea of “the central dogma of molecular biology” to describe the general flow of biological information through the three major biomolecules: DNA, RNA, and proteins (Crick, 1956, 1970). Through transcription, information along chromosomal DNA gets copied into messenger RNA (mRNA), and through translation, proteins are synthesized using the mRNA as a template. The entire transcriptional process to generate mRNA molecules is regulated by a group of proteins known as transcription factors. More specifically, these transcription factors can recognize the DNA sequences encoded within the chromosomes and control the level of mRNA expression. This transcription regulation is involved in many fundamental aspects of biology, including embryonic development, cellular differentiation, and cell fate. In other words, gene regulation processes control what each cell in the organism is doing, how a tissue functions, and how organisms survive and reproduce.

### 1.1.2 The non-coding human genome

Interestingly, less than 2% of the human genome actually encodes protein sequences (Elgar and Vavouri, 2008). Thus, it was not surprising that as a follow-up to the Human Genome Project, the Encyclopedia of DNA Elements (ENCODE) project was launched in 2003 and aimed to identify all “functional elements” in the human genome using a broad range of experimental assays to study the functions of non-coding genomes such as chromatin immunoprecipitation, DNase hypersensivity profiling, cap analysis gene expression for localization of transcription start sites, and temporal profiling of DNA replication sites. In 2012, ENCODE reported that over 80% of

the human genome has biochemical functions (Bernstein et al., 2012), and many (95%) of these are expressed, non-protein-coding regions that may function to regulate the expression of protein-coding loci.

### 1.1.3 Non-coding genetic variants and their functional impact

In recent years, considerable progress in genome-wide association studies (GWAS) and the development of high-throughput sequencing technologies have allowed the discovery of an incredible collection of genetic variants within the human population, many of which are significantly correlated with human disease phenotypes. These discoveries of disease-risk loci have brought genetic researchers toward the post-GWAS era where there are more unanswered questions, such as how a non-protein-coding locus affects cellular fate or influences a phenotype, than there are answered ones.

To date, it has been shown that non-coding DNA regions can produce functional types of RNA molecules. These include transfer RNAs (tRNAs), which are involved in translation of mRNA, Piwi-interacting RNAs (piRNAs), which are linked to post-transcriptional gene silencing in animal germ line cells, and microRNAs (miRNAs), which function in the post-transcriptional regulation of gene expression and/or translation repression.

In addition to being transcribed into functional RNA molecules, non-coding DNA regions can also regulate the transcriptional and/or translational activities of protein-coding genes via different mechanisms. Those DNA regions, called regulatory elements, function as footprints of transcription factors and thereby regulate mRNA expression and in turn, protein production. In eukaryotic organisms, typical types of regulatory elements include promoters, enhancers, and insulators.

### 1.1.4 Enhancer elements: DNA regulatory sequences that control distal gene expression

In vertebrate and mammalian genomes, protein-coding genes are not only regulated by specific sequence elements proximal (near) to their transcription start sites or promoter regions, but also by distal (far) and orientation-independent elements such as enhancers. Of the types of regulatory elements noted above, enhancers are considered to be long-range regulatory elements because they can be located millions of nucleotides (nts) away from the gene(s) they regulate and even reside on a different chromosome (Banerji et al., 1981; Geyer et al., 1990; Lomvardas et al., 2006). Enhancers typically range in size from ~100 base pairs (bp) (Catena et al., 2004) to several kilobases (kb) (Chi et al., 2005), with an average length of 500 bp (Marshall et al., 2001).

In many eukaryotic genomes, chromatin looping between regulatory elements that are distant from one another along the chromosome is widely observed and appears to be a general mechanism for establishing long-range functional interactions. Enhancers can activate transcription of target genes that are located far away through binding of a protein complex that facilitates the formation of loops between the enhancer and these genes. However, this kind of long-range regulatory element interaction is extremely rare in yeast systems. For instance, the GAL4 activator can only enhance gene expression when the DNA binding site is within a few hundred bp of the regulated locus (Guarente and Hoar, 1984; Struhl, 1984).

It is believed that enhancer elements mostly function to regulate gene expression in a tissue-specific manner and thus effect different sets of genes in different tissues (Visel et al., 2009). These regulatory events can also contribute to the determination of cell fate, the potential for differentiation into new cell types throughout development, and the maintenance of specific gene expression programs through epigenetic modification of the genome (Creyghton et al., 2010). In this way, enhancers are thought to be central regulators of gene expression during

animal development (Levine, 2010; Wilber et al., 2011). Therefore, discovering long-range regulatory elements genome-wide is necessary to fully understand gene expression regulation in animals as well as during their development.

### 1.1.5 Enhancer elements: DNA regulatory sequences that affect phenotype by controlling distal gene expression

An example where an enhancer element variant affects a human phenotype can be found within the enhancer that regulates the sonic hedgehog gene (*Shh*) (Lettice et al., 2003). This variant disrupts the binding of a transcription factor to an enhancer element that regulates *Shh*, resulting in limb malformation. More specifically, it has been demonstrated that a genetic variant located in the 5$^{th}$ intron of the *Lmbr1* gene does not affect expression of this gene, but affects the *Shh* locus located 1 megabase (Mb) away. Thus, these variants in the *Lmbr1* intron were recognized to affect an enhancer element of *Shh*, a gene whose tight regulation is necessary for proper limb development in mammals (**Figure 1-1**). This example suggests that an enhancer, working as a *trans*-regulatory element, may not necessarily regulate the protein-coding gene containing it or located nearby, but may instead control the expression of a gene that is distant along the chromosome and thereby affect a phenotype change (for more discussion, please see **Section 1.4.1**).

(a)


[Lettice et al., *HMG*. 2003]

(b)

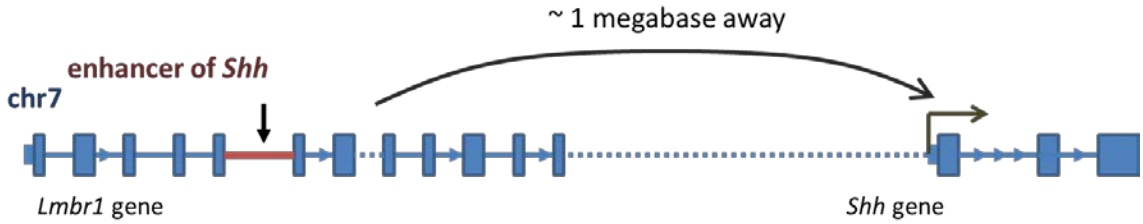



Figure 1-1. An example of genetic variants in an enhancer element affecting a human phenotype.

(a) A single nucleotide variant can cause paraxial polydactyly. (b) The *Shh* enhancer and variant are located in the 5$^{th}$ intron of the *Lmbr1* gene (~1 Mb away along the chromosome).

### 1.1.6 Chromatin signatures of enhancer elements

It has recently been demonstrated that chromatin modifications can be regarded as indicators of the transcriptional regulatory function and activity of certain types of genomic loci. It became possible to describe the chromatin architecture of regulatory elements using the molecular biology technique known as chromatin immunoprecipitation (ChIP) with antibodies that directly detect specific covalent modifications of histone proteins. For example, this approach revealed that monomethylated histone 3 lysine 4 (H3K4me1) is spanning a broader region and strongly enriched at nearly all active enhancer elements, and occurrences of these H3K4me1 marks correlate with the transcriptional activity of nearby genes (Heintzman et al., 2007). In addition, it has been suggested that acetylation of histone H3 lysine 27 (H3K27ac) is an important enhancer mark that can be used to distinguish whether an enhancer element is at the state of being active or being poised (predetermined). Thus, it is not surprising that genomic locations with H3K27ac enrichment are highly diverse between various cell types in humans (Creyghton et al., 2010). With the development of high-throughput sequencing technology, the use of ChIP followed by sequencing (ChIP-seq) (Johnson et al., 2007) has become highly prevalent and this approach has made it possible to survey chromatin modifications on a genome-wide scale (Barski et al., 2007). In December 2008, ENCODE started to integrate and aggregate ChIP-seq datasets of histone modifications (e. g. H3K4me1, H3K27ac) for a number of commonly studied human cell lines including lymphoblastoid cells (GM12878), chronic myelogenous leukemia cells (K562), and many others. This provides genomic researchers with the resources to identify and explore putative enhancer elements on a genome-wide scale.

An additional genomics technique that has been used for regulatory element identification is the mapping of DNase I hypersensitive (DHS) sites, a method that reveals chromatin regions of eukaryotic genomes that are open and accessible to DNA-binding transcription factors (Gross and Garrard, 1988; Wu et al., 1979). Again, by combining DHS mapping with high-throughput sequencing (Crawford et al., 2006), it has become possible to survey open chromatin regions on

a genome-wide scale. In addition, formaldehyde-assisted identification of regulatory elements followed by high-throughput sequencing (FAIRE–seq) can identify open chromatin regions by depleting histone-bound 'closed' DNA after being chemical crosslinking with formaldehyde (Giresi et al., 2007). Thus, previous studies have used these approaches to identify active enhancers, defined by open chromatin regions that are distal to protein-coding genes, and active promoters, defined by open chromatin regions that are proximal to the transcription start sites of protein-coding gene regions (Thurman et al., 2012). In total, genome-wide datasets of chromatin modifications and open chromatin states are important resources for predicting active enhancer elements (Ernst et al., 2011; Hoffman et al., 2012).

## 1.2 Experimental approaches to identify functional regulatory elements

### 1.2.1. Luciferase reporter assay

To determine whether a DNA element of interest displays regulatory activity, one can clone and fuse the putative regulatory element to a luciferase reporter gene (McNabb et al., 2005). Because the reporter construct contains a promoter with minimal activity (e.g. SV40), reporter expression is directly correlated with the activity of the included regulatory element. Using this reporter gene assay, one can estimate the strength of the regulatory element. However, one cannot use this technique to identify the specific promoters that the regulatory element activates in the cell.

### 1.2.2. Creating element knockouts by targetable DNA cleavage engineering

To identify the function of a locus, one can make mutations (often times a deletion) specifically in that locus, and observe the resulting gene expression change or phenotype. This strategy is known as reverse genetics. Below, I review two popular genetic engineering approaches to study the function of specific genetic loci.

***Zinc-Finger Nucleases (ZFN)***

ZFNs originated from the observation by Li et al. (Li et al., 1992) that the natural type IIS restriction enzyme *Fok*I has physically separable DNA-binding and cleaving activities. The cleavage could be redirected to other DNA sequences by substituting alternative DNA-recognition domains for the natural one. The most useful DNA-binding domain that has been combined with the *Fok*I DNA cleavage activity consists of three $Cys_2His_2$ zinc fingers (ZFs), each of which can recognize 3 bp of DNA in a modular fashion (Pavletich and Pabo, 1991). This remarkable modularity of ZFs suggested that many different combinations of ZFs can be assembled that would recognize different DNA sequences for cleavage. Adding more fingers can improve specificity, but there is also the possibility that fingers in a polydactyl domain will bind to off-target sites. There are also specificity challenges for ZF binding since some fingers can bind equally well to any triplet nt sequence (Carroll, 2011). In total, it can be challenging to design and validate ZF proteins for specific DNA locus binding.

***CRISPR-Cas9***

In selected bacteria and archaea, the functions of Clustered Regularly Interspaced Short Palindromic Repeats (CRISPRs) and CRISPR-associated (Cas) genes are essential for adaptive immunity (Mojica et al., 2000). More specifically, this is an essential mechanism by which these organisms respond to invading genetic material (e.g. bacteriophage genomes). The CRISPR-Cas9 system consists the Cas9 DNA endonuclease and a single-stranded guide RNA (gRNA) to

which it binds. The gRNA anneals to target DNA sequences and thereby can be used to direct Cas9's endonuclease activity to any desired site in the genome. Recently, using engineered gRNAs, this system has been found to be an efficient tool that can successfully manipulate and edit the mouse and human genomes with site-specific changes (Cong et al., 2013; Doudna and Charpentier, 2014; Jinek et al., 2012; Mali et al., 2013).

Overall, ZFN and CRISPR-Cas9 are useful tools for knocking down and thereby discovering the function of DNA regulatory elements. However, these approaches will not reveal the target gene(s) of these regulatory elements.

## 1.3 Experimental technology to determine the physical interactions of DNA regions

### 1.3.1 Traditional and small scale methods — chromosome conformation capture and fluorescence in situ hybridization

Chromosome conformation capture (3C) and fluorescence *in situ* hybridization (FISH) are the two major molecular approaches for studying genome organization and nucleus compartmentalization. These strategies can successfully identify long-range DNA interactions and estimate the likelihood that two DNA regions along the genome interact with each other in the three-dimensional space of a nucleus. In the next few paragraphs, we highlight these low-throughput but high-quality methods as well as the insights gained from each of them.

***Chromosome Conformation Capture (3C)***

3C can be used to detect whether a pair of DNA fragments are interacting via a protein complex and does so by stabilizing this interaction through formaldehyde crosslinking. This

crosslinking is followed by digesting the genomic DNA with a restriction enzyme and then performing a proximity ligation reaction between the two interacting DNA regions. To detect if a pair of DNA regions are interacting, PCR primers are designed to amplify the ligation products of the suspected interacting regions. The first use of this protocol was to demonstrate that the yeast third chromosome forms a 3D contorted ring structure (Dekker et al., 2002). When performing a 3C experiment, it is important to perform internal control experiments to address possible biases such as PCR efficiency differences between primer pairs, assess the level of random background interactions, and properly normalize data (Dekker, 2006).

***Fluorescent in situ hybridization (FISH)***

FISH is a cytogenetic technique that was first developed and introduced in 1982 (Langer-Safer et al., 1982) that can be used to localize and visualize the presence of DNA sequences of interest along the chromosome, even during interphase or in an intact nucleus in living cells (Edelmann et al., 2001). By designing fluorescent probes hybridizing to the complementary DNA sequences of interest and using fluorescence microscopy, one can visualize where the probe is bound on the genomic DNA. FISH is often used for studying features in DNA in both medical diagnostics and basic research (Nath and Johnson, 2000).

By designing FISH probes for pairs of DNA sequences of interest, the presence or absence of a spatial interaction relationship between the two DNA regions can be directly visualized. However, this technique requires severe treatment that may affect the organization of the chromosomes (Dekker et al., 2002), can analyze only a limited number of DNA loci simultaneously (Williamson et al., 2014), and does not reveal if the DNA–DNA interaction is due to transcription factor binding or other mechanisms.

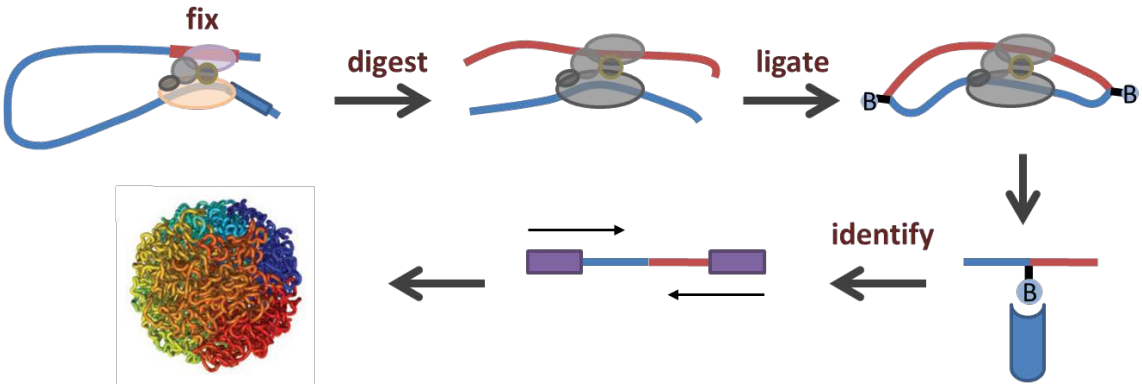



Figure 1-2. A schematic of the Hi-C protocol that can be used to detect the 3D structure of chromosome folding in eukaryotic nuclei.

Protein–DNA interactions are first fixed by formaldehyde treatment and then the genomic DNA is digested by a specific restriction enzyme (e.g. HindIII, NcoI, or MboI). The digested restriction sites are then filled in with biotinylated dNTPs and the interacting DNA fragments undergo proximity ligation (blunt-end ligation). The biotinylated DNA molecules are then pulled down using streptavidin beads and these are used for paired-end sequencing library construction for identifying DNA–DNA interactions globally. The chromatin architecture model figure was adapted from a figure by Lieberman-Aiden et al. (Lieberman-Aiden et al., 2009).

Table 1-1. Experimental protocols to capture chromatin conformation.

| | 3C | 4C | Capture-C | 5C | ChIA-PET | Hi-C |
|---|---|---|---|---|---|---|
| Scale | One-by-one | One-by-all | Many-by-all | Many-by-many | Many-by-many | All-by-all |
| Detection | PCR or sequencing | Inverse PCR sequencing | Sequencing | Multiplexed LMA sequencing | Sequencing | Sequencing |

### 1.3.2 3C variants – 4C, 5C, and ChIA-PET

There are several variations of 3C that extend the protocol from low-throughput to high-throughput detection of DNA–DNA interactions. While they all take advantage of the development of high-throughput sequencing techniques to identify DNA–DNA interactions, each one has a different focus (**Table 1-1**). One example, 4C is used for identifying all the interacting partner regions for one single locus at a time (Simonis et al., 2006; van de Werken et al., 2012; Zhao et al., 2006). Additionally, 5C detects all the interacting partners amongst a set of loci (Dostie et al., 2006), while ChIA-PET combines chromatin immunoprecipitation with a crosslinking and ligation procedure to capture the DNA–DNA interactions that are formed by a protein of interest (Fullwood et al., 2009). Capture-C is another variation that allows us to identify the interacting partner regions for a specific set of DNA regions of interest by designing probes for purifying those loci, which is followed by generation of sequencing libraries. These 3C variants allow us to connect chromatin structure to gene regulation (Dixon et al., 2012; Kalhor et al., 2012; Li et al., 2012; Rao et al., 2014; Shen et al., 2012; Zhang et al., 2013), to study DNA replication timing (Ay et al., 2014; Pope et al., 2014; Ryba et al., 2010), or to explore somatic copy number alterations (De and Michor, 2011; Fudenberg et al., 2011). Lastly, Hi-C captures and surveys DNA–DNA interactions genome-wide in a non-biased manner (**Figure 1-2**), that requires no prior knowledge for the interacting DNA regions. (Duan et al., 2010; Kalhor et al., 2012; Lieberman-Aiden et al., 2009). I discuss more details about Hi-C in the next section (**Section 1.1.3**).

### 1.3.3 Hi-C and long-range regulatory interactions

In 2009, Lieberman-Aiden and colleagues. developed a genome-wide version of chromosome conformation capture (Hi-C) (Lieberman-Aiden et al., 2009), which uses high-throughput sequencing technology that allows the global identification of DNA–DNA interactions within a genome of interest. The technique has been used to reveal conserved chromatin

organizing principles and chromatin folding as well as the three-dimensional (3D) architecture of the human genome (Dixon et al., 2012; Lieberman-Aiden et al., 2009; Nagano et al., 2013; Rao et al., 2014).

In Hi-C, the interacting DNA regions and their binding proteins are cross-linked by formaldehyde, restriction sites are digested, and proximity ligation is performed with biotinylated dNTPs filling in the restriction enzyme cut sites (**Figure 1-2**). The biotinylated dNTP allows for streptavidin bead-mediated pull down to capture pairs of ligated DNA fragments. After reversing of the crosslinks and size-selection of the ligated artificial DNA interacting molecules, sequencing libraries are constructed to interrogate the sequences and utilized to create an interaction map on a genome-wide scale.

Other than investigating the folding principle of chromatin, Hi-C data can also be applied to other applications such as genome assembly and haplotyping (Burton et al., 2013; Selvaraj et al., 2013) and locating centromeres and ribosomal DNA (Marie-Nelly et al., 2014; Varoquaux et al., 2015). Currently, it is recognized that the 3D architecture of chromatin can affect gene regulation and genome function. Unfortunately, the resolution of this original experimental iteration is limited by the distribution of cleavage sites for the specific restriction enzymes used within the genome. Ideally, this approach could be a powerful means for identifying the long-range interactions between enhancer elements and the promoters they regulate, but it was not used for this purpose initially because of this methodological shortcoming.

Other than numerous human and mouse cell lines, Hi-C-like data are also available for organisms such as yeast (Duan et al., 2010; Marie-Nelly et al., 2014; Mizuguchi et al., 2014; Tanizawa et al., 2010), bacteria (Le et al., 2013), *Drosophila melanogaster* (Hou et al., 2012; Li et al., 2015; Sexton et al., 2012), and *Arabidopsis thaliana* (Feng et al., 2014; Wang et al., 2015). Understanding how Hi-C analysis methods work has become important with the increasing number and variety of Hi-C datasets (Ay and Noble, 2015).

## 1.4 Computational methods and challenges for predicting enhancer–promoter pairs

### 1.4.1 Nearest gene

The most common approach for predicting enhancer–promoter interactions is to assign the nearest *cis*-promoter along the chromosome as the target of the enhancer element. An improvement to this method can be performed by adding insulator sites as a constraint on deciding which gene is regulated (Ernst et al., 2011; Heintzman et al., 2009). However, the example described in **Section 1.1.5** has shown that the gene (*Lmbr1*) which is harboring the enhancer is not its regulated target gene, but a linearly far away (around 1 Mb) gene (*Shh*) is. Additionally, only 27% of enhancers have an interaction with their nearest promoter based on recent ENCODE analysis (Sanyal et al., 2012). This suggests that the nearest gene along the chromosome is not always the only candidate for describing the function of a distal regulatory element.

### 1.4.2 Expression quantitative trait loci (eQTL)

eQTL studies can identify putative regulatory variants that influence gene expression. These studies determine if transcript levels of a set of protein-coding genes vary among a panel of individuals in correlation with their genotypes, which is usually at the risk alleles at GWAS loci (Cheung et al., 2003; Gilad et al., 2008; Stranger et al., 2007). Thus, eQTL analyses can be applied to detect candidate regulatory links between SNPs and their target genes.

Although eQTLs can suggest which genes are regulated by the intergenic alleles, the interpretations of the results can be challenging. First, the association could result from the confounding effects of the genetic background and thus the observation needs to be robust.

Second, the SNP associated with the target gene may not be the causal SNP. Any SNPs in the same haplotype or linkage disequilibrium (LD) block as the eQTL could be the actual causal SNP. Third, the association is correlative and may reflect indirect regulation between SNPs and genes (Fehrmann et al., 2011).

### 1.4.3 Correlation of DHS or histone modifications

The mapping of DNase I hypersensitive sites (DHS) has been utilized to identify regulatory regions including enhancers, promoters, insulators, and so on (Gross and Garrard, 1988; Wu et al., 1979). In combination with high-throughput sequencing technologies, the mapping of DHS was boosted to a genome-wide scale and called DNase-seq (Crawford et al., 2004; Sabo et al., 2004). Using DHS profiles examined across multiple cell types, one can observe whether pairs of enhancer and promoter DNA regions are simultaneously accessible or inaccessible to DNase (Thurman et al., 2012). Similarly, one can search for a correlation between the histone modification patterns at active enhancers and the transcript levels of their nearby coding-genes and thereby predict linkages between enhancer states and target genes (Ernst et al., 2011). These approaches demonstrate that epigenetic data can be used to assist the prediction of functional enhancer–promoter interactions. However, these correlations are limited as they do not prove direct regulation, and it would be computationally expensive to ascertain all possible pairs of DNA regions in the human genome.

### 1.4.4 Co-evolution between the elements

If a pair of DNA sequences (an enhancer and a promoter) function concordantly, it is expected that the sequence pair would be evolutionary conserved together among species, and the evolutionary constraints between them may result in higher levels of synteny among genomes

(Ahituv et al., 2005; Kikuta et al., 2007). One can utilize these concepts to predict enhancer–promoter interactions by identifying pairs of DNA sequences that are co-conserved and having a level of synteny across species (He et al., 2014). Nonetheless, this estimation based on evolutionary constraints lacks information on whether a pair of shared synteny DNA sequences is functioning directly as forming physical interactions between the enhancer and the promoter. In addition, since enhancer may change rapidly during evolution (Shibata et al., 2012), observing co-conserved DNA sequences may be insufficient for discovering all possible pairs of enhancer–promoter elements in the human genome.

## 1.5 Outline of dissertation

On the whole, the analyses of data derived from Hi-C and its related approaches have yielded important insights into the functional and regulatory significance of the three-dimensional structure of DNA. Moreover, these datasets can also be important resources for high-throughput identification of long-range regulatory elements and their interacting promoter partners. In this thesis, I focus on identifying enhancer–promoter interactions within Hi-C datasets to improve our understanding of the mechanisms by which enhancer elements regulate gene expression. To do this, I have developed a series of novel approaches toward identifying high-throughput DNA–DNA interactions using pre-existing Hi-C datasets. I also further characterized these important regulatory interactions, and discovered many of the protein complexes that bridge and mediate these events.

In **Chapter 2**, I describe our analytical approach for re-analyzing Hi-C data and calling interacting hotspots along the chromosome that are interacting with other hotspots. Using this approach, I specifically identify putative enhancer elements and their target gene promoters throughout the human genome.

In **Chapter 3**, I characterize the identified enhancer–target gene pairs by investigating the transcription factor binding sites contained within these enhancer elements and their target gene promoters. I also examine the conservation level of the identified enhancer elements and the extent to which their targeted genes were expressed in a cell-type-specific manner. Additionally, I discover that the relationship between enhancer elements and their target genes are many-to-many. I also discover transcription factor binding sites that are overrepresented in the enhancer elements when they are in contact with promoters.

In **Chapter 4**, I present an automated pipeline called HIPPIE that processes raw Hi-C data through the steps of mapping, normalization, integrating epigenetic marks, and reporting candidate enhancer elements and their target promoters. This pipeline is designed to run on high-performance computing clusters.

In **Chapter 5**, I switch my focus towards the task of understanding how enhancers select the correct promoters with which to interact genome-wide. To approach this, I describe a model I developed that explicitly bins the genome into meaningful physically-interacting DNA regions (PIRs). I then determine the transcription factor binding sites (TFBSs) that are overrepresented in enhancer PIRs when they are interacting with promoters, enhancers, exons, introns, or other intergenic PIRs. Furthermore, I reveal pairs of TFBSs that are overrepresented in enhancers and their interacting promoters, respectively, as well as identified candidate transcription factor complexes that can mediate enhancer–promoter interactions.

Finally, in **Chapter 6**, I summarize the computational methods I developed and the biological characteristics of the long-range regulatory pairs I have uncovered. I also highlight potential applications of our identified enhancer–promoter pairs with their characteristics, and possible uses of the transcription factor binding motifs we discovered to be overrepresented in driving the interactions between enhancers and their target promotors.

# Chapter 2 : A genome-wide approach for predicting enhancer and promoter interactions

This Chapter references work from:

Hwang, Y.-C., Zheng, Q., Gregory, B.D., and Wang, L.-S. (2013). High-throughput identification of long-range regulatory elements and their target promoters in the human genome. *Nucleic Acids Res*. 41, 4835–4846. doi:10.1093/nar/gkt188

## 2.1 Introduction

### 2.1.1 Traditional high-throughput methods on identifying enhancer elements — ChIP-seq and conservation

While the Human Genome Project was declared complete in 2003, many regulatory elements still remain undefined. Enhancers are one such class of elements because true definition of an enhancer requires identification of both the regulatory sequence as well as its interacting promoter region(s). Enhancer–target identification is further complicated by the fact that they interact in an orientation-independent manner, can be millions of base pairs away from each other, or even reside on different chromosomes (Banerji et al., 1981; Geyer et al., 1990; Lomvardas et al., 2006). Enhancer elements also have dynamic regulatory activities under various developmental and environmental conditions. For instance, they can activate gene expression in a tissue- and temporal-specific manner. Thus, they affect different sets of genes in different tissues (Visel et al., 2009), and/or play variable regulatory roles during animal development (Levine, 2010; Wilber et al., 2011). One well-studied example of this dynamic

property is the locus-control region (LCR) that regulates the cluster of five human β-type globin genes on 11p15.4 (Wilber et al., 2011). These globin genes are exclusively expressed in erythroid cells, and are expressed differentially in fetal and adult cells mediated by the LCR that is located about four kilobases (kb) upstream.

Recent studies reveal that in eukaryotes, histone modifications such as histone 3 lysine 27 acetylation (H3K27ac), histone 3 lysine 4 mono-methylation (H3K4me1), di-methylation (H3K4me2), and tri-methylation (H3K4me3) can play crucial roles in the activation of enhancer elements under different environmental conditions, cell lineages, tissue types, or developmental stages (Creyghton et al., 2010; Heintzman et al., 2007, 2009; Roh et al., 2007; Visel et al., 2009). These activating histone marks tend to be present in enhancer elements that are activated and absent when they are repressed. Additionally, activated regulatory elements are more likely to be located within the context of accessible (open) chromatin where they can be bound by transcription factors. The accessibility of specific DNA sequences can be determined by their sensitivity to digestion by DNase I, with open chromatin being highly digested and vice versa. Recently, large scale studies of activating (e.g. H3K27ac) histone marks and DNase I hypersensitive sites (DHSs) such as those from the Encyclopedia of DNA Elements (ENCODE) (Dunham et al., 2012; The ENCODE Project Consortium, 2004) have been used in various human cell-types to predict enhancer elements (Creyghton et al., 2010; Heintzman et al., 2009). Additionally, other high-throughput studies assaying E1A binding protein p300 and CREB binding protein (CBP) interaction sites have also been used to discover putative enhancers (Lee et al., 2011; Rödelsperger et al., 2011; Visel et al., 2009). Although these studies can predict enhancer elements on a large scale, they suffer from the inability to globally identify the target gene promoters of the identified enhancer elements.

### 2.1.2 From Hi-C sequencing to DNA–DNA interactions

Although the mechanism of enhancer–target promoter interaction formation is still not well understood, it is commonly accepted that enhancers and promoters interact with each other through a transcription factor protein complex (Schoenfelder et al., 2010). Based on this model, the chromosome conformation capture (3C) approach can be used to identify enhancer elements as well as their target genes simultaneously by detecting two linearly independent DNA segments that are bound to one another via a protein complex. One major drawback of the 3C approach is that it requires prior knowledge of the putative enhancer and promoter elements to allow design of specific PCR primers, which is often unknown. To address this limitation, a high-throughput version of 3C was developed (Hi-C) (Lieberman-Aiden et al., 2009) to detect genome-wide DNA–DNA interaction events. This approach avoids multiple PCR steps by ligating interacting DNA elements followed by high-throughput sequencing in order to provide unbiased identification of DNA–DNA interacting pairs. Several variants have been developed by other groups to identify the chromosome organization and regulatory sites of the human, yeast, and *Drosophila melanogaster* genomes (Duan et al., 2010; Kalhor et al., 2012; Sexton et al., 2012; Tanizawa et al., 2010). However, these original studies focused on determining large-scale, chromosomal organization, and did not demonstrate whether the high-throughput sequencing variant of 3C is sensitive or specific enough for prediction of enhancer–promoter interactions.

More recently, Chepelev et al. (Fullwood and Ruan, 2009) developed ChIA-PET, which is a strategy that combines 3C with ChIP-seq, for an enhancer associated histone modification (H3K4me2) to identify intra-chromosomal enhancer–promoter interactions (Chepelev et al., 2012). This led to the successful identification of only intra-chromosomal enhancer–promoter interactions that were associated with a specific histone modification (H3K4me2). Another recent study applied the variant 3C method (carbon-copy chromosome conformation capture (5C)) to identify ~one hundred enhancers and their specific target genes by designing ~6000 primers along the ENCODE pilot project regions (Sanyal et al., 2012). Although none of these previous

studies were at the genome-wide scale, they have demonstrated that datasets produced by the 3C method can be used for genome-wide identification of enhancer–target promoter interactions.

Here, we revisit the original Hi-C experimental data with the goal of identifying enhancer–target gene interactions on a genome-wide scale for humans. To do this, we developed a new analysis framework for Hi-C experiments that integrates multiple genome-wide, enhancer-defining datasets to identify enhancer–target gene pairs. Using this approach, we identified thousands of high confidence enhancer–target promoter interactions in two different human cell types. We validated these interaction pairs by demonstrating our putative enhancer elements are highly correlated with known p300 binding sites, and their target gene promoters are enriched in RNA Polymerase II (Pol II) binding. Furthermore, we found that the predicted enhancer elements are conserved in the mammalian lineage, and their target genes are expressed in a highly cell-type-specific manner. In total, our pipeline has allowed the first robust and genome-wide discovery of thousands of novel enhancer–promoter interactions in the human genome.

## 2.2 Identification of candidate enhancer elements and their target genes

### 2.2.1 Workflow for discovering enhancer elements and their target genes

We built an analysis workflow that extracts high-quality DNA interacting sites from Hi-C datasets. Figure **2**-**1** shows the overall workflow for identifying these DNA interacting hotspots, which we analyzed further in order to identify putative enhancer elements and their promoter partners. All three samples from the original Hi-C study (Lieberman-Aiden et al., 2009) were used in our analyses (cell line GM06990 with restriction enzymes HindIII and NcoI, as well as cell line K562 with HindIII (referred to as GM/HindIII, GM/NcoI, and K562/HindIII, respectively)). The

original Hi-C study used a 1 megabase (Mb) window size to uncover the three-dimensional organization of human nuclear chromosomes. However, this resolution is far too coarse for studying regulatory elements, which requires single nucleotide resolution. To improve resolution for our purposes of identifying DNA interacting hotspots, we applied our genomic distribution-based analysis for identification of these specific genomic regions (Zheng et al., 2010). Briefly, our algorithm first identifies clusters of Hi-C reads that are closer to each other than what the background geometric distribution dictates. We then labeled the resulting clusters as hotspots if their lengths on the chromosomes are longer than 99% of all clusters (for more discussion, please see **Section 2.2.2**). We found that a hotspot is on average ~1 kb in length, and between 107,059 and 185,042 total hotspots were identified in each of the three samples. The Hi-C method dictates that sequencing reads will start at or near the sites of the restriction enzyme used in the experiment rather than the actual DNA–DNA interaction site. Therefore, the resolution of this method is limited to the distance between the restriction sites of the particular restriction enzyme (RE) used for that study (**Figure 2-2**). To account for this shortcoming, we extended the length of the originally identified DNA interacting hotspots based on the estimated length between RE site positions on each human nuclear chromosome, while also allowing each nucleotide of an extended hotspot to represent the true interaction site. We found that on average an extended hotspot is 3–3.3 kb long (**Table 2-1**), indicating that our resolution has improved ~300-fold compared to the 1 Mb window size used in the original study.

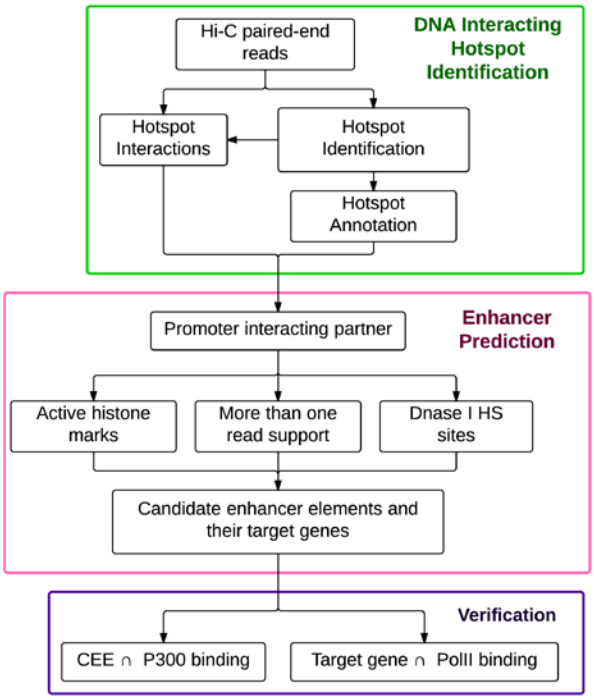



Figure 2-1. Genome-wide enhancer element identification workflow.

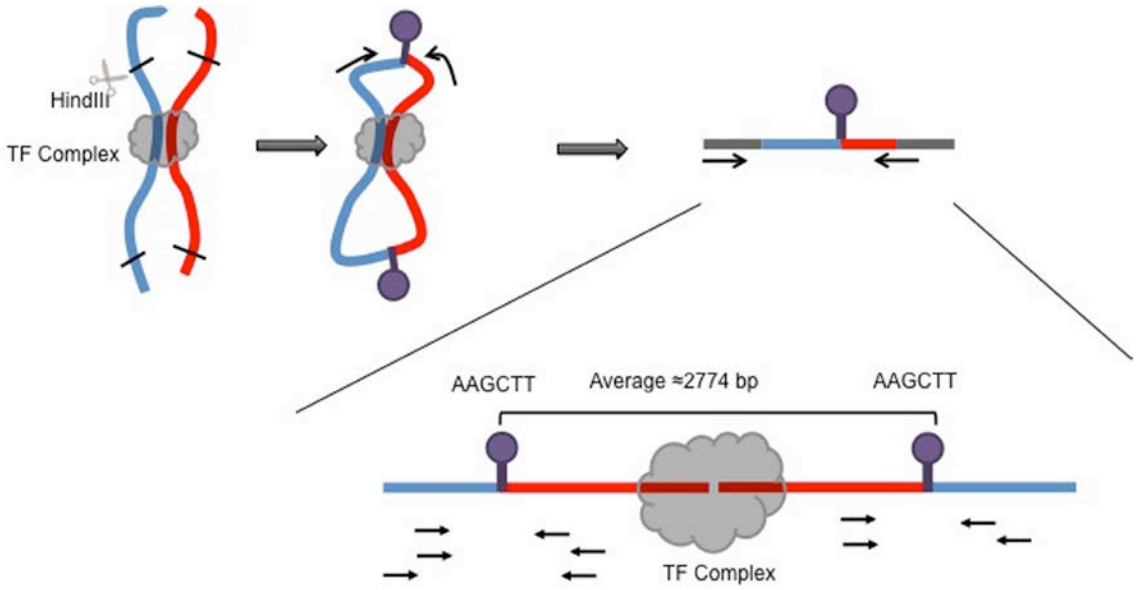



Figure 2-2. Identification of potential enhancer elements by our novel analysis pipeline requires an extension of regions that have been termed DNA interacting hotspots from the 3C data as depicted.

Table 2-1. Characterization of extended hotspots.

| Samples | GM/HindIII | GM/NcoI | K562/HindI |
|---|---|---|---|
| # Raw reads (paired-spots) | 30,009,111 | 28,659,279 | 36,823,509 |
| # Unique mapped pairs | 18,728,707 | 18,891,283 | 21,744,849 |
| Percentage of mapped to raw | 62% | 66% | 59% |
| # Unique mapped single-end | 37,457,414 | 37,782,566 | 43,489,698 |
| # Clusters (merged by gap length) | 4,973,281 | 5,076,539 | 6,247,694 |
| Average cluster length (bp) | 172.4 | 168.9 | 160.2 |
| # Hotspots | 107,059 | 166,990 | 185,042 |
| Average hotspot length (bp) | 1047.9 | 1007.8 | 964.1 |
| # Extended hotspots | 96,800 | 137,611 | 150,611 |
| Average extended hotspot length (bp) | 3065.8 | 3349.5 | 3282.1 |

GM = GM06990

### 2.2.2 Geometric-based model for identifying hotspot and extended hotspot

We first identified significant clusters in the Hi-C data using a geometric distribution-based model (Zheng et al., 2010). To do this, we first assembled all mapped reads for a given dataset (GM/HindIII, GM/NcoI, or K562/HindIII) into consecutive contigs (made up of overlapping reads) for each nuclear chromosome, without initially considering the read pairing information for these libraries. This approach allowed us to determine the gap regions between the identified contigs. These gap lengths should follow a geometric distribution:

$$P(X_i = k) = (1 - p_i)^{k-1} p_i$$

$$P(X_i \leq k) = 1 - (1 - p_i)^k$$

where $X_i$ and $p_i$ are the gap lengths and the probability of a position covered by any read on chromosome *i*, respectively. Accordingly, we fit the gap lengths to a geometric distribution for each chromosome (**Figure 2-3**) and estimated $p_i$ based on the mean gap length. We then grouped contigs into clusters by merging nearby contigs based on the gap distances between them. Specifically, contigs were merged into significant clusters if they are closer to each other than the 5% quantile according to the fitted geometric distribution.

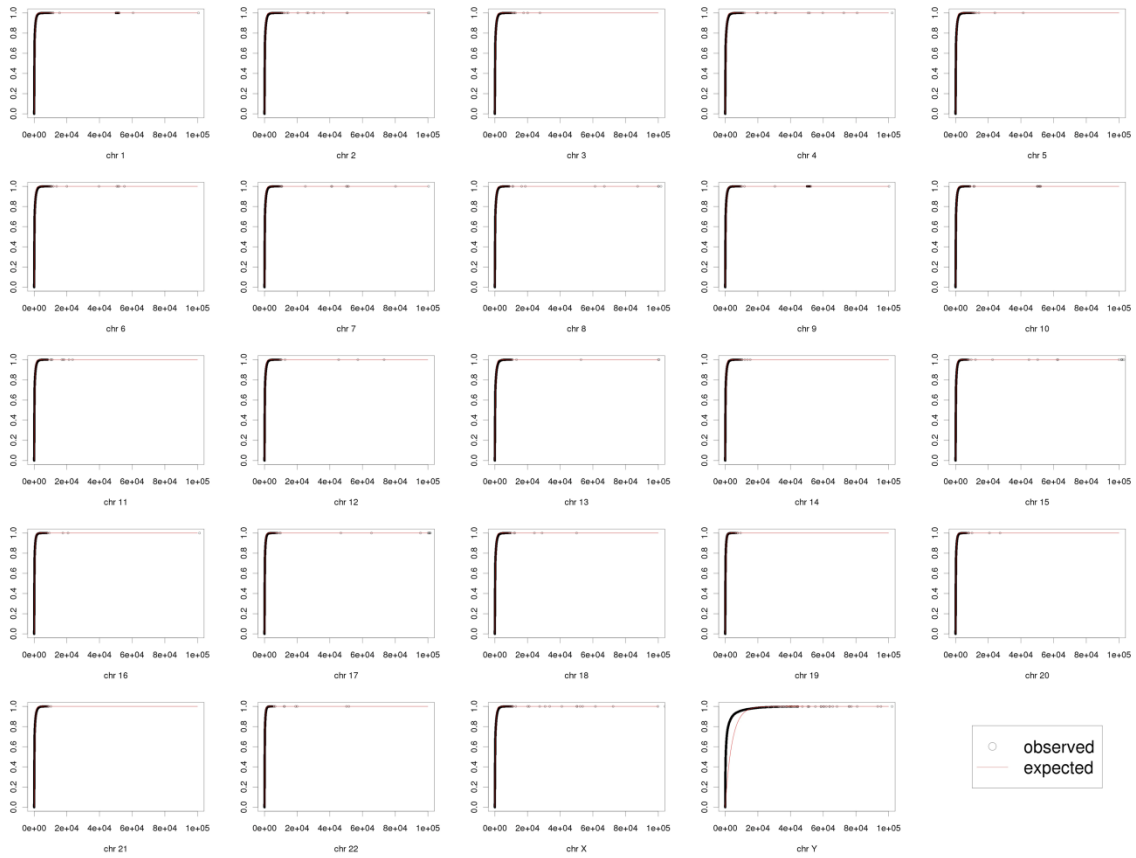



Figure 2-3. Gap lengths follow a geometric distribution along each chromosome in the Hi-C sample GM/NcoI.

The black circles represent the observed distance between each consecutive contig (gap lengths), and the red line represents the empirical cumulative geometric distribution estimated using the observed mean. The distributions for the samples GM/HindIII and K562/HindIII are similar (data not shown).

Next, we identified high confidence DNA interacting hotspots by fitting cluster lengths to an additional geometric distribution for each nuclear chromosome (**Figure 2-4**), where the $X_i$ value is based specifically on cluster length and the $p_i$ value is the emission probability based on the mean cluster length calculated for the Hi-C data for chromosome *i*. Only the significant clusters (length greater than or equal to the 99$^{th}$ percentile) identified with this second geometric distribution-based test were retained and defined as DNA interacting hotspots. It is worth noting that we did not take into account the Hi-C interaction data for these hotspots during this analysis step, but only looked for interacting partners during our analysis to identify those hotspots that are putative enhancer elements (see below). We also analyzed DNA interacting hotpots identified using the quantiles of 98% and 99.9%, and the results of these analyses are presented in **Figure 2-5** and **Figure 2-6**, respectively.

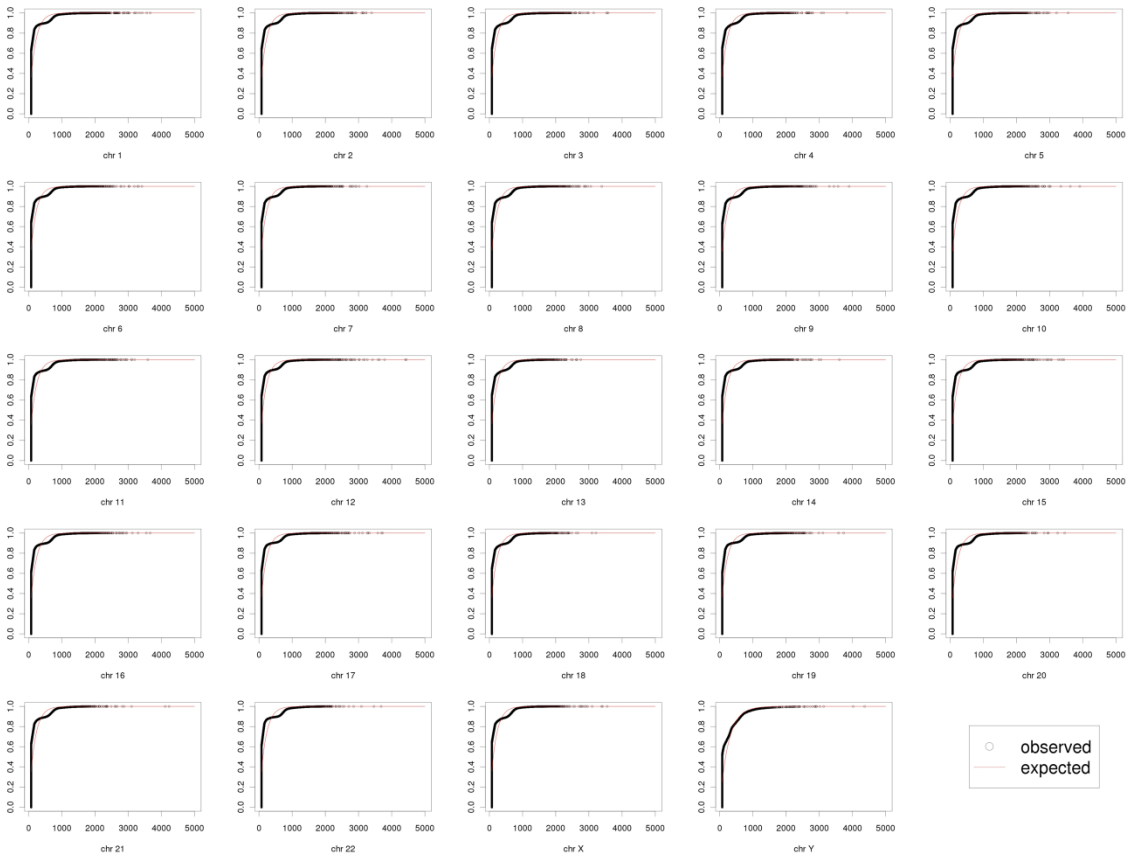



Figure 2-4. Cluster lengths follow a geometric distribution along each chromosome in the Hi-C sample GM/NcoI.

The black circles represent the observed cluster lengths, and the red line represents the empirical cumulative geometric distribution estimated using the observed mean. The distributions for the samples GM/HindIII and K562/HindIII are similar (data not shown).

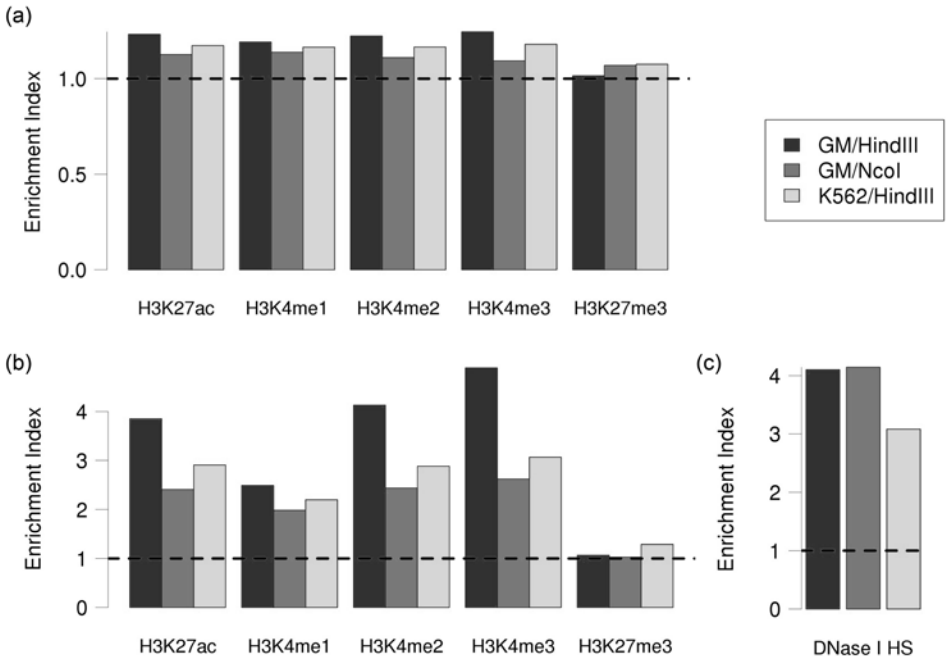



Figure 2-5. Additional analyses with a more relaxed cutoff (98% geometric distribution-based test) for identifying DNA–DNA interacting hotspots.

(a–b) Fold enrichment for activating (H3K27ac and H3K4me1 – 3) and repressive (H3K27me3) histone marks with (a) all CEEs that have a promoter interaction, and (b) CEEs whose promoter interaction is supported by > 1 read (*P* values < 0.001). (c) Fold enrichment of DNase I hypersensitive sites in CEEs with a promoter interaction supported by > 1 read and enriched in activating histone marks (*P* values < 0.001). The three samples are marked as follows; black bars: GM/HindIII; gray bars: GM/NcoI; and light gray bars: K562/HindIII. Dashed line is expected value based on genomic control.

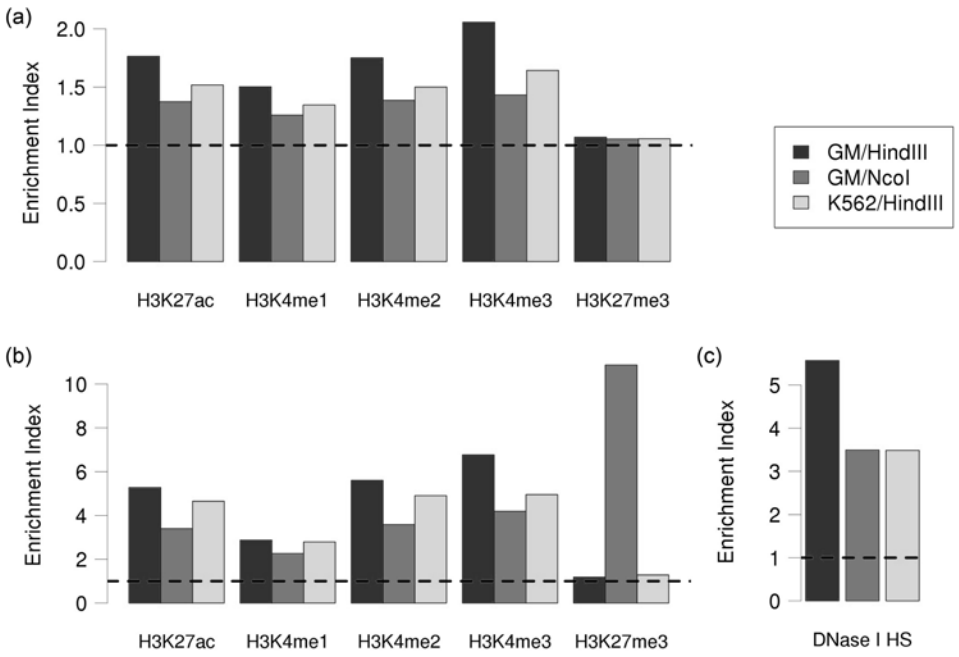



Figure 2-6. Additional analyses with a more relaxed cutoff (99.9% geometric distribution-based test) for identifying DNA–DNA interacting hotspots.

(a–b) Fold enrichment for activating (H3K27ac and H3K4me1 – 3) and repressive (H3K27me3) histone marks with (a) all CEEs that have a promoter interaction, and (b) CEEs whose promoter interaction is supported by > 1 read (*P* values < 0.001). (c) Fold enrichment of DNase I hypersensitive sites in CEEs with a promoter interaction supported by > 1 read and enriched in activating histone marks (*P* values < 0.001). The three samples are marked as follows; black bars: GM/HindIII; gray bars: GM/NcoI; and light gray bars: K562/HindIII. Dashed line is expected value based on genomic control.

### 2.2.3 Characterization of DNA interacting extended hotspots

We classified all extended hotspots based on human genome annotations and found that many of them are located within protein-coding genes, functional RNAs, and tandem repeats, etc., suggesting that some of the interaction hotspots may be involved in regulatory processes (**Figure 2-7**). Interestingly, we observed that extended hotspots were located within 5% – 20% of total promoter regions (defined as the 500 base pairs upstream of protein-coding gene transcription start sites) of the human genome. This led us to speculate that some of the extended hotspots from our reanalysis of Hi-C data may actually reflect target promoters that are interacting with enhancer elements in the human genome.

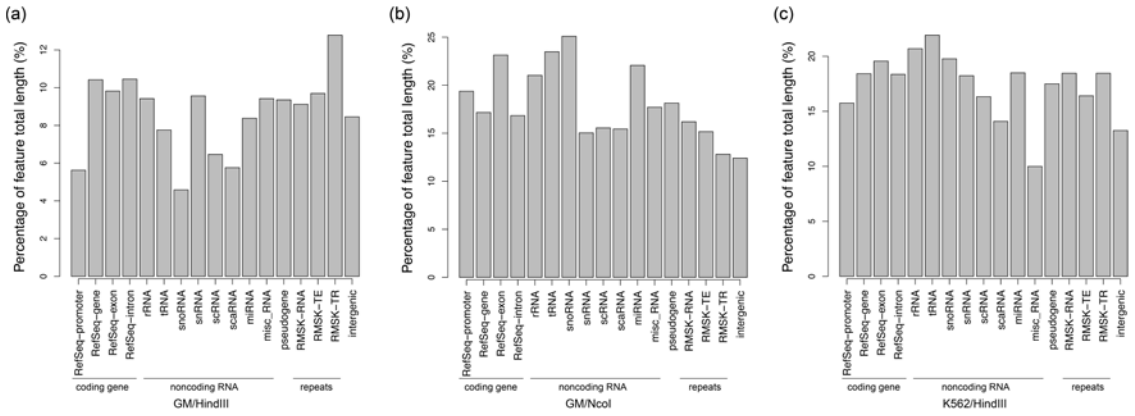



Figure 2-7. Functional annotation of extended hotspots for sample. (a) GN/HindIII, (b) GM/NcoI, and (c) K562/HindIII.

Each bar (as labeled) represents the percent of total length of each genomic feature that overlaps with extended hotspots.

### 2.2.4 Candidate enhancer elements are enriched in activating histone marks

***Prediction of candidate enhancer elements (CEEs)***

To identify candidate enhancer elements (CEEs), we first considered extended hotspots that interact with a protein-coding gene promoter region(s) (defined as the 500 base pairs (bp) upstream of the annotated transcription start site). As shown in **Table 2-2**, 22%–62% of the extended hotspots interact with a protein-coding gene promoter. The variation in promoter interactions is likely a consequence of the number of promoters that are covered by extended hotspots, which is influenced by both the total sequencing depth in a particular sequencing library and the restriction enzymes and cell types used in the Hi-C experiments. We next examined the enrichment of promoter-interacting extended hotspots in four activating histone modifications known to be associated with enhancer elements (H3K27ac, H3K4me1, H3K4me2, and H3K4me3), and a heterochromatic histone modification (H3K27me3) as a negative control (Raney et al., 2011; Rosenbloom et al., 2011). As expected, we found that promoter-interacting extended hotspots are enriched (permutation test, *P* values < 0.001) in all four activating histone modifications but not with H3K27me3 (**Figure 2-9**) when compared to the random background control. These results suggest that many of the promoter-interacting extended hotspots are human enhancer elements.

Table 2-2. Number of CEEs present after each filtering step

| Filtering step | GM/HindIII | GM/NcoI | K562/HindIII |
|---|---|---|---|
| Promoter partners | 22,818 | 90,200 | 93,109 |
| Strong interactions (> 1 read) | 1,757 | 11,001 | 9,955 |
| Activating histone mark enrichment | 928 | 5,617 | 5,814 |
| DNase I HS sites | 823 | 4,809 | 5,033 |

To further improve our confidence that we are detecting bona fide enhancer–target gene promoter interactions, we added an additional quality control step where we only retain promoter-interacting extended hotspots if their promoter interaction is supported by more than one read (n > 1) in the sequencing results (**Figure 2-8**). This filtering step dramatically reduced the number of potential enhancer elements in all three samples. In fact, only 7.7% - 12.2% of the promoter-interacting extended hotspots were retained as potential enhancer elements (**Table 2-2**). This step likely reduced the number of false positives in our dataset, since we found it substantially increased the enrichment in the four enhancer-associated activating histone modifications (H3K27ac, H3K4me1, H3K4me2, and H3K4me3) in the remaining promoter-interacting extended hotspots (**Figure 2-9b**). Taken together, these results indicate that increased read support for the promoter-extended hotspot interactions is necessary for high confidence identification putative enhancer elements and their targets from Hi-C experimental data.

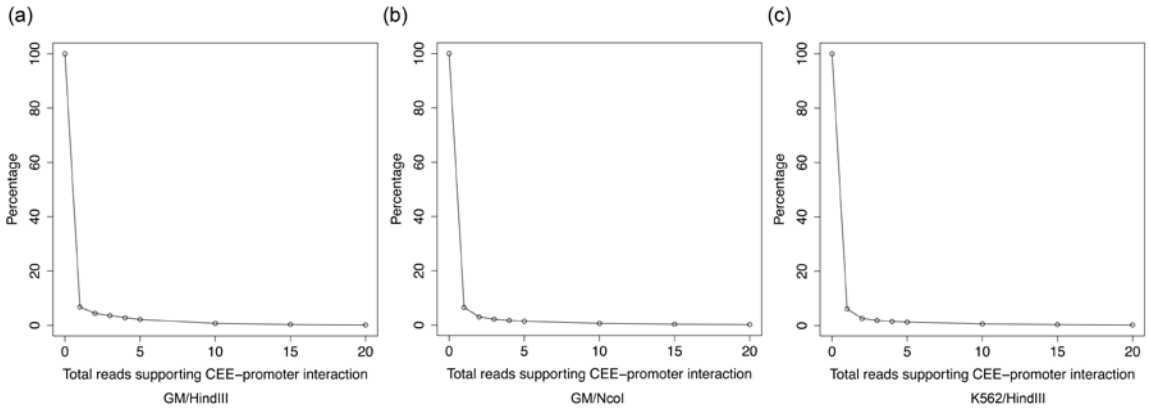



Figure 2-8. Read support for CEE–promoter interactions in the three Hi-C samples. (a) GM/HindIII, (b) GM/NcoI, and (c) K562/HindIII.

The final filtering step in our pipeline to identify candidate enhancer elements (CEEs) was to determine the enrichment of DNase I hypersensitive sites (DHSs) within the subset of high confidence promoter-interacting extended hotspots (supported by > 1 sequencing read) using previously published datasets (Gross and Garrard, 1988; Wu et al., 1979). From this analysis, we found that the set of high confidence promoter-interacting extended hotspots from all three original Hi-C experiments were enriched (*P* values < 0.001) in DHSs (**Figure 2-9c**). The tendency of high confidence promoter-interacting extended hotspots to co-localize with DHSs provides further evidence of the reliability of our analysis strategy to identify bona fide enhancer element–target promoter pairs in the human genome. In summary, the combination of these results has led us to incorporate all three of these analysis steps in our pipeline for genome-wide prediction of candidate enhancer elements (CEEs) and their interacting target promoters in the human genome.

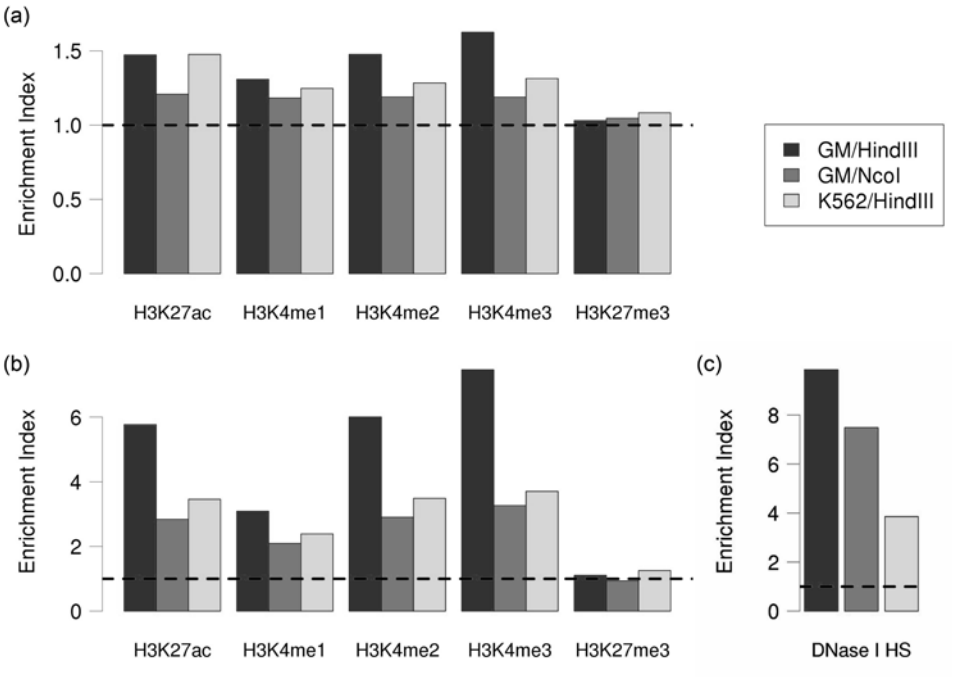



Figure 2-9. Potential enhancer elements are enriched for activating histone marks and DNase I hypersensitive sites.

(a – b) Fold enrichment for activating (H3K27ac and H3K4me1 – 3) and repressive (H3K27me3) histone marks with (a) all CEEs that have a promoter partner, and (b) CEEs whose promoter partner is supported by > 1 read. (c) Fold enrichment of DNase I hypersensitive sites in CEEs with a promoter interaction supported by > 1 read and enriched in activating histone marks. The three samples are marked as follows; black bars: GM/HindIII; gray bars: GM/NcoI; and light gray bars: K562/HindIII. Dashed line is expected value based on genomic control.

## 2.3 Discussion

The original Hi-C article suggested that this experimental approach could identify regulatory elements with better sequencing throughput, although this analysis was never performed. In this study, we show that with careful analyses and comprehensive integration of publicly available functional genomic datasets, Hi-C data can be used to comprehensively identify enhancer–target gene interactions genome-wide. We first employed a geometric distribution model to identify DNA interacting hotspots instead of using a sliding window to probe 1 Mb segments of the human genome. This change in analysis methods significantly improved our genomic resolution (~3.3 kb or 300-fold improvement). This increase in resolution is necessary for identifying the actual sequences of regulatory elements in the human genome.

From this initial list of DNA interacting hotspots, we focused on intergenic sequences that interact with protein-coding gene promoters, and found these elements overlap significantly with enhancer-associated chromatin marks such as H3K27ac and H3K4me1 that have been previously used to identify enhancer elements (Creyghton et al., 2010; Heintzman et al., 2009). Interestingly, a recent study by Chepelev et al. utilized this property by combining Hi-C with H3K4me2 immunoprecipitation to identify enhancer–promoter interactions (Chepelev et al., 2012). However, this study focused solely on *cis*- interactions and did not examine other enhancer-associated epigenetic marks. Here, we used multiple chromatin marks as well as DHS datasets to identify thousands of candidate enhancer elements in two human cell types with high confidence. We also uncovered that not all epigenetic marks are equal for these purposes. Specifically, we found that all four activating histone marks are enriched on the putative enhancers, but they demonstrate distinct levels of enrichment (**Figure 2-9**). In total, our analysis pipeline incorporates data for multiple histone modifications and DHSs, which increases confidence that bona fide enhancer elements are truly being identified.

In addition, our analysis is unique when compared to three other studies that were recently published. Specifically, Lan *et al.* integrated histone modification data that overlapped sites enriched with reads from Hi-C experiments for the K562 cell line, and found 12 clusters of Hi-C sites with different combinations of histone modifications (Lan et al., 2012). However, their analyses were limited only to these overlapping regions and did not also interrogate all of the other relevant datasets as we have done here. Furthermore, their study only examined enhancer–promoter interactions on a specific subset of the human genome (GATA1/GATA2 target genes). In another recent study, 5C experiments were performed to study enhancer–promoter interactions. However, they focused entirely on the 44 ENCODE pilot genomic regions instead of performing a genome-wide analysis (Sanyal et al., 2012). This is because a global study of enhancer–promoter interactions is not feasible with the 5C protocol, since this approach requires the design of specific primers for a select group of targeted regions. ChIA-PET, another recently developed protocol that detects chromosomal interactions using high-throughput sequencing (Chepelev et al., 2012) was also used to study enhancer–promoter pairs. However, as pointed out by the developers of this method, their approach is different from unbiased approaches like Hi-C because it requires an antibody to a specific histone modification, protein, etc. Thus, this method will not detect any enhancers that are not in close proximity to the histone modification, protein, etc. being immunoprecipitated.

The Hi-C protocol has an inherent limitation for enhancer discovery as we have described (**Figure 2-2**). Specifically, we have revealed that the data from the Hi-C protocol actually detects restriction enzyme sites around the bona fide DNA-DNA interaction regions. While increasing the read coverage is still essential for obtaining high-quality enhancer–promoter interaction data, it does not solve this particular limitation. To accommodate this lapse in resolution, we had to extend the identified DNA interacting hotspots in both directions, since we could not predict which direction to extend based on the Hi-C sequencing data alone. Thus, our analysis workflow is the

first to allow confident prediction of enhancer–target promoter interactions from Hi-C data, and provides the framework for future studies that will use this approach for these same purposes.

## 2.4 Materials and methods

### 2.4.1 Comparisons between replicates

To comprehensively identify putative enhancer–target promoter interactions in the human genome, we first downloaded the original genomic alignments for the paired-end Hi-C sequencing data from two different human cell lines, a lymphoblastoid (GM06990) and a chronic myelogenous leukemia (K562) cell line (GEO accession number GSE18199). After an initial analysis, we found that the overlap of extended hotspots (defined below) between biological replicates of the GM/HindIII sample is 50.6% (*P* value < 2.2e-16, chi-square test). Furthermore, the sequencing reads from these same samples were also combined in the original Hi-C study. Therefore, to more comprehensively identify DNA-DNA interacting pairs, mapped reads from biological replicates of the GM/HindIII sample were merged to single datasets. In total, we looked at the interacting patterns for three different sample sets from the original Hi-C study: GM06990 cells with HindIII (GM/HindIII), GM06990 with NcoI (GM/NcoI), and K562 with HindIII (K562/HindIII) digestion.

### 2.4.2 Identification of CEEs enriched in activating histone modifications

We began selecting for candidate enhancer elements (CEEs) by focusing on the extended hotspots that: 1) had at least one of its interacting regions overlapping a protein-coding gene promoter and 2) the particular CEE–promoter interaction was supported by > 1 paired-end sequencing read in the corresponding Hi-C dataset. We then determined the overlap (see below

for description of enrichment analyses) between these promoter-interacting extended hotspots and the four activating histone marks (H3K4me1, H3K4me2, H3K4me3, and H3K27ac) that are known to be associated with enhancer elements in the human genome (Creyghton et al., 2010; Heintzman et al., 2009; Roh et al., 2007; Visel et al., 2009). We also examined the overlap of these promoter-interacting extended hotspots with H3K27me3, a heterochromatic histone modification (Young et al., 2011) that is not enriched at enhancer elements. To further select for CEEs that are likely bona fide enhancer elements, we only maintained promoter-interacting extended hotspots containing known enhancer-related histone modifications that are also enriched in DNase I hypersensitive sites (DHSs). Thus, CEEs are defined as highly confident promoter-interacting extended hotspots enriched in known enhancer-related histone modifications and DHSs. For these enrichment analyses, the histone modification and DHS data was downloaded from the UCSC ENCODE production phase (hg18 assembly) (Raney et al., 2011; Rosenbloom et al., 2011). It is of note that we used the lymphoblastoid cell line (GM12878) and chronic myelogenous leukemia (K562) from the ENCODE project in our study, as they are the most closely related cell lines to those used in the original Hi-C study (GM06990 and K562). As a control, we generated 1000 sets of the same number of extended hotspots randomly selected from the human genome (random extended hotspots), and used them as a background to evaluate the significance of enrichment for all subsequent analyses.

# Chapter 3 : Global characterization of long-range regulatory elements and their target genes

To understand characteristics of how long-range regulatory elements function and access our prediction of the enhancer–target gene interaction, we identified candidate enhancer element and their target gene in using the Hi-C data in **Chapter 2**. In this chapter, we then focus on understanding the characteristics of enhancer element and their target genes.

This Chapter references work from:

Hwang, Y.-C., Zheng, Q., Gregory, B.D., and Wang, L.-S. (2013). High-throughput identification of long-range regulatory elements and their target promoters in the human genome. *Nucleic Acids Res*. 41, 4835–4846. doi:10.1093/nar/gkt188

## 3.1 Enhancers and their target genes are enriched in binding activities associated with gene expression

To provide further evidence that our CEEs are bona fide enhancer elements, we examined the enrichment of p300 binding within these regions. We focused on p300 because it is a known enhancer-associated co-activator that mediates the regulation of target gene expression (Eckner et al., 1994; Maston et al., 2006). We found that the CEEs from all three Hi-C experiments were enriched (*P* values < 0.001) in p300 binding compared to a background control of all extended hotspots (**Figure 3-1**). This enrichment in p300 binding within CEEs strongly suggests we have identified bona fide enhancers, and by using the Hi-C data in this analysis we also identify the gene promoter(s) that each element can target.

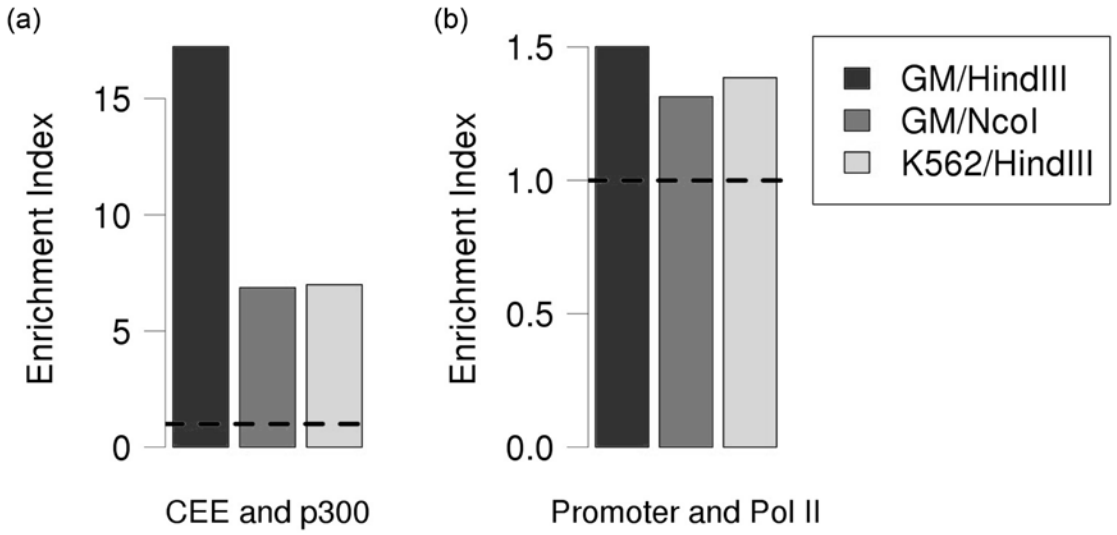



Figure 3-1. Potential enhancer elements are enriched for p300 binding, and their target genes are highly bound by Pol II.

(a) p300 binding site enrichment in candidate enhancer elements (CEEs). (b) Pol II enrichment observed for the genes targeted by CEEs.

Enhancer elements generally activate gene expression through direct interaction with target promoters (Maston et al., 2006; McKnight and Kingsbury, 1982; Nolis et al., 2009) that results in increased RNA Pol II association. Therefore, we tested whether the CEE target gene promoters were enriched for Pol II binding. From this analysis, we found that CEE target promoter regions are more than 20% enriched (*P* value < 0.001) for Pol II binding when compared to the promoters of all other protein-coding genes. These results suggest that transcription initiation is increased at promoters that are in contact with the CEEs compared to all other gene promoters in the human genome.

To further test if transcription is generally higher from target genes of the CEEs, we also investigated the enrichment of Pol II Ser2 phosphorylation, which marks elongating Pol II, within these loci for the K562 dataset using previously published data (GSM935547). Interestingly, we observed a 1.47-fold enrichment (*P* value < 0.001) in Pol II Ser2 phosphorylation within target genes of our CEEs compared to all other protein-coding genes. These results indicate that the CEE–target gene interaction not only increases Pol II promoter binding, but also effects transcription elongation.

In summary, the consistent enrichment of the CEEs in p300 binding as well as their target genes with initiating and elongating Pol II strongly suggests that we have identified bona fide enhancer–target gene pairs by reanalyzing the previous Hi-C results. We also compared our CEEs to the enhancers predicted in the recently published study using 5C with primers designed to the ENCODE pilot project regions covering only ~1% of the human genome (Sanyal et al., 2012). We found that our CEEs overlap significantly with these enhancer elements compared to all extended hotspots lying within the ENCODE pilot region as a background control (**Table 3-1**). Thus, our method provides an important improvement over previous approaches for identifying human enhancer elements because we not only identify enhancers, but we also uncover their specific regulatory targets on a genome-wide scale.

Table 3-1. Comparison of the CEEs predicted using Hi-C and enhancer predictions in 5C (Sanyal et al., 2012)

| Samples | Hi-C #CEEs | 5C # Enhancers | # Intersects | *P* value of intersects |
|---|---|---|---|---|
| GM*/HindIII | 19 | 87 | 1 | 0.1316 |
| GM*/NcoI | 37 | 87** | 5 | 0.0001 |
| K562/HindIII | 137 | 119 | 9 | < 0.0001 |

* 5C study uses GM12878; Hi-C study uses GM06990.

** 5C study uses only HindIII as the RE, here we are comparing using the GM/HindIII data set.

# CEEs shows the number of CEEs that is overlapped with the 5C primers along the ENCODE pilot regions. *P* value is calculated by permutation tests using the extended hotspots that overlap the ENCODE primer sets as background.

We found that CEEs interact with 1.17–1.62 target gene promoters on average (**Table 3-2)**, which is consistent with recent results (Sanyal et al., 2012) and suggests that human enhancer elements can interact pleiotropically. Additionally, we found that most target gene promoters interacted with multiple (1.17–2.36) CEEs (**Table 3-2**), suggesting the existence of enhancer redundancy in the human genome.

Table 3-2. Characteristics of enhancer–target interactions

| Samples | # CEEs | Average CEE interactions | # of target promoters | Average target promoter interactions | # of enhancer–promoter |
|---|---|---|---|---|---|
| GM/HindIII | 823 | 1.17 | 820 | 1.17 | 953 |
| GM/NcoI | 4,809 | 1.62 | 3,444 | 2.27 | 7,757 |
| K562/HindIII | 5,033 | 1.42 | 3,032 | 2.36 | 7,104 |

We also determined the interaction characteristics of CEEs and their target genes. We found that the vast majority of these interactions are intra-chromosomal (on the same chromosome), while fewer than 13% are inter-chromosomal (**Figure 3-3a**) with very little read support for these latter associations (**Table 3-3**). Interestingly, we found that > 95% of the intra-chromosomal interactions occurs within a range of 1 Mb (**Figure 3-3**b). In total, these results indicate that the majority of the CEEs that we have identified from the Hi-C data are in relatively close proximity to their target promoters.

Table 3-3. Average number of reads support for intra- and inter-chromosomal interactions of CEEs and their target promoters.

| Sample | Intra-chromosomal | Inter-chromosomal |
|---|---|---|
| GM/HindIII | 6.56 | 2.20 |
| GM/NcoI | 6.09 | 2.29 |
| K562/HindIII | 6.78 | 2.39 |

## 3.2 Enhancers and target promoters are enriched in enhancer-associated motifs

To identify specific sequence motifs in the CEEs and their target promoters, we further searched for overrepresented sequences using HOMER (Heinz et al., 2010). Not surprisingly, a quick search using a random genomic background yielded the recognition site of HindIII (AAGCTT) as top motif in the CEEs identified in by the original GM/HindIII Hi-C experiment (**Table 3-4a**). These results suggest that as expected the Hi-C experimental approach identifies DNA interaction sites that are localized near restriction sites in the human genome (**Figure 2-2**). To minimize this bias for RE sites, we performed the motif searches with a background of all extended hotspots. As a result, we identified 38, 54, and 39 motifs from each experiment, including the binding motifs of known enhancer-associated transcription factor families such as Sp1, NRF1, E2F, GATA, and ETS (**Table 3-4b**). Remarkably, we found that in all three CEE datasets there was significant enrichment for the binding sequence of the E26 transformation-specific (ETS) family binding domain-containing proteins. These proteins act as transcription factors that bind to specific enhancers and promoters, and facilitate the assembly of transcription machinery to initiate gene expression (Gutierrez-Hartmann et al., 2007; Hollenhorst et al., 2011). Thus, our CEEs are enriched in sequences known to bind enhancer specific proteins.

Table 3-4. (a) Top 3 most enriched motifs for all CEEs using the whole-genome as the background sequence in the K562/HindIII library. (b) Top 10 most enriched motifs in CEEs from the GM/NcoI library using extended hotspots as the background.

(a)

| Motif | *P* value | % of Targets | % of Background |
|---|---|---|---|
|  | <1e-300 | 87.9 | 55.2 |
|  | <1e-258 | 84.5 | 80.7 |
|  | <1e-228 | 60.2 | 55.6 |

(b)

| Transcription Factor (DNA binding domain) | Motif | *P* value | % of Targets | % of Background |
|---|---|---|---|---|
| Sp1(Zf) |  | 1e-134 | 38.5 | 23.6 |
| NRF1(NRF) |  | 1e-111 | 15.0 | 6.4 |
| ETS(ETS) |  | 1e-69 | 41.6 | 30.3 |
| ELF1(ETS) |  | 1e-64 | 58.2 | 46.8 |
| GFY-Staf |  | 1e-54 | 7.5 | 3.1 |
| NRF1 |  | 1e-51 | 19.2 | 12.0 |
| YY1(Zf) |  | 1e-45 | 9.1 | 4.61 |
| E2F |  | 1e-45 | 30.2 | 22.0 |
| GFY |  | 1e-38 | 10.0 | 5.6 |
| GABPA(ETS) |  | 1e-32 | 77.3 | 70.1 |

## 3.3 Enhancers are conserved within vertebrates

Functional elements are often under evolutionary selection because of their cellular function(s) (Blanchette and Tompa, 2002). To study if the CEEs are under evolutionary selection, we investigated the conservation score in these elements across the mammalian clade (cons44way conservation score (Pollard et al., 2010)) compared to their upstream and downstream flanking sequences. We found that the CEEs tend to be more conserved than their flanking regions (*P* value < 0.05 for all datasets) (**Figure 3-2a**). In total, these results revealed that the CEEs that we have identified are under purifying selection in the human genome, suggesting that they are functional enhancer elements.

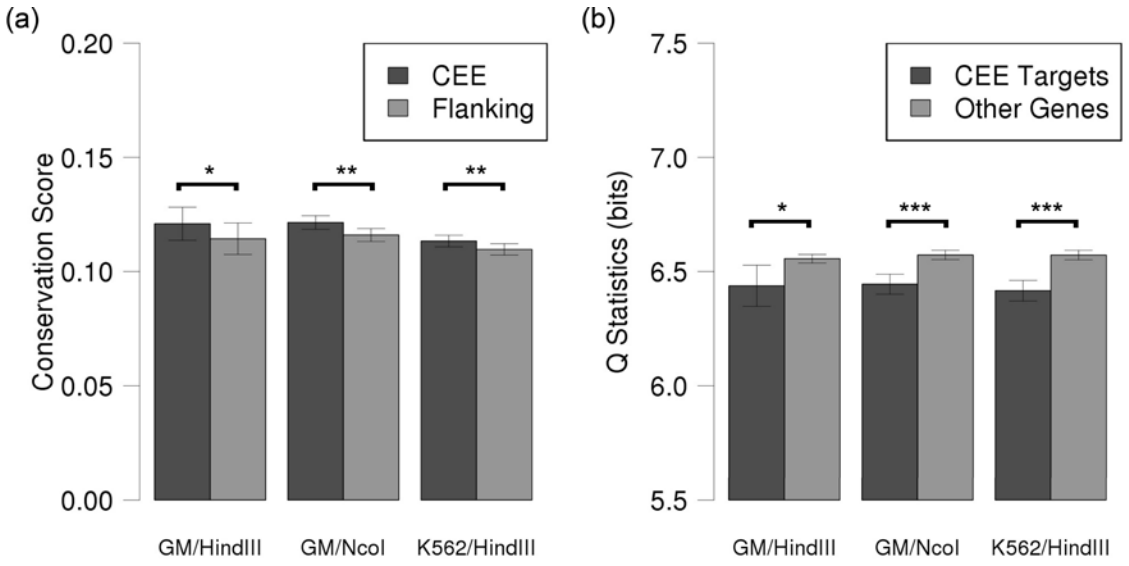



Figure 3-2. Potential enhancer elements are evolutionarily conserved, and their target genes are expressed in a cell-type-specific manner.

(a) The conservation score of CEEs (black bars) compared to similarly-sized flanking regions (gray bars) from the three different Hi-C experiments (as specified). (b) The Q statistic values for CEE target (black bars) compared to non-target (gray bars) genes from the three different Hi-C experiments (as specified). Error bars indicate s.e.m. Differences are statistically significant (* denotes *P* value < 0.05, ** denotes *P* value < 0.01, and *** denotes *P* value < 0.001, Wilcoxon rank-sum test)

## 3.4 Tissue-specific expression of the target genes

Enhancer elements are known to function in a cell-type-specific manner (Heintzman et al., 2009), so the expression profiles of their target genes are likely to display a similar pattern. To determine whether genes targeted by CEEs exhibit this cell-type-specific tendency, we computed the *Q* statistic (Schug et al., 2005) for every human gene expressed in nine ENCODE cell types (see Methods for descriptions), and then compared CEE target genes to all other loci in the cell types (GM12878 or K562) most closely corresponding to those used in the original Hi-C experiment (GM06990 or K562). We found that genes targeted by the CEEs have significantly lower *Q* values, indicating that these loci are expressed in a cell-type-specific manner. This is true for CEEs identified using all three Hi-C experiments (*P* value = 0.01, 2.11e-05, 4.24e-16 for GM/HindIII, GM/NcoI, and K562/HindIII, respectively). In total, all of our results suggest that we have identified thousands of bona fide enhancer–target gene interactions. A significant amount of future attention can now be focused on determining the biological functions and significance of these newly identified interactions in human cells.

## 3.5 Discussion

Our analyses revealed that unannotated long-range and inter-chromosomal enhancer–target gene interactions can be detected using Hi-C data. This is in strong contrast to previous studies of short-range enhancer–target gene interactions, namely predicting *cis*-targeted genes within a small fixed window (Rödelsperger et al., 2011) or by defining a variable but local transcriptional domain (Dixon et al., 2012) around the identified enhancer elements. We found inter-chromosomal interaction to be much less frequent than both *cis* and *trans* intra-chromosomal interactions (**Figure 3-3a** and **b**). This may be because the inter-chromosomal and long-range interactions are underestimated due to the very limited sequencing depth of the initial Hi-C experiments, or to these being less stable and/or transient interactions. Thus, we may

identify more of these interactions with future Hi-C experiments with much greater sequencing depth.

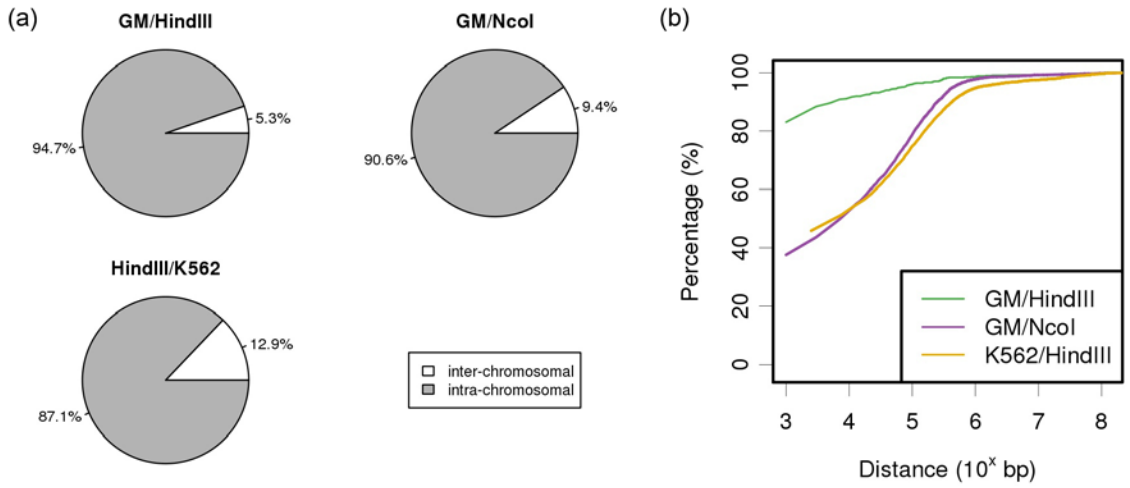



Figure 3-3. Characterization of CEE–target gene interaction distance.

(a) The portion of inter- and intra-chromosomal CEE–target interactions for the three different Hi-C samples as denoted. (b) The distance distribution of intra-chromosomal CEE–target interactions.

We have also uncovered both one-to-one and multiple-to-multiple CEE–target interactions **(Table 3-2**). These results reveal the extreme complexity of enhancer–target promoter relationships in the human genome. Interestingly, genes targeted by the same enhancer element could be co-regulated, competing, or activated in different developmental stages or tissue types. Similarly, enhancer elements that target the same gene could also be cooperative or competing in maintaining gene expression homeostasis or for altering expression activities of the target gene. Comprehensive time-course studies with high read coverage (for better sensitivity) will be necessary to further elucidate the regulatory mechanisms behind each enhancer–target promoter interaction.

Applying our analysis workflow to identify DNA interaction information from Hi-C, allows identification of candidate enhancers and their associated target genes. In the future, comparing datasets similar to the ones provided here with findings from GWAS and eQTL studies is likely to provide mechanistic insights into how many intergenic SNPs can be associated with a certain disease. In total, a comprehensive list of enhancer–promoter interactions is likely to significantly improve the resources available to future genetic studies focused on human disease.

## 3.6 Materials and methods

***Enrichment analyses***

All enrichment analyses for CEEs and their target promoters (e.g. p300 binding) were performed by computing the enrichment index (*ERI*) as a ratio of the two proportions:

$$ERI(A) = C(A)/P(A)$$

where *A* is the set of intervals for a particular histone modification or other genomic feature (e.g. DHS, p300 binding, or Pol II binding) determined using ENCODE ChIP-seq or DNase-seq

experiments (Encode and Consortium, 2011). *C(A)* is the total length in base pairs of CEEs (or interacting promoter hotspots if we are examining target promoter characteristics) that overlap with *A*, *and P(A)* is the mean of total lengths that overlap with *A* from 1000 random control sets (see **Section 2.4.2**). It is worth noting that in enrichment analyses for CEEs, each permuted set is selected randomly from the collection of all extended hotspots with the additional constraints that they must have similar chromosomal and length distributions as the set of CEEs being analyzed (Quinlan and Hall, 2010). For enrichment analyses of CEE target promoters, each control set is selected randomly from the promoter regions of the 21,522 non-redundant protein-coding genes in the hg18 assembly. Thus, a high *ERI* for a set of CEEs or their target promoters indicates that they tend to be overlapping with a particular histone modification or binding feature when compared to all extended hotspots or non-redundant protein-coding gene promoters, respectively.

***Characterizing p300 binding to CEEs and RNA Polymerase II binding to their target promoters***

We downloaded the previously identified p300 and RNA Pol II binding sites from the UCSC ENCODE database for GM12878 and K562 cell lines (hg18 assembly) (Encode and Consortium, 2011; Raney et al., 2011; Rosenbloom et al., 2011). We then calculated the *ERI* for p300 binding within CEEs as well as RNA Pol II binding to CEE interacting promoters as described above.

***Determining cell-type-specific expression of CEE target genes***

We downloaded the previously published gene expression profiles for the nine ENCODE human cell lines (GSE26312) (Ernst et al., 2011). Data were normalized by RMA (Bolstad et al.,

2003; Irizarry et al., 2003a, 2003b) and $\log_2$-transformed. We aggregated probeset-level to gene-level expression values for each cell line as follows. For each probeset, we computed the average expression level across replicates. For each gene, we then computed the average expression across multiple probesets (if applicable). The gene expression profiles of the nine ENCODE cell lines were combined into a common gene set (13,436 genes), and between sample expression values were normalized again to eliminate any array-specific bias using quantile normalization (*normalize.quantiles* function in R/affy package) (Bolstad et al., 2003). To determine if a gene has strong tissue-specific expression in either GM12878 or K562 cells compared with the other seven cell types, we used an entropy-based metric (Schug et al., 2005) as follows. For each gene *g* we computed $p_{c|g}$ as the expression level in cell type *c* divided by the sum of expression levels across all nine cell lines. The entropy (Shannon, 1948) for *g* is defined as $H_g = - \sum_{1 \le c \le N} p_{c|g} \log_2(p_{c|g})$, where $N = 9$ is the total number of cell types in this study. $H_g$ ranges between 0 (gene *g* is expressed in only one cell type) and $\log_2(N)$ (gene *g* is expressed uniformly in all cell types). To measure the specificity for a particular cell type *c*, we computed $Q_{g|c} = H_g - \log_2(p_{c|g})$. The quantity $-\log_2(p_{c|g})$ has a range between 0 (when gene *g* is only expressed in cell type *c*) and infinity (when gene *g* is not expressed in cell type *c*).

***Sequence motifs in CEEs***

We examined the sequence motifs of the CEEs using the HOMER software package (Heinz et al., 2010), and only considered 8, 10, and 12 bps for the motif length in each sample. We used all extended hotspots as the background when searching for overrepresented motifs *(-bg* parameter in HOMER) in an effort to reduce potential biases introduced towards restriction sites due to the original Hi-C protocol**.** Significance levels were set as *P* value < 0.05.

# Chapter 4 : A high-throughput identification pipeline for promoter interacting enhancer elements (HIPPIE)

This Chapter references work from:

Hwang, Y.-C., Lin C.-F., Valladares O., Malamon J., Kuksa P. P., Zheng Q., Gregory, B.D., and Wang, L.-S. (2015). HIPPIE: a high-throughput identification pipeline for promoter interacting enhancer elements. *Bioinformatics.* 31, 1290–1292. doi:10.1093/bioinformatics/btu801

## 4.1 Introduction

Genome-wide chromosome conformation capture (Hi-C) has been utilized to reveal three-dimensional connectivity of chromatin regions in eukaryotic nuclei (Lieberman-Aiden et al., 2009). Due to its capability to capture all possible chromatin interactions in a genome, it has been recently employed to observe long-range regulatory elements with their geographically proximal target gene promoters (Hwang et al., 2013). Although there have been workflows successfully expediting the analysis of one-dimensional high-throughput sequencing results such as whole-exome sequencing, ChIP-seq, DNase-seq, and RNA-seq; there are limited tools to un-tangle two-dimensional DNA–DNA physical interactions using Hi-C datasets. In an effort to reduce the obstacle of processing these large-scale datasets, and to establish an analysis protocol to detect candidate long-range regulatory elements, we implemented an automated workflow that processes Hi-C results starting from read mapping with quality controls, and corrects for biases in interactions based on the linear distance, mappability, GC content, and fragment lengths of each pair of Hi-C reads. This pipeline identifies candidate promoter-interacting enhancer elements by

integrating Hi-C results with epigenomics data such as histone modifications and DNase hypersensitivity sites.

## 4.2 Integration of multiple genomic datasets in ENCODE

HIPPIE takes Hi-C raw reads as the input and generates a list of enhancers with their interacting target gene(s) as the output. We built HIPPIE with five step-wise phases (Figure **4**-**1**): (I) read mapping, (II) quality control, (III) identification of significant DNA–DNA interacting regions, (IV) enhancer–target gene predictions, and (V) characterization of these long-range interactions. Although HIPPIE is streamlined and automated, each phase of HIPPIE can be independently called with commonly used file formats generated by different platforms and programs, such as FASTQ, SAM, BAM, or BED. Thus, it can readily be combined with other upstream processing and/or downstream analyses. The implementations of each phase are described below.

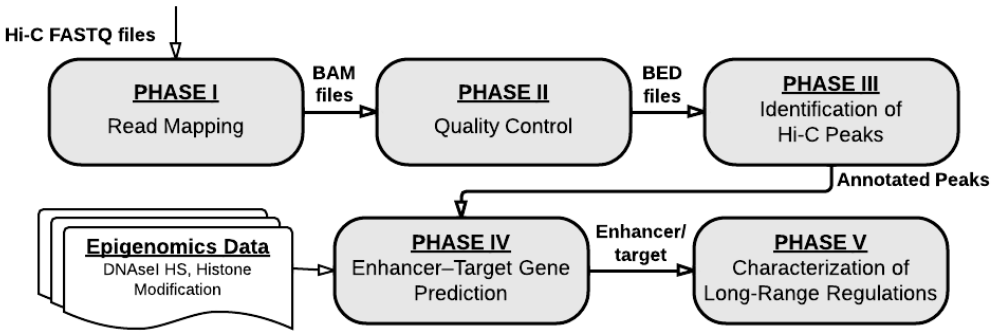



Figure 4-1. An overview of HIPPIE

Read mapping in HIPPIE uses the sequence alignment package BWA (Li and Durbin, 2009). It takes raw Hi-C paired-end sequencing reads in FASTQ format as input, and applies SAMtools (Li et al., 2009) to compress the read alignment SAM files to BAM files and produces mapping quality metrics. The quality control steps discard reads not passing a user-defined mapping quality criterion (default minimum quality score = 30), remove potential PCR duplicates, ignore mitochondrial sequences, and exclude random contigs.

Identification of interacting DNA fragments consists of calling significant Hi-C peaks and annotating their genomic features. Because the resolution of Hi-C is constrained by the length distribution of the fragments produced by the chosen restriction enzyme (the sequence between two consecutive restriction sites along the genomic DNA), we retained the restriction fragments that harbor significantly higher specific than nonspecific read coverage (**Section 4.5**) as "Hi-C peaks". Next, we applied BEDtools (Quinlan and Hall, 2010) to annotate these peaks with genetic features downloaded from the UCSC Genome Browser (Karolchik et al., 2014), including annotations for promoters, exons, introns, other functional RNAs, etc.

Enhancer–target gene prediction reveals the interactions of the annotated peaks, and produces a list of candidate enhancer elements (CEEs) and the gene(s) with which they interact as supported by Hi-C reads. To correct for Hi-C experimental biases in their linear distance between restriction fragments, GC content, mappability and length reported in (Yaffe and Tanay, 2011), we implemented the algorithm introduced by (Jin et al., 2013) and extracted statistically significant DNA–DNA interactions ($P$ value ≤ 0.1, negative binomial distribution test). For enhancer prediction, our pipeline selects Hi-C peaks that interact with a promoter, reside in a DNase hypersensitive region, as well as harbor high levels of enhancer-associated histone modifications (H3K27ac or H3K4me1) but not promoter-associated marks or repressive marks (H3K4me3 and H3K27me3). An option of using ENCODE genome segmentations (Hoffman et al., 2013) for candidate enhancers is also provided. This step is followed by characterization of enhancer–

promoter interactions, which summarizes the overall properties of the interactions such as their linear distance distribution, as well as reports the enrichment of specific histone modifications and GWAS single nucleotide polymorphisms (SNPs) within the CEEs. Note the phases are not only streamlined with error control, but also modularized for individual calls. For instance, users can map their Hi-C reads with other algorithms, and call peaks with HIPPIE starting at phase III; or one can directly import the interaction regions and utilize HIPPIE for enhancer–target gene identifications (phase IV).

## 4.3 Using HIPPIE

HIPPIE was built specifically for long-range enhancer–gene pair interaction detection upon the architecture of our previous DNA sequencing workflow (Lin et al., 2013). For instance, we implemented job dependencies and error checking to automate the entire process. To run HIPPIE, users first prepare a configuration file describing the software and data paths, as well as their Hi-C library information. For each library, HIPPIE generates a corresponding bash script for Open Grid Scheduler job submission commands that can be invoked at the command line. When errors occur, all following jobs will be held for users to troubleshoot and re-execute the stalled phase or step. This modular architecture reduces the potential for unnecessary, repeated jobs. A complete run of HIPPIE produces candidate enhancer elements (CEEs) in BED format that are annotated with their target gene symbol(s), together with Hi-C read count supporting the interaction and interaction *P* values.

To run HIPPIE, users can either install the package on their own cluster system, or simply access a pre-created Amazon Machine Image (AMI) from Amazon Web Services (AWS) on an Elastic Compute Cloud (EC2) instance (AMI ID: ami-3b0fb252).

We evaluated HIPPIE on our cluster using publicly available Hi-C datasets (Dixon et al., 2012). These datasets are 36 and 100 base pair (bp) paired-end sequencing with a total of 59.4 giga bases (1.35 billion single reads) from the Illumina GA II platform (GEO accessions GSM862723 and GSM892306). The total CPU time required for HIPPIE to process these datasets is 437.26 core-hours. The breakdown of CPU time for each phase is as follows: read mapping: 64.4%, quality control: 5.8%, identification of peaks: 26.8%, enhancer–target gene interaction prediction: 2.8%, and characterization: 0.1%. The maximum memory usage is 4.77G for read mapping. We identified 3,707 candidate enhancer elements with 3,190 targeted RefSeq genes.

## 4.4 Comparison with other tools

While there are publicly available pipelines for processing Hi-C reads, there are no open-source software packages that take raw reads as input and ultimately identify enhancer–target gene pairs along with their interaction characteristics (**Table 4-1**). Among them, Hicpipe takes mapped reads and corrects the contact maps based on possible experimental biases (Yaffe and Tanay, 2011). HiC-inspector aligns reads and generates a contact matrix with user-defined read densities but does not have statistical filtering steps for the identified fragments (https://github .com/HiC-inspector). HiCUP maps reads with filtering out artifacts and self-interacting reads without any statistical model (http://www.bioinformatics.babraham.ac.uk/projects/hicup/). None of those identify long-range regulatory elements; nor provide error checking.

Table 4-1. Comparison among Hi-C processing pipelines

| | HIPPIE | HiCUP | HiC-inspector | hicpipe |
|---|---|---|---|---|
| **DNA – DNA Interactions** | | | | |
| Mapping algorithm | BWA | Bowtie | Bowtie | - |
| PCR artifacts filtering | ✓ | ✓ | - | - |
| Restriction Fragment size | Exact size | - | User-defined max. size | Bias correction |
| User-defined threshold for peak calling | ✓ | - | - | ✓ |
| GC-content normalization | ✓ | - | - | ✓ |
| **Enhancer–target gene prediction** | | | | |
| Epigenomics Annotation | ✓ | - | - | - |
| Enhancer–target distance | ✓ | - | - | - |
| Enhancer GWAS enrichment | ✓ | - | - | - |
| Enhancer histone modification enrichment | ✓ | - | - | - |

## 4.5 Materials and methods

### 4.5.1 Coverage threshold for restriction fragments (Hi-C peak identification)

For quality control, as shown in **Figure 4-2**, we first identified specific and non-specific read pairs by the distances of each mapped read in a pair from the closest restriction site ($d_1$ and $d_2$). When $d_1 + d_2 \leq 500$ nt, both reads are considered specific reads, while $d_1 + d_2 > 500$ nt, both reads are considered non-specific reads as previously described (Yaffe and Tanay, 2011). We then calculated the specific and non-specific read coverage of each restriction fragment. A Hi-C peak is called if the specific read coverage for a restriction fragment is higher than the 95$^{th}$ percentile of the non-specific read coverage distribution. Non-specific reads are subsequently discarded and only Hi-C peaks and the specific reads denoting their interaction partners are used for further analysis.

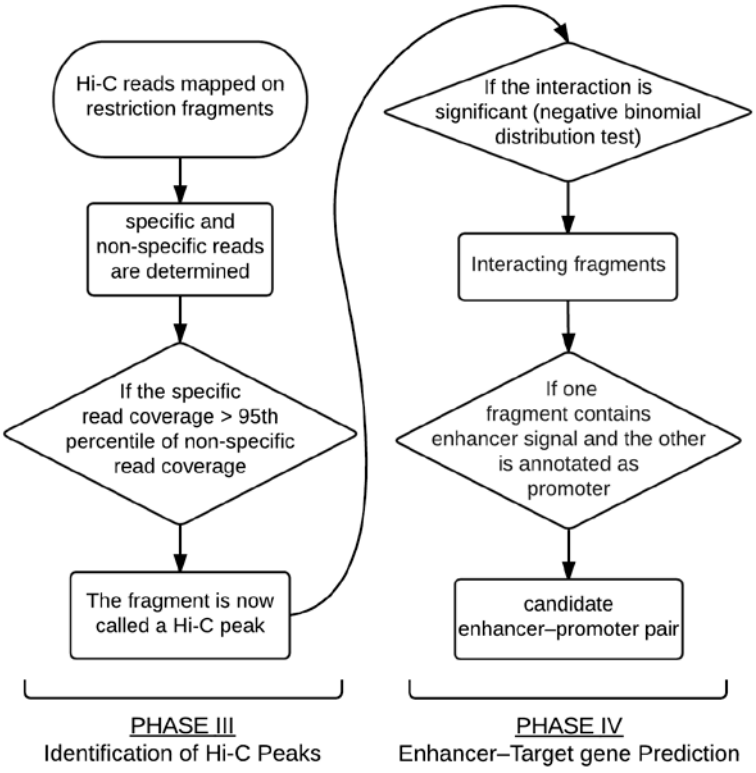



Figure 4-2. The quality control flow for HIPPIE phase III and phase IV.

### *Calculating mappability and GC content*

We generated a set of 36 base pair (bp) pseudo-reads using a one nucleotide sliding window to extract reads from both ends (500 nt in length) of each restriction fragment (i.e. the region between two consecutive restriction enzyme cut sites along the human genome). The pseudo-reads are then mapped back to the genome using BWA and the mappability of each restriction fragment is determined by the fraction of uniquely mapped pseudo-reads (MAPQ ≥ 30 and with SAM tag of "XT:A:U" in BWA) for both of its ends. The GC content of each fragment is calculated as the percentage of total bases that are either guanine or cytosine within the 500 nts that make up their ends (Jin et al., 2013).

### *Estimating random contact frequencies and significance calculation*

We implemented the calculation of random contact frequencies described previously (Jin et al., 2013) into our HIPPIE analysis pipeline. Specifically, the expected number of read pairs ($\mu_{i,j}$) for each interaction (*i, j*) on the same chromosome is:

$$\mu_{i,j} = m_i \times m_j \times F_{i,j}^{gc} \times L_{i,j}$$

We set $x_{i,j}$ as the observed (actual) read pair supporting the interaction of restriction fragments *i* and *j*. In order to account for the inherent biases of the Hi-C methodology, we first binned the restriction fragments by the GC content of their ends into 20 bins (with break points 0, 0.05, 0.01… 0.90, 0.95, and 1). For the length of the restriction fragments, we took the log of this value and binned by 2 orders of magnitude. If two restriction fragments are on the same chromosome, we binned the distance of each by a 5000 bp window size. We then let $B_i^{len}$ be the bin assignment of restriction fragment *i* by its length, $B_i^{gc}$ to allow the bin assignment of restriction fragment *i* by its GC content, and $B_{i,j}^{gc}$ be the bin assignment for the fragment interaction of *i* and *j* based on their linear distance.

Then $L_{i,j}$ is the expected frequency of contacts between fragment *i* and *j*, using a correction factor for restriction fragment length bias:

$$L_{i,j} = \frac{\sum_{k,l} \frac{x_{k,l}}{m_k \times m_l}}{\sum_{k,l} 1}$$

Where

$\forall\{k,l\}$ satisfy: $B_k^{len} = B_i^{len}, B_l^{len} = B_j^{len}, B_{k,l}^{dist} = B_{i,j}^{dist}, chr(k) = chr(l), m_k > 0.2, \text{and } m_l > 0.2$

Where $F_{i,j}^{gc}$ is a correction factor for GC content bias (contact fraction for the corresponding GC bin among all possible GC bins) between fragment *i* and *j* :

$$F_{i,j}^{gc} = \frac{\sum_{k,l} \frac{x_{k,l}}{m_k \times m_l} / \sum_{k,l} 1}{\sum_{u,v} \frac{x_{u,v}}{m_u \times m_v} / \sum_{u,v} 1}$$

Where $\forall\{k,l\}$ satisfy: $B_k^{gc} = B_i^{gc}, B_l^{gc} = B_j^{gc}, B_{k,l}^{dist} = B_{i,j}^{dist}, chr(k) = chr(l), m_k > 0.2, \text{and } m_l > 0.2$ , and $\forall\{u,v\}$ satisfy: $chr(u) = chr(v), m_u > 0.2, \text{and } m_v > 0.2$.

Note for inter-chromosomal interactions, the same estimation equations are used as above, except the requirements of $chr(k) = chr(l)$, $chr(u) = chr(v), \text{and } B_{k,l}^{dist} = B_{i,j}^{dist}$ .

With the estimation of $\mu_{i,j}$ of each restriction fragment pair (*i, j*), we then fit all $X_{i,j}$ to a negative binomial distribution to estimate the statistical significance of the interaction between each pair:

$$X_{i,j} \sim NB(u_{i,j}, p)$$

Where *p* is the fixed value ($\beta$-1/$\beta$), where $\beta$=2.057 as derived by (Jin et al., 2013).

# Chapter 5 : Identifying the transcription factors mediating enhancer–target gene regulation in the human genome

## 5.1 Abstract

The majority of reported genetic variations associated with diseases or traits reside in non-coding genomic regions. One class of these non-coding elements is enhancers, which regulate gene expression by recruiting transcription complexes and forming long-range interactions with the promoters of protein-coding genes. However, the mechanisms underlying these long-range regulatory interactions as well as the protein complexes involved in the formation of such interactions in the three-dimensional space of the nucleus remain poorly understood.

To screen for all possible enhancers and the genes they regulate genome-wide, we analyzed the latest Hi-C sequencing datasets and identified long-range regulatory (e.g. enhancer–promoter) interactions and protein factors mediating these interactions. While previous work aimed at understanding genome organization and identifying chromatin interactions at different scales (1MB compartment), we developed a novel methodology aimed at accurately identifying and delineating interacting DNA regions with physical interaction mediated by binding of transcription or other protein factors. We identified 1,194,010 physically-interacting DNA regions (PIRs). These regions have been identified using multiple sources of information ranging from Hi-C read-out, including genomic mapping positions, distances of mapped reads to their flanking restriction sites, and strand orientations of the mapped read-pairs. We found 1,193,987 significant intra-chromosomal PIR interactions genome-wide involving 602,671 PIRs. We then identified significantly enriched protein-binding sites in these PIRs and discovered 30 DNA-binding proteins involved in the formation of long-range regulations. Our novel analyses identified 338,791/1,193,987 (25%) DNA–DNA interactions with evidence of protein binding among the

observed chromatin interactions. With these protein-binding regions, we discovered motifs recruiting transcription factors that participate and facilitate in 2,466 enhancer–target gene interactions. These significantly over-represented interactions between these protein factors recapitulated over 86% of known protein–protein interactions.

## 5.2 Introduction

Enhancers are non-coding DNA elements that regulate gene expression and affect phenotype by recruiting transcription factors that directly interact with promoter elements in the DNA. The genome-wide relationship between enhancers and their target genes remains obscure because the three-dimensional DNA looping associated with enhancer–promoter interactions is challenging to detect. In 2009, Lieberman-Aiden et al. developed Hi-C, a non-biased "all-to-all" protocol utilizing high-throughput sequencing to capture chromosome conformations genome-wide that resolves the chromosome architecture with 1 Mb resolution (Lieberman-Aiden et al., 2009). In Hi-C, the physically-interacting DNA regions and their binding proteins are cross-linked, followed by restriction enzyme cleavage and proximity ligation of the interacting DNA fragments to localize and capture pairs of interacting DNA fragments. The sequencing library of these ligated DNA fragments are then built and utilized to identify the chromatin interaction map genome-wide.

Recently, Rao et al. modified the Hi-C protocol by applying a more frequent restriction enzyme (4-cutter, e.g. MboI) instead of a 6-cutter (e.g. HindIII or NcoI) to achieve higher resolution in localizing interacting DNA fragments, and by performing the DNA–DNA proximity ligation in intact nuclei to generate denser Hi-C contact matrix (Rao et al., 2014). With their new protocol, and so far the highest sequencing read depth (~3 billion reads/sample), they successfully resolved the DNA–DNA interaction map with resolution of only a few kilobases (1 kb,

5 kb, or 10 kb). With kilo-base resolution of the DNA–DNA interaction contact, it becomes possible to delineate long-range regulatory interactions more accurately compared to traditional Hi-C experiment. At 10 kb-resolution, Rao et al discovered chromatin loops are enriched for CTCF binding motifs in a convergent orientation, as well as for other transcription factor ChIP-seq peaks. They also resolved sub-compartments of the chromatin contact domains in 25 kb-resolution maps. However, the regulatory interactions and the DNA-binding proteins mediating these interactions were not explored as even higher (<5 kb) resolution is required to delineate these interactions.

Identification of the chromatin interactions from Hi-C data is complicated by many possible systematic biases including GC content of the interacting DNA fragments, fragment length after cleavage, as well as read mappability (Imakaev et al., 2012; Rao et al., 2014; Yaffe and Tanay, 2011). To acquire a corrected DNA–DNA contact matrix from Hi-C raw read counts, two major approaches have been proposed to account for these biases. One approach is to compute correction factors for each locus based on average read frequencies for equally-sized bins of genome-wide GC content, fragment length and mappability (Jin et al., 2013; Yaffe and Tanay, 2011). The other approach is to learn bias vector from balancing the raw contact matrix (Imakaev et al., 2012; Rao et al., 2014). It has been shown that the two approaches for bias corrections give comparable corrected results (Imakaev et al., 2012; Jin et al., 2013; Yaffe and Tanay, 2011). On the other hand, to account for the typically observed decrease in the read count with the increase in the linear genomic distance between interacting DNA regions (Imakaev et al., 2012; Kaplan and Dekker, 2013), Ay et al. introduced Fit-Hi-C as spline-based fitting method calling significant DNA–DNA interactions by the linear genomic distance between two DNA sites and the normalized read count (Ay et al., 2014).

Despite advances in analyzing Hi-C data, there are still limitations of the binning system for constructing the chromatin interaction map. Jin et al. has shown it is possible to study

interactions with the restriction fragment as a unit (i.e. DNA region between two consecutive restriction sites) with a 6-cutter restriction enzyme (Jin et al., 2013). Restriction fragment-based binning might be problematic for more frequent cutters such as 4-cutter (e.g. MboI), since each restriction fragment length is much smaller on average compared to a 6-cutter and interacting DNA regions are more likely to span more than one restriction fragment. The read distribution for each restriction fragment could become too sparse to deal with. Other groups have been working with a uniform binning scheme, in which the linear genome is partitioned into fixed-width bins (Lieberman-Aiden et al., 2009; Rao et al., 2014). However, this uniform binning scheme with fixed-width partition may not be capturing the actual physically-interacting DNA regions. As a result, this could complicate the precise identification of interacting regulatory elements (promoters, enhancers, etc.), the analysis of long-range regulatory interactions, as well as impede the discovery of protein-binding motifs and protein factors mediating these interactions.

It has been shown that architectural proteins, such as CTCF and cohesin, contribute to both global chromosome architecture and regulatory interactions (Gibcus and Dekker, 2013). Additionally, Rao et al. has identified binding sites of transcription factors in specific DNA loci that participate in DNA looping interactions at a genomic region resolution of 5 kb and 10 kb (Rao et al., 2014). However, the mechanisms by which transcription factor complexes facilitate and mediate long-range enhancer–promoter interactions, including their formation, regulation and maintenance genome-wide are not yet characterized (Maksimenko and Georgiev, 2014).

In this work, we developed a methodology to identify DNA physically-interacting regions (PIRs) in the genome from Hi-C sequencing read-out with evidence of protein mediated physical interaction. Unlike the typically used fixed-width uniform binning, the proposed dynamic binning model allows us to more accurately describe the interacting DNA elements (e.g., promoters, enhancers, etc.). We also discovered the sequences that are overrepresented in these regions

and identified the protein factor binding sites which can suggest the underlying mechanisms of formation, regulation and maintenance of these long-range interactions.

## 5.3 Results

### 5.3.1 Hi-C processing pipeline for identifying physically-interacting regions

We analyzed high read depth Hi-C sequencing datasets (Rao et al., 2014) for a lymphoblastoid cell line (GM12878), determined DNA physically-interacting regions (PIRs) with evidence of protein-mediated physical interaction, and identified significant DNA–DNA interactions between PIRs as shown in **Figure 5-1**. We first aligned the ~3 billion Hi-C paired-end reads to the human genome (hg19 assembly) using STAR aligner (Dobin et al., 2013). Each of the single-end read from a read pair was first mapped separately. To improve mapping, both contiguously mapped and chimeric reads have been identified and paired (see **Section 5.5**, **Table 5-1**). In total, we found that 77% of the paired-end reads were uniquely mapped. To remove potential random ligation events, including un-cut, self-ligated, or re-ligated read-pairs, we filtered out the read-pairs that are less than 5,000 bps apart from each other (see **Section 5.5**) as suggested in previous research (Jin et al., 2013; Lajoie et al., 2015). In addition, to correct for all possible Hi-C experimental biases including length of the crosslinked DNA fragments, restriction site accessibility, or ligation rate of the restriction enzyme digested fragments, we normalized the read counts using a matrix normalization method by Knight and Ruiz (Knight and Ruiz, 2012) as used by Rao et al. (Rao et al., 2014).

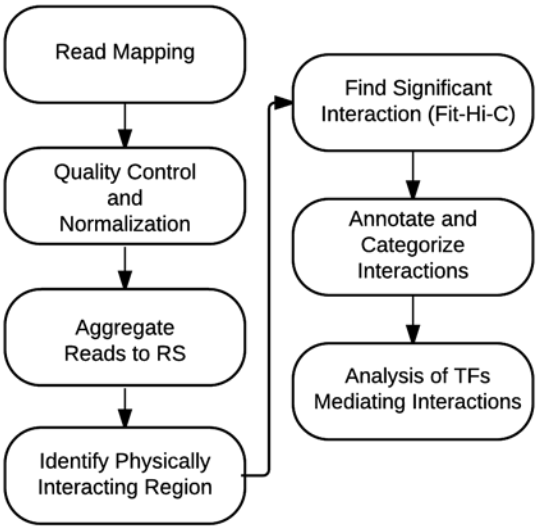



Figure 5-1. Hi-C re-analysis workflow for finding physically-interacting regions (PIRs) and identifying protein factors mediating regulatory interactions. RS: Restriction Site.

Table 5-1. Hi-C data and mapping result.

| Sample | Total sequenced paired-end reads | Contiguously mapped pair-end reads | % | Paired-end reads with chimeric reads involved | % | Total mapped % |
|---|---|---|---|---|---|---|
| Rep1 | 2,971,864,405 | 1,916,327,449 | 64.5 | 382,768,558 | 12.9 | 77.4 |
| Rep2 | 2,623,020,446 | 1,668,938,202 | 63.6 | 410,058,030 | 15.6 | 79.3 |

Since the resolution for identifying interacting DNA regions in the Hi-C protocol is defined by the restriction enzyme cutting frequency (Jin et al., 2013), we developed a strategy to more precisely resolve PIRs as DNA regions flanked by restriction sites (RSs) on both sides, and those observed to be consistently cleaved/ligated in Hi-C sequencing library. To find these flanking RSs, we utilized the information from Hi-C reads, including mapping coordinates, distances to their nearest RSs, strand orientations (+/-) of mapped read-pairs, and relative locations of DNA interaction sites with respect to mapped reads (see **Section 5.5**, **Figure 5-2**). This strategy allowed us to more precisely localize interacting DNA regions where chromosomal physical interaction is mediated by protein binding. Compared to the proposed dynamic binning, typical fixed-width binning of the genome (e.g. 1 kb, 5 kb, 10 kb, or 1 Mb) identifies chromosomal interaction sites using locations of the mapped reads (Jin et al., 2013; Rao et al., 2014; Yaffe and Tanay, 2011) and may have limited accuracy in detecting the precise location of actual interacting DNA regions.

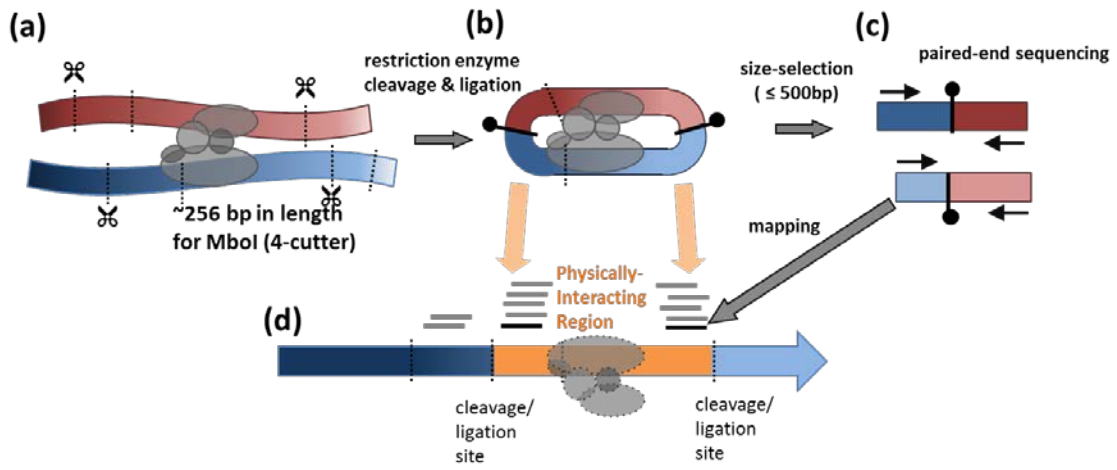



Figure 5-2. Hi-C model and identification of DNA physically-interacting region.

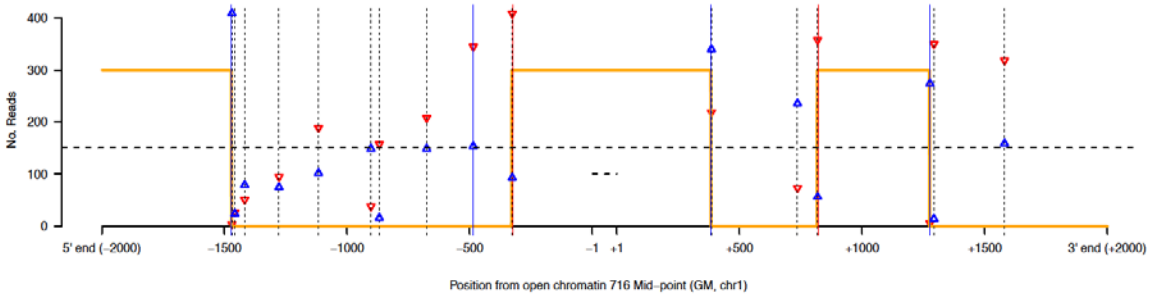
(a) Crosslinked DNA–DNA interaction mediated by protein complexes; red and blue boxes: interacting DNA fragments; group of ellipses: protein complex; doted lines: restriction sites; black vertical arrows: restriction enzyme cut sites. (b) Ligated DNA fragments. The ligation sites are formed by the restriction enzyme cut sites. (c) Sequencing library built after reverse crosslinking, sonication, and DNA fragment size selection (300–500 bp). Paired-end reads from the opposite strands of the ligated DNA construct (horizontal black arrows) is reported by the sequencer. Note, a read could span through the ligation junction (left read of the upper DNA construct) as chimeric read, or get very close to the ligation junction (right read of the upper fragment or both reads from the bottom fragment). (d) Landscape of the reads piling up along the chromosome and the locus of physically-interacting region (PIR). Each read represents a cleavage/ligation site and the read resides in-between the cleavage/ligation site and the actual protein-binding site. The PIR is determined by two proximal consistently cut and ligated restriction sites.

The consistent cleavage/ligation site corresponds to the restriction site with local maximum number of reads. Blue box with arrow: orientation of the genome (small coordinate to large, pointing from p-terminus to q-terminus); gray horizontal lines: reads; gray vertical lines: consistent cleavage/ligation site; dotted ellipses: inferred transcription factor binding locus. Orange line: identified PIR. We only show the blue chromosome for simplicity and its paired physically-interacting region is not shown.

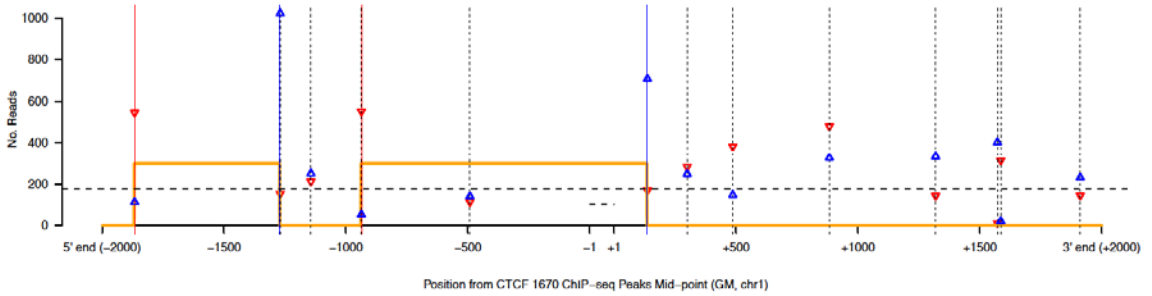
In total, we called 1,785,342 PIRs from all 22 autosomes (chromosomes 1–22) and X-chromosome, with average length of 994 nucleotides (i.e. 1–2 restriction fragments in length). The developed strategy provides a way to dynamically partition the genome into interacting regions. The identified PIRs cover 58.4% of the 23 chromosomes, and each region on average spans 1–2 restriction fragments. Note that PIRs do not necessarily correspond to the fixed-width uniform binning regions and detecting PIRs could greatly increase the resolution and more accurately determine the real DNA–DNA interacting sites compared to fixed-width binning. This is because fixed-width binning could falsely aggregate Hi-C reads to a bin while the read would have suggested the interacting site is located in its upstream or downstream neighboring region. Utilizing the relative PIR position from Hi-C reads becomes more critical with the increase in the resolution (e.g., from 10 kb to 1 kb or even less) to avoid assigning reads to a wrong DNA region. We note that at high resolution (1 kb or less) the interaction site is equally likely to be located in the assigned bin or its upstream or downstream bins, i.e. there is 2/3 chance of false positive discoveries when using uniform binning.

We initially evaluated our PIRs by investigating the overlap with known transcription factor binding sites genome-wide (**Figure 5-3a and b**). We first looked at open chromatin regions from ENCODE (Bernstein et al., 2012) that are likely to be accessible by DNA binding proteins. We found that identified PIRs cover 194,314 out of 231,242 (84.0%) open chromatin regions. On average, 79.1% of the open chromatin region is covered by PIR. Investigating whether PIRs are enriched in CTCF binding sites, we found that PIRs align well with over 88% (39,547 out of 44,597) of CTCF ChIP-seq peaks from ENCODE. This finding is consistent with studies that suggest CTCF has a role in mediating chromatin interactions (Phillips and Corces, 2009).

(a)




(b)




(c)

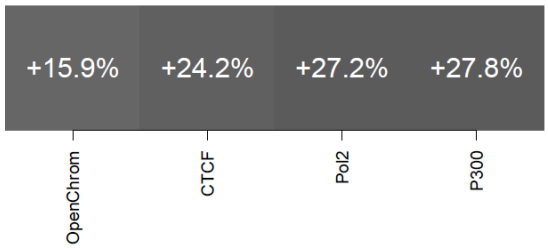



Figure 5-3. PIRs cover (a) open chromatin and (b) CTCF binding sites. (c) Increase percentage test for DNA regions overlapping with open chromatin, CTCF, and occupancies of RNA Polymerase II and p300.

### 5.3.2 Detecting and annotating significant regulatory interactions

To detect significant pairs of PIRs at the chromosomal level, we applied Fit-Hi-C by considering the normalized read count and the linear distance (in nucleotides) of the pairs of interaction sites (Ay et al., 2014) (see **Section 5.5**). For chromosomes 1–22 and X-chromosome, we found that 1,194,010 out of 334,305,544 intra-chromosomal PIR–PIR interacting pairs (>5 kb apart) are significantly interacting (adjusted *P* value ≤ 0.05).

Comparing our data to the Hi-C loci identified by Rao et al. (10 kb binned DNA loci interacting with higher contact frequency than their linear neighborhood) (Rao et al., 2014), we re-discovered 11,793 out of their 12,278 loci (96.0%) corresponding to 35,390 significantly interacting PIRs. The majority (4,496 out of 8.054 (55.8%)) of their Hi-C interactions were also re-discovered by us and correspond to 45,851 significant PIR–PIR interactions with over 10 interactions per each of the 10 kb–10 kb loops on average. Since the methods for identifying their interacting Hi-C peaks and significant PIR interactions are fundamentally different (see **Section 5.5**), we did not expect to re-discover all Hi-C peaks called by Rao et al. Due to the high resolution used in our PIR discovery (average ~1 kb as opposed to 10 kb resolution used by Rao et al), we got more precise interaction loci with protein-binding events. In addition to the 10 kb DNA–DNA interactions from Rao et al., we also discovered 1,148,159 more PIR–PIR interactions (96% of our significant interacting pairs). The set of interactions we discovered would correspond to both architectural genome interactions as well as regulatory interactions.

We found the significantly interacting PIRs tend to be enriched for open chromatin regions, and regulatory elements associated with histone modifications such as H3K4me1, H3K4me2, etc. (**Figure 5-4**). This is consistent with findings from us and others showing that DNA loops can be associated with regulatory elements (Hwang et al., 2013; Rao et al., 2014).

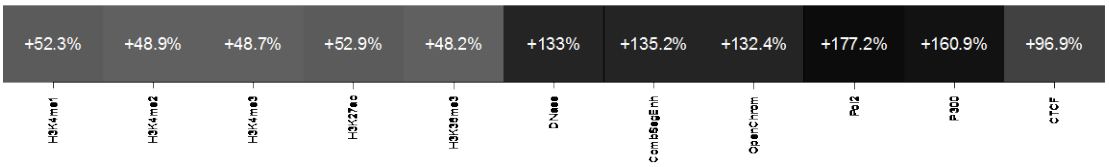



Figure 5-4. Regulatory epigenetic marks are enriched at PIRs with significant intra-chromosomal interactions compared to all PIRs as the background (including the ones that are not significantly interacting with other PIRs).

### 5.3.3 Interactions between regulatory elements overrepresented in PIR–PIR interactions

To further analyze the nature of the detected interactions, we integrated annotations of Genome Segmentation tracks from ENCODE (Hoffman et al., 2013) and RefSeq Genes tracks from UCSC genome browser (Pruitt et al., 2005) and classified each of the significant interacting PIRs as enhancer, promoter, exon, intron, or intergenic site (in this prioritized order). Since enhancers are generally found in open chromatin regions and marked by histone modifications such as H3K4me1 and H3K27ac,(Calo and Wysocka, 2013; Thurman et al., 2012), we annotated PIRs as enhancer elements if they interact with at least one promoter region and overlapped with an open chromatin site and either H3K4me1 or H3K27ac or both. We excluded PIRs as enhancers if they overlapped DNA regions marked by H3K4me3 (active promoter mark) and H3K27me3 (repressive mark). In addition, the promoter-interacting PIRs intersecting with an “E” (enhancer) or “WE” (weak enhancer) annotation from Genome Segmentation tracks were also considered as enhancers.

We found that the percentage of regulatory interactions (enhancer–promoter, enhancer–enhancer, or promoter–promoter pairs) accounted for 5% of the significant interactions, and are significantly more enriched compared to all detectable interactions (only 1%) (**Table 5-2** and **Table 5-3**). This indicates that regulatory interactions may be more stable than other, non-regulatory interactions.

Table 5-2. Annotation quantities for significant interactions.

| Type of PIR | Enhancer | Promoter | Exon | Intron | Intergenic |
|---|---|---|---|---|---|
| Enhancer | 42,140 | 11,848 | 51,351 | 112,308 | 114,926 |
| Promoter | 11,848 | 2,098 | 18,695 | 24,468 | 23,548 |
| Exon | 51,351 | 18,695 | 45,814 | 116,035 | 91,343 |
| Intron | 112,308 | 24,468 | 116,035 | 158,513 | 160,540 |
| Intergenic | 114,926 | 23,548 | 91,343 | 160,540 | 220,383 |
| Sum | 332,573 | 80,657 | 323,238 | 571,864 | 610,740 |

Table 5-3. Annotation quantities on all interactions (including non-significant interactions).

| Type of PIR | Enhancer | Promoter | Exon | Intron | Intergenic |
|---|---|---|---|---|---|
| Enhancer | 2,301,282 | 799,815 | 5,033,846 | 16,116,564 | 18,761,453 |
| Promoter | 799,815 | 101,855 | 1,209,335 | 3,133,731 | 3,634,571 |
| Exon | 5,033,846 | 1,209,335 | 4,079,178 | 22,603,506 | 24,460,019 |
| Intron | 16,116,564 | 3,133,731 | 22,603,506 | 46,744,095 | 92,412,481 |
| Intergenic | 18,761,453 | 3,634,571 | 24,460,019 | 92,412,481 | 92,913,813 |
| Sum | 43,012,960 | 8,879,307 | 57,385,884 | 181,010,377 | 32,182,337 |

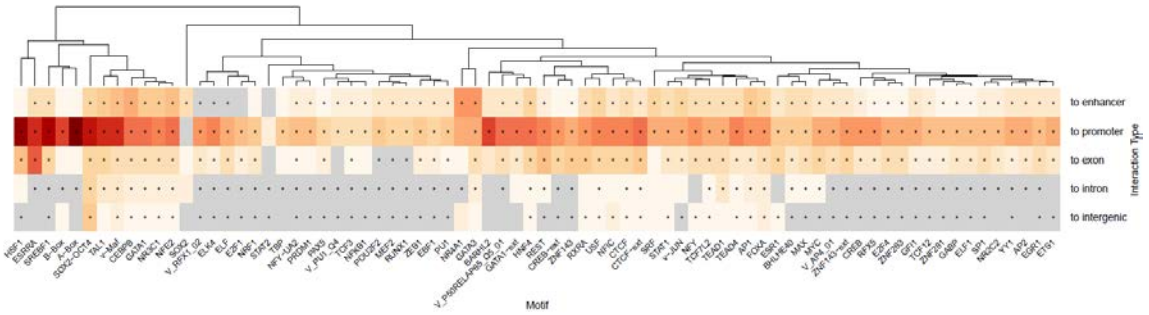
### 5.3.4 Transcription factor binding motif occurrences in PIR–PIR interactions

In order to elucidate the mechanisms underlying the observed enhancer–promoter interactions, we investigated potential transcription factor (TF) binding that likely mediates these regulatory interactions. To achieve this, we interrogated a transcription factor binding site database, Factorbook, which provides a collection of motif loci for 119 DNA-binding proteins based on ChIP-seq experiments that overlap with the computationally-discovered sequence motifs (Wang et al., 2013). This comprehensive repertoire of direct protein–DNA binding motif and loci based on ChIP-seq peak calling allowed us to identify the specific protein-binding events in our PIRs with high confidence.

**Figure 5-5a** shows enrichment for transcription factor binding motifs in enhancers for each of the five categories of the enhancer-interacting DNA regions. The frequency odds ratio is shown for all five different types of enhancer-interacting elements (promoter, enhancer, intron, exon, or intergenic). We found that TF binding is most enriched at enhancers when the enhancer is also interacting with a regulatory element (enhancer and/or promoter) and the frequency odds ratio is even higher when the regulatory element is a promoter, suggesting there is specific TF machinery mediating enhancer–promoter interactions. Additionally, in **Figure 5-5b**, TF binding motifs are highly enriched in promoters interacting with enhancers, promoters, exons, or introns, while most of the motifs are depleted in promoter–intergenic region interactions. We then further investigated the possible mechanisms of long-range enhancer–promoter interactions. As shown in **Figure 5-6**, we constructed the motif–motif pair matrix and screened for significantly enriched motif–motif pairs co-occurring in enhancer–promoter interacting regions. We identified 30 protein factors corresponding to the significantly enriched motif–motif pairs. From the identified protein–protein interactions, we recovered over 86% of known physical interactions according to the protein–protein interaction database, BioGRID (Chatr-Aryamontri et al., 2015; Stark et al., 2006). Among the motif–motif pairs that are enriched, we found that YY1, SP1 and MYC are all enriched in both enhancers and their interacting promoters. YY1 is a transcription factor that mediates

long-distance interactions by binding to an enhancer in B-cell (matching the cell-type we used to discover TF binding for regulatory DNA–DNA interactions). It can recruit cohesin and CTCF for DNAlooping during V(D)J somatic rearrangement of the immunoglobulin loci to produce functional immunoglobulin genes (Atchison, 2014). In addition, SP1 has been shown to function as a link of both sides of enhancer and promoter DNA and is able to form multimers (tetramers and assemblies of multiple tetramers) that facilitate a DNA looping structure (Mastrangelo et al., 1991). Recently, it has been found that another TF, EBF1, is essential for maintaining the identity of B-cells, and this function may be due to EBF1 binding to an enhancer element (Nechanitzky et al., 2013). RUNX1 was discovered to be bound to enhancers in the hematopoietic lineage and may affect transcriptional regulations for leukocyte activation (Laguna et al., 2015). In summary, this suggests that other transcription factors identified by our analysis are also likely to be involved in facilitating and mediating long-range regulatory element interactions.

(a)




(b)

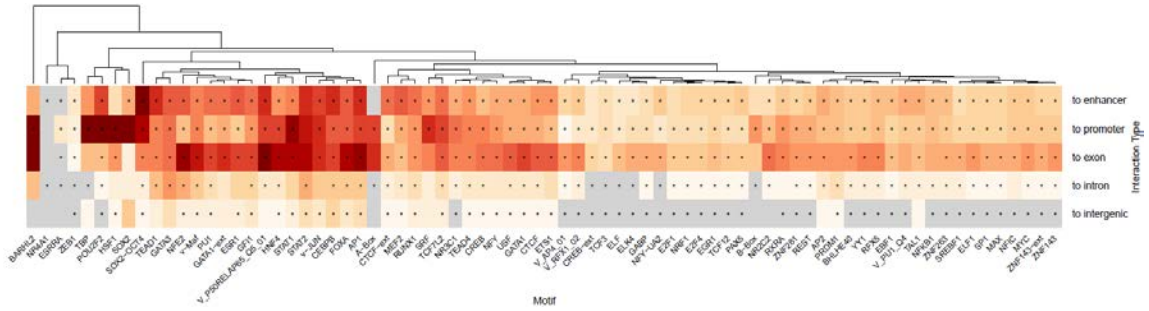



Figure 5-5. (a) Motif discovered in enhancers interacting with other elements. (b) Motif discovered in promoters interacting with other elements.

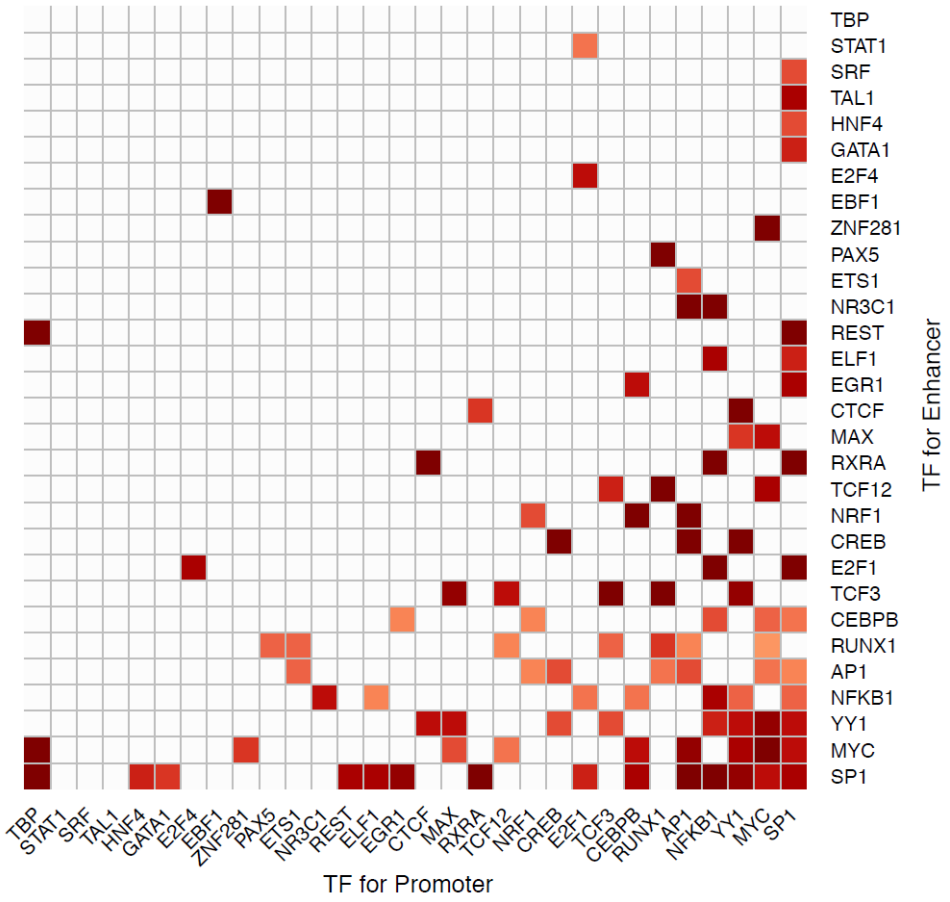



Figure 5-6. Motif pairs discovered in enhancer–promoter interactions with corresponding protein pairs co-exist in the same complex as shown by experimental evidence for physically interactions (e.g. co-IP, two-hybrid, co-localization, etc.).

## 5.4 Discussion

We introduced a novel methodology to discover PIRs in the genome with ~1 kb resolution utilizing Hi-C sequencing data. The proposed dynamic genome partition based on read piling for restriction sites (**Figure 5-2**) allowed us to more precisely delineate interacting DNA regions compared to uniform fixed-width binning shown by others.

We also identified protein-binding sequences that are overrepresented in interactions among regulatory elements, such as enhancer–promoter, enhancer–exon, and promoter–promoter interactions. The specific preferences of the interaction types for TF binding suggests that there is specific transcription factor machinery involved in different types of interactions. With high-confidence ChIP-seq data available for DNA-binding proteins, the power of detecting the transcription factors that are mediating DNA–DNA interactions can be substantially increased.

Previous studies have shown and identified that the DNA binding domain of a protein factor controls how an enhancer selects its cognate target promoter to be regulated in pre-erythroid cells (Deng et al., 2014). The TF–TF interacting pairs for enhancer–promoter interactions identified in this study can suggest general mechanisms on how an enhancer controls the expression of its target gene. Further study of other classes of genic or intergenic interactions (e.g. enhancer–exon, promoter–intron) may help us to elucidate the mechanisms and factors involved in the formation on these interactions. Also, with decreasing sequencing costs, it may become more feasible to perform Hi-C experiments with higher read coverage and in multiple cell and tissue types. Repeating and perhaps re-running our novel methodology detecting TF binding along PIRs would allow the discovery of protein complexes that identify the specific genomic bridges within each specific cell/tissue types.

## 5.5 Materials and methods

***Hi-C Data and mapping to the genome***

For our analysis, we used Hi-C datasets (Rao et al., 2014) from a lymphoblastoid cell line (GEO database under accession number GSE63525). We combined 16 in situ Hi-C sequencing libraries (HIC003 through HIC018) as the primary dataset. Libraries HIC020 through HIC029 were used as replicates. We first aligned the paired-end reads by STAR aligner (Dobin et al., 2013) to a standardized human genome assembly (GRCh37/hg19) (Lander et al., 2001) and only allowed for uniquely mapped reads. We first mapped each single-end (101 nt in this study) of a read-pair separately (i.e. as two independent single-end reads). When a single-end read spanned through a ligation junction and split-mapped to two distant genomic loci, we reported all such single-end reads as chimeric reads. Both halves of a chimeric read were required to map uniquely and have a minimum mapped length of 22 nt. All other reads mapped to a single contiguous locus along the genome were reported as contiguous, non-chimeric reads. We then paired these separately mapped single-end reads with their corresponding paired-end partners. For those paired-end reads with a chimeric read involved, we required that the pairing partner of the chimeric read (a single-end read) mapped in the proximity of one of the two split halves spanned by the chimeric read.

***Hi-C data pre-filtering***

To filter out reads resulting from self-ligated, un-cut or re-ligated DNA products, we removed the read pairs that are mapped to two loci that are less than 5 kb apart. After the filtering, the four strand combinations for the remaining mapped read-pairs had almost equal observed probability (**Table 5-4**), indicating that the remaining read pairs represented DNA

fragments that were legitimately digested and ligated as expected according to Hi-C protocol (Jin et al., 2013).

Additionally, to avoid any biases in detection of the region that cannot be mapped as a unique genomic locus, we also removed from the analysis restriction sites (RSs) that have mappability of less than 0.8 (see **Section 5.5**). We found that 96% of the RSs have mappability higher than 0.8, suggesting that most of RSs had high mappability given a relatively long read length (101 nts).

Table 5-4. Equal probability strand combination after (5kb) distance filtering

| All read pairs | | | |
|---|---|---|---|
| read1 | read2 | count | % |
| + | + | 351,356,711 | 20.4 |
| + | - | 507,746,008 | 29.5 |
| - | + | 505,952,205 | 29.4 |
| - | - | 355,039,391 | 20.6 |

| Read pair distance >= 5000 bp | | | |
|---|---|---|---|
| read1 | read2 | count | % |
| + | + | 329,957,629 | 24.9 |
| + | - | 331,958,821 | 25.0 |
| - | + | 330,499,655 | 24.9 |
| - | - | 333,438,805 | 25.1 |

***Mappability***

To calculate the mappability spanning ±500 bp flanking region of each RS, we first generated simulated reads with 1 bp sliding window with read length of 101 bp (**Figure 5-7**). The mappability of each RS flanking region is calculated as the percentage of uniquely mapped simulated reads among all simulated reads within 500 bp upstream and downstream of the RS.

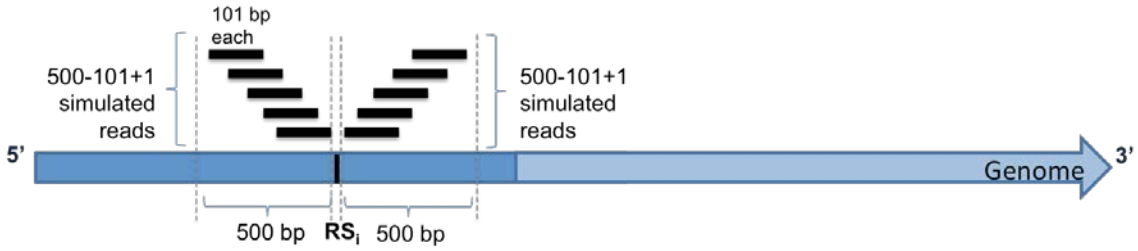



Figure 5-7. Definition of mappability.

***Hi-C read normalization***

We used matrix balancing method (Knight and Ruiz, 2012) at 1 kb resolution (Rao et al., 2014) to normalize the read counts using KR normalization vector. Each observed Hi-C read-pair was assigned a weight based on the mapped genomic locations of the corresponding single-end reads in the pair.

***Identification of physically-interacting regions***

To identify physically-interacting DNA regions (PIRs), we utilized the idea that each single-end Hi-C read is always located in the proximity of a restriction site (RS) that serves as the cleavage/ligation site in the Hi-C protocol.

The RSs correspond to sites in the genomic DNA containing sequence that can be recognized by the restriction enzyme, e.g., “GATC” for restriction enzyme MboI. Each of the single-end reads is indicating a possible physically-interacting region (e.g. transcription factor binding locus) from the cleavage/ligation site. In other words, once we mapped the reads and determined their cleavage/ligation sites, we can infer the relative position (upstream or downstream from the RS) for the DNA-interacting region.

The cleavage/ligation site is attainable from the mapping information of Hi-C sequencing paired-end reads because (1) proper DNA ligation is required to form a phosphodiester bond between the 5' phosphate of the donor DNA and the 3' hydroxyl of the acceptor DNA, and (2) the strand orientation pattern reported by Illumina sequencer restricts the combinations of flanking upstream and downstream regions from the ligation site for each read pair. The workflow for identifying all PIRs along the genome includes three major phases: (I) find all read-pairs with ligation junctions, (II) identify physically-interacting regions, and (III) find PIR-PIR interactions. Each phase is described below.

*I.* *Find all read-pairs with ligation junctions*

As shown in **Figure 5-8**, we investigated pair of strand orientations and the distances from mapped reads to their nearest (both upstream and downstream) restriction sites to achieve candidate ligation junctions. For instance, for a read-pair with strand orientations reported as +/- or -/+, the candidate ligation junction would be formed by two RSs, one RS laying upstream from one read of the read-pair and the other RS laying downstream from the other end read of the read-pair. Similarly, for read-pair with strand orientations as +/+ or -/-, the candidate ligation junctions are expected to be either both located upstream from each of the single-end read of a read-pair or both being downstream from each of the single-end read of the read-pair. Finally, by estimating the distances from both reads of a read-pair to their nearest upstream and/or downstream RSs, we can identify the feasible ligation case(s) from two candidate ligation junctions. We report that the only proper case that fulfills the criteria is while sum-of-distances of the mapping positions is shorter than the maximum Hi-C fragment length (in this study, 500 bp), which is determined by the size-selection step from the Hi-C experiment. Note if both candidate ligation junctions have sum-of-distances meeting the size selection threshold, we discarded the read-pair for further analysis to avoid ambiguity and for simplicity. For the paired-end reads with single ligation junction determined, we report the pairs of RSs that form the junction.

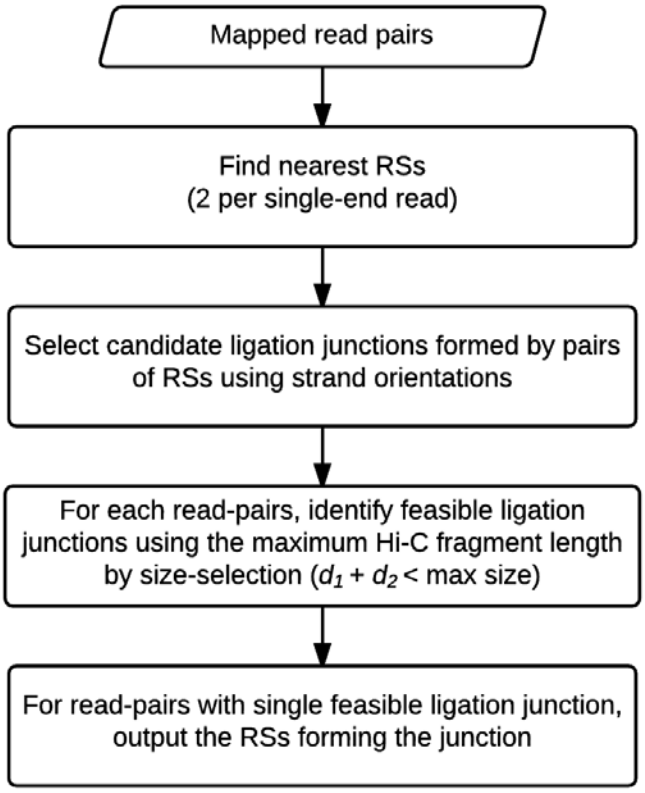



Figure 5-8. Find all read-pairs with ligation junctions.

*II.* *Identify physically-interacting regions*

With the identified RSs that form ligation junctions, we further identified physically-interacting regions (**Figure 5-9**). First, we note the sum of upstream and downstream read counts (single-end reads from read-pairs) for each RS. We clustered RSs separately for upstream and downstream read counts by thresholds of the maximum gap ($d_{cluster}$) and the minimum read ($r_{threshold}$). The maximum gap is defined as the third-quantile of the restriction fragment distance distribution, and the minimum read requirement is defined as the median of the normalized read distribution for each chromosome. Within each corresponding cluster, we identified the RSs with the maximum read count as the candidate flanking ends for a PIR. Finally, we matched the nearest upstream and downstream candidate flanking ends with a max-gap algorithm (in this study, the max-gap is 4000 bp), and reported the PIRs that are bracketed by the flanking ends.

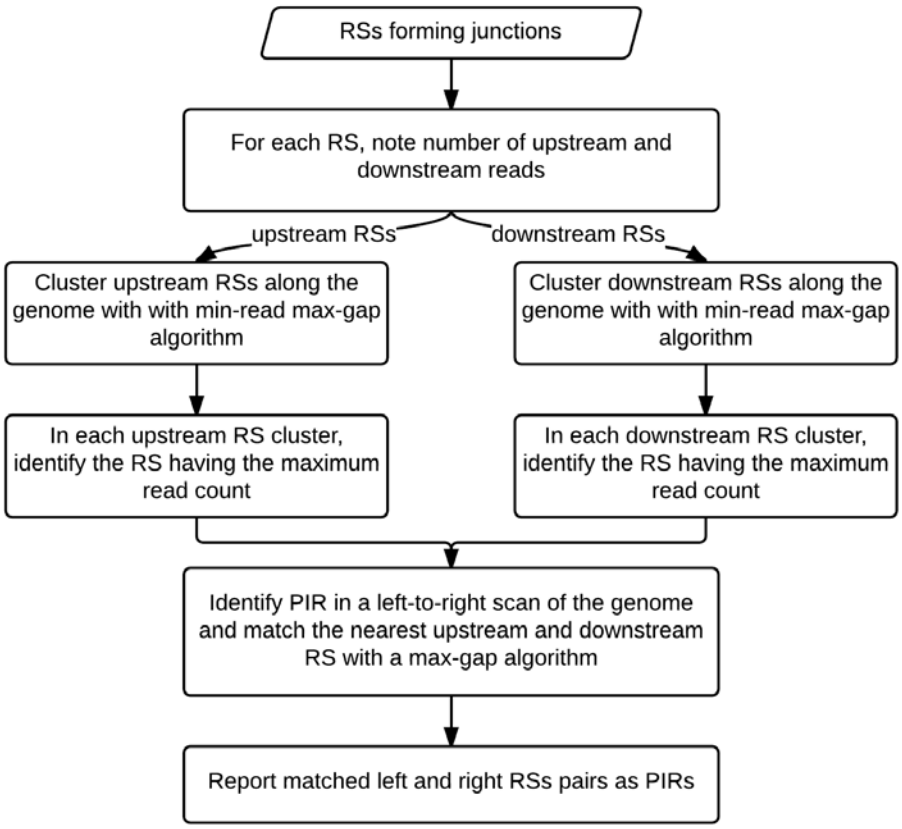



Figure 5-9. Identify physically-interacting regions.

*III.* *Find all PIR–PIR interactions*

We found the interactions between PIRs by tracing the Hi-C read-pairs that participated in the identification of PIRs (**Figure 5-10**). For each PIR, we listed all the corresponding paired-end read IDs (read names) of single-end reads that piled up on both side of the RS clusters those are flanking it.

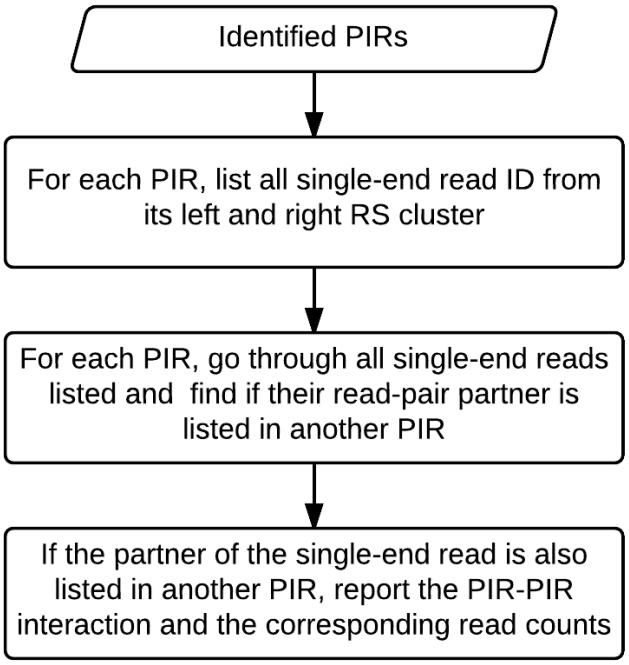



Figure 5-10. Find all PIR–PIR interactions.

***Identifying significant PIR–PIR interactions genome-wide***

To identify significant intra-chromosomal PIR–PIR interaction pairs, we implemented Fit-Hi-C method (Ay et al., 2014) in R. For each of the autosomal chromosomes (1–22) and chromosome X, we split all observed PIR–PIR interactions into 2,000 distance groups according to the linear distance (in nucleotides) between interacting PIRs. We filtered out the PIR pairs that are less than 5,000 nucleotides apart. For each distance group, we calculated the average distance and the average normalized read counts of the interacting PIRs. With the 2,000 aggregated data points, we fit the normalized read counts by the function of distance using *smooth.spline* function in R. After the first spline fitting, we removed the outliers as described in (Ay et al., 2014) and fit the second spline function. We then reported PIR–PIR interactions that are significant after Benjamini–Hochberg correction (adjusted *P* values <= 0.05).

***Overlap with DNA loops (Rao et al.)***

We compared PIR–PIR interactions in our study with the set of DNA loops identified in (Rao et al., 2014) using HiCCUPS. We downloaded the set of loops and Hi-C loci (DNA regions that participate in the formation of significant DNA loops) at 10 kb resolution from the GEO database under accession number GSE63525.

We first measured how well the set of PIRs overlapped with the set of HI-C loci. Our PIRs overlapped with 11,793 of the 12,278 (96.0%) Hi-C loci; the overlap corresponded to 35,390 PIRs we identified. We then overlapped the set of PIR–PIR interaction with the set of DNA loops (Hi-C peak locus to Hi-C peak locus). The significant number of overlapped DNA loops (4,496 out of 8,054, 55.8%) corresponded to 45,851 significant PIR–PIR interactions (10 PIR–PIR interactions per DNA loop on average).

***Functional and genomic annotation data***

We downloaded the ChIP-seq peak data for histone modifications (H3K4me1, H3K4me2, H3K4me3, H3K27ac, and H3K36me3), DNase I hypersensitive sites, transcription factors (RNA Polymerase II, p300) and DNA-binding proteins (i.e. CTCF), and annotation tracks (Combined Segmentation and open chromatin) from UCSC Genome Browser (hg19 assembly). For the histone modifications, RNA Polymerase II, p300, and CTCF, we downloaded the ChIP-seq uniform peaks from UCSC Genome Browser 2011 freeze. The open chromatin track from ENCODE is a synthesis of multiple assays such as DNaseI-seq, FAIRE-seq, and ChIP-seq data (OpenChromSynth, release 2 (Feb 2012)).

***Enrichment analysis and histone modification annotation***

To estimate the extent of overlap between PIRs and regulatory epigenetic marks genome-wide, we calculated the sum of overlapped nucleotides between PIRs and each signal track (regulatory/epigenetic mark) genome-wide as the observed value. We sampled (1000 times) random genomic regions from the genome with length distribution matched with the length of PIRs. We calculated the average of 1,000 sums of overlaps between the sampled regions and each of the signal tracks. We then reported the percentage differences between the observed value and the averaged value from the background as the enrichment of the PIRs for each of the signal tracks. All the region intersections were performed using bedtools (Quinlan and Hall, 2010).

***Regulatory and genetic annotation of the interacting site***

We annotated called PIRs as enhancers, promoters, exons, introns, or intergenic elements. To do this, the annotation for enhancer regions was downloaded from UCSC genome

segmentation track (Hoffman et al., 2013) and the gene models were downloaded from RefSeq. We also downloaded the ChIP-seq peak data for H3K4me1, H3K4me3, H3K27ac, and H3K27me3 epigenetic marks and open chromatin tracks from UCSC uniform peak calling track.

We annotated enhancers as the promoter-interacting PIRs that overlapped the enhancer (marked as "E") or weak enhancer (marked as "WE") annotation from the genome segmentation track (Hoffman et al., 2013). We also annotated all promoter-interacting PIRs as an enhancer if they overlapped an open chromatin region with H3K4me1 or H3K27ac ChIP-seq peak, while not overlapping with H3K4me3 and H3K27me3 peaks. The rest of the PIRs were annotated as promoters, exons, introns, and intergenic elements. To perform genomic annotations, we used RefSeq gene models (hg19 assembly). The promoters were defined as 500 bp-long regions upstream of the RefSeq TSS of protein-coding genes. We then annotated PIRs with the prioritized order of promoters, exonic, intronic, or intergenic elements.

***Motif analysis of the annotated PIR–PIR interactions***

To identify PIRs with evidence of protein factor-binding, we used Factorbook data (Wang et al., 2013) that integrates ChIP-seq experimental data from ENCODE with computationally-predicted TF binding sites to comprehensively survey protein-DNA binding genome-wide. The Factorbook data were obtained from UCSC hg19 database (factorbookMotifPos table, release 4). The data contains a list of 161 factors and their motifs discovered from 91 cell types. We focused on 76 known DNA-binding transcription factors. We filtered out the TFs with less than 10 binding sites with PIRs genome-wide. For each PIR, we reported all TFs that have at least one binding site within that PIR.

We also reported enrichment for each of the surveyed binding motifs in PIR–PIR interactions. To do this, we categorized PIR–PIR interactions according to the classes of

interacting PIR elements (enhancers, promoters, exons, introns, or intergenic elements). We estimated binding-motif enrichment as observed/expected frequency odds ratio. We computed expected probability as the probability of the first type of a PIR having the first motif times the probability of the second type of a PIR having the second motif as shown in the equation below.

$$\text{Prob}(M_k \text{ observed in } C_i, C_i) = P(M_k|C_i, C_j)$$

$$= P(M_k|C_i) \times P(C_j), \quad \text{since } C_j \text{ does not does not required to contain } M_k$$

$$= \frac{P(M_k, C_i)}{P(C_i)} \times P(C_j)$$

$$= \frac{\#(C_i \text{ containing } M_k)}{\#C_i} \times \frac{\#C_j}{\#(\text{total elements})}$$

We also estimated the possibility for enhancer and promoter interaction with the motif preferences on both sides of PIRs with the equation below:

$$\text{Prob}(M_k, M_l \text{ observed in } C_i, C_i) = P(M_k|C_i) \times P(M_l|C_j)$$

$$= \frac{P(M_k, C_i)}{P(C_i)} \times \frac{P(M_l, C_j)}{P(C_j)}$$

$$= \frac{\#(C_\text{i} \text{ containing } M_k)}{\#C_i} \times \frac{\#(C_j \text{ containing } M_l)}{\#C_j}$$

We performed binomial distribution test to report the significance of observed binding motifs in each type of PIR–PIR interaction.

# Chapter 6 : Conclusions and future directions

In this dissertation, I have described novel approaches for delineating DNA units and their chromatin interactions using Hi-C data, with the goal of providing a better understanding of enhancer–promoter interactions. I have detected DNA-interacting units in Hi-C data by using a number of different approaches (identification of interaction hotspots, use of restriction fragment-based technologies, and identification of physically-interacting regions) to study their global characteristics in long-range regulation of the human genome. I also developed an automated pipeline with improved mapping of raw Hi-C data and automatic identification of candidate enhancer–promoter interactions.

## 6.1 Summary of findings

In **Chapter 2**, I developed an approach to identify DNA interacting hotspots, which are genomic regions that have higher read coverage in Hi-C datasets than expected, in the only publically-available human Hi-C data from two cell-types, GM06990 (B-Lymphocytes) and K562 (myelogenous leukemia cells) (Lieberman-Aiden et al., 2009). I then identified hotspot-interacting pairs within the top 5% of contact strength (with more than 1 supporting read-pair) and interacted with a promoter region as candidate interacting pairs. Additionally, I extracted a candidate list of active enhancer–promoter interactions based on their genetic and epigenetic annotations. As expected, I found that hotspots with the top contact strength are strongly enriched with those histone modifications known to be associated with active regulatory elements, compared with other remaining hotspots that are interacting with lower contact strength.

In **Chapter 3**, I evaluated and explored the interactions of candidate enhancer elements and their target genes with p300 and RNA Polymerase II, respectively. I found that candidate

enhancer elements are preferred for p300 occupation. p300 is a co-activator that can occupied enhancer elements and increases expression of its target genes (Eckner et al., 1994; Maston et al., 2006). I also found that promoters of protein-coding genes that are touched by an enhancer are significantly enriched for RNA polymerase II occupation, an indicator of transcriptional activity. I additionally found that candidate enhancer elements are more conserved than their flanking regions and that enhancer-touched genes are expressed in a tissue-specific manner. Supporting genomic evidence suggests that the enhancer–target gene pairs I identified based on Hi-C data could be a good candidate list for future validation for both their regulatory function and tissue-specificity.

Following the successful identification of enhancer–target genes using Hi-C data, in **Chapter 4**, I developed and implemented a high-throughput identification pipeline for promoter interacting enhancer elements (HIPPIE) that can take raw Hi-C reads as input and identify candidate enhancer elements. It runs efficiently on typical Linux servers and grid computing environments and is available as open-source software. Since higher-coverage Hi-C data has come out (Jin et al., 2013), making it possible to study DNA-interacting units with higher resolution at the restriction fragment level, I implemented both hotspot-based and restriction fragment-based methods for detecting DNA-interacting regions.

Finally, in **Chapter 5**, in addition to identifying genome-wide enhancer–target gene pairs, I was also interested in uncovering the mechanisms of how regulatory elements form interactions and the proteins that mediate these interactions. I shifted to the latest Hi-C data by Rao et al. (Rao et al., 2014), which had the highest sequencing depth of any published dataset to date. I refined my approach for calling DNA-interacting units from hotspot or fragment-based methods to detect physically-interacting DNA regions (PIRs). This Hi-C data was generated with a more frequently cutting endonuclease (4-cutter) than previous data sets (6-cutter). This allows better resolution for pin-pointing the actual DNA–DNA interaction sites as well as protein complex

binding sites. I then discovered a set of enhancer–promoter interactions and annotated those using functional genomics data. My analysis revealed the putative transcription factor binding events that facilitate and mediate these long-range regulatory element interactions.

## 6.2 Future directions: applications to genetic research

By developing novel strategies to determine DNA interacting regions and DNA–DNA physical interactions in Hi-C datasets, my approaches identified putative enhancer elements and their target gene promoters genome-wide. I also explored and discovered the characteristics of those enhancer–promoter interactions as well as transcription factor binding motifs that are preferred within enhancer–promoter interactions. These identifications and characteristics can help improve building tools for predicting more regulatory interactions, and facilitate the interpretation of how non-coding variants result in human diseases. In the following sections, we summarize three areas showing how our studies can lead to further discoveries.

### 6.2.1 Predicting regulatory interactions

It remains challenging to predict the target genes of enhancers computationally. Common approaches (as discussed in **Section 1.4**) are to assign the nearest promoter of an enhancer element as its target gene, to correlate the DNase I hypersensitive sites of enhancer and promoter regions across cell- and tissue- types (Thurman et al., 2012), or to utilize pairwise expression correlation between enhancer RNAs with messenger RNAs (Andersson et al., 2014). Recently, it has been shown that one can integrate multiple epigenomics data sets (e.g. integrated histone marks or transcription factor binding motifs) to predict enhancer–promoter pairs using machine learning methods (Whalen et al., 2015). The physically interacting enhancer–

promoter pairs identified in **Chapter 2** and **Chapter 5** can serve as positive training samples for the community to train their enhancer–promoter prediction methods.

Additionally, in this study, I have characterized physically interacting enhancers and their target promoters in terms of their p300 occupancy, conservation level, RNA Polymerase II binding, and linear distances as described in **Chapter 3**. I have also discovered pairs of transcription factor binding motifs corresponding to enhancer–promoter interactions as described in **Chapter 5**. Integrating the additional information learned from this study as attributes of interacting enhancers and promoters could enable us to build a better computational prediction model for studying unknown pairs of enhancer–promoter interactions in any cell-type of interest by utilizing general knowledge (e.g. overrepresented TFBS, conservation level, linear distance, and so on) and cell-type-specific information (e.g. p300 occupancy, RNA Polymerase II occupancy, transcription factor expression level, and so on). One can run the same analysis I performed to discover TF pairs from Hi-C data in several other cell types and identify the TFs that are enriched across cell types in order to find a general mechanism of enhancer–promoter interactions. Moreover, by integrating expression data of the transcription factors that are involved in enhancer–promoter interactions across cell-types could improve predictions of whether a given interaction involving those factors will occur. In sum, this study provides the basis for a resource of regulatory information that can help shed light on long-range interactions.

### 6.2.2 Interpreting disease-related non-coding genetic variants using enhancer–promoter interacting pair information

The majority of GWAS-identified genetic variants reside in non-coding regions of the genome (Manolio et al., 2009), making it challenging to understand how these variants cause diseases or phenotypes. One possible mechanism is that non-coding variants lie in an enhancer

element, and affect phenotypes through regulating distal protein-coding gene expression. With the candidate enhancer elements detected in **Chapter 2** and **Chapter 5,** we can predict which GWAS hits reside in enhancer elements and affect phenotypes by that mechanism.

The most common method for finding the possible candidate target genes of enhancers is taking their nearest genes (as discussed in **Section 1.4.1**). However, as mentioned in **Chapter 1**, the enhancer variants located within the *Lmbr* gene affect the expression of the *Shh* gene, that lies 1 Mb away, but not the expression of *Lmbr* (Lettice et al., 2003). If a GWAS hit is predicted to lie within this enhancer element, using its closest chromosomal neighbor may identify the wrong candidate target gene. Other methods, such as correlations of epigenetic marks or expression values between DNA regions across cell and tissue types could help find the candidate targets of an enhancer element, but none of these provide direct information about whether the target gene is physically interacting with the enhancer, as correlation is not causation. Alternately, the enhancer–target gene pairs predicted in **Chapter 2** and **Chapter 5** by Hi-C data can help to deduce candidate mechanisms regarding the non-coding variants discovered by GWAS through annotating the target gene affected by an enhancer variant.

### 6.2.3 Cell differentiation and tissue-specificity long-range interactions

For the work described throughout **Chapter 2**–**Chapter 5**, we used publicly available genomic data. It may be that there are other chromosomal features that would have a strong association with the long-range regulation of enhancers and target promoters, although this data is not yet available. In addition, increasing amounts of Hi-C data for other cell- or tissue-types with high read depth and good quality are rapidly becoming available. This may make it more feasible to compare the enhancer–promoter relationships across different cell conditions and thereby understand the interaction specificity among cell-types, tissue-types, disease cells versus normal cells, or stem cells versus differentiated cells. By incorporating other available data, performing

analyses similar to those described in **Chapter 2** and **Chapter 5** would enable the discovery of features associated with these interactions and thereby better understand the long-range regulatory elements that are involved in determining and maintaining specific cell fates.

## 6.3 Concluding remarks

Enhancers are a major group of functional DNA elements that play a fundamental role in cell development and contribute to disease when malfunctioning. In this work, I have developed computational methods for analyzing Hi-C data to identify enhancer–promoter interactions. Our approaches include an improved mapping strategy, better DNA-interaction unit binning, and the discovery of specific motifs involved in enhancer and promoter interactions on a genome-wide scale. I also developed an automatic pipeline (HIPPIE), which can be run efficiently on typical Linux servers, for the community to process their own or publically available Hi-C data and identify regulatory elements.

# BIBLIOGRAPHY


Ahituv, N., Prabhakar, S., Poulin, F., Rubin, E.M., and Couronne, O. (2005). Mapping cis-regulatory domains in the human genome using multi-species conservation of synteny. Hum. Mol. Genet. *14*, 3057–3063.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., et al. (2014). An atlas of active enhancers across human cell types and tissues. Nature *507*, 455–461.

Atchison, M.L. (2014). Function of YY1 in Long-Distance DNA Interactions. Front. Immunol. *5*, 45.

Ay, F., and Noble, W.S. (2015). Analysis methods for studying the 3D architecture of the genome. Genome Biol. *16*, 183.

Ay, F., Bailey, T.L., and Noble, W.S. (2014). Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. Genome Res. *24*, 999–1011.

Banerji, J., Rusconi, S., and Schaffner, W. (1981). Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. Cell *27*, 299–308.

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. Cell *129*, 823–837.

Bernstein, B.E., Birney, E., Dunham, I., Green, E.D., Gunter, C., and Snyder, M. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Blanchette, M., and Tompa, M. (2002). Discovery of regulatory elements by a computational method for phylogenetic footprinting. Genome Res. *12*, 739–748.

Bolstad, B.M., Irizarry, R.A., Astrand, M., and Speed, T.P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics *19*, 185–193.

Burton, J.N., Adey, A., Patwardhan, R.P., Qiu, R., Kitzman, J.O., and Shendure, J. (2013). Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nat. Biotechnol. *31*, 1119–1125.

Calo, E., and Wysocka, J. (2013). Modification of enhancer chromatin: what, how, and why? Mol. Cell *49*, 825–837.

Carroll, D. (2011). Genome engineering with zinc-finger nucleases. Genetics *188*, 773–782.

Catena, R., Tiveron, C., Ronchi, A., Porta, S., Ferri, A., Tatangelo, L., Cavallaro, M., Favaro, R., Ottolenghi, S., Reinbold, R., et al. (2004). Conserved POU binding DNA sites in the Sox2 upstream enhancer regulate gene expression in embryonic and neural stem cells. J. Biol. Chem. *279*, 41846–41857.

Chatr-Aryamontri, A., Breitkreutz, B.-J., Oughtred, R., Boucher, L., Heinicke, S., Chen, D., Stark, C., Breitkreutz, A., Kolas, N., O'Donnell, L., et al. (2015). The BioGRID interaction database: 2015 update. Nucleic Acids Res. *43*, D470–D478.

Chepelev, I., Wei, G., Wangsa, D., Tang, Q., and Zhao, K. (2012). Characterization of genome-wide enhancer-promoter interactions reveals co-expression of interacting genes and modes of higher order chromatin organization. Cell Res. *22*, 490–503.

Cheung, V.G., Conlin, L.K., Weber, T.M., Arcaro, M., Jen, K.-Y., Morley, M., and Spielman, R.S. (2003). Natural variation in human gene expression assessed in lymphoblastoid cells. Nat. Genet. *33*, 422–425.

Chi, X., Chatterjee, P.K., Wilson, W., Zhang, S.-X., Demayo, F.J., and Schwartz, R.J. (2005). Complex cardiac Nkx2-5 gene expression activated by noggin-sensitive enhancers followed by chamber-specific modules. Proc. Natl. Acad. Sci. U. S. A. *102*, 13490–13495.

Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L.A., et al. (2013). Multiplex genome engineering using CRISPR/Cas systems. Science *339*, 819–823.

Crawford, G.E., Holt, I.E., Mullikin, J.C., Tai, D., Blakesley, R., Bouffard, G., Young, A., Masiello, C., Green, E.D., Wolfsberg, T.G., et al. (2004). Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites. Proc. Natl. Acad. Sci. U. S. A. *101*, 992–997.

Crawford, G.E., Holt, I.E., Whittle, J., Webb, B.D., Tai, D., Davis, S., Margulies, E.H., Chen, Y., Bernat, J.A., Ginsburg, D., et al. (2006). Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). Genome Res. *16*, 123–131.

Creyghton, M.P., Cheng, A.W., Welstead, G.G., Kooistra, T., Carey, B.W., Steine, E.J., Hanna, J., Lodato, M. a, Frampton, G.M., Sharp, P. a, et al. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc. Natl. Acad. Sci. U. S. A. *107*, 21931–21936.

Crick, F. (1956). Ideas on Protein Synthesis.

Crick, F. (1970). Central Dogma of Molecular Biology. Nature *227*, 561–563.

De, S., and Michor, F. (2011). DNA replication timing and long-range DNA interactions predict

mutational landscapes of cancer genomes. Nat. Biotechnol. *29*, 1103–1108.

Dekker, J. (2006). The three "C" s of chromosome conformation capture: controls, controls, controls. Nat. Methods *3*, 17–21.

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. Science *295*, 1306–1311.

Deng, W., Rupon, J.W., Krivega, I., Breda, L., Motta, I., Jahn, K.S., Reik, A., Gregory, P.D., Rivella, S., Dean, A., et al. (2014). Reactivation of Developmentally Silenced Globin Genes by Forced Chromatin Looping. Cell *158*, 849–860.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 1–5.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Dostie, J., Richmond, T.A., Arnaout, R.A., Selzer, R.R., Lee, W.L., Honan, T.A., Rubio, E.D., Krumm, A., Lamb, J., Nusbaum, C., et al. (2006). Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome Res. *16*, 1299–1309.

Doudna, J.A., and Charpentier, E. (2014). The new frontier of genome engineering with CRISPR-Cas9. Science (80-. ). *346*, 1258096–1258096.

Duan, Z., Andronescu, M., Schutz, K., McIlwain, S., Kim, Y.J., Lee, C., Shendure, J., Fields, S., Blau, C.A., and Noble, W.S. (2010). A three-dimensional model of the yeast genome. Nature *465*, 363–367.

Eckner, R., Ewen, M.E., Newsome, D., Gerdes, M., DeCaprio, J.A., Lawrence, J.B., and Livingston, D.M. (1994). Molecular cloning and functional analysis of the adenovirus E1A-associated 300-kD protein (p300) reveals a protein with properties of a transcriptional adaptor. Genes Dev. *8*, 869–884.

Edelmann, P., Bornfleth, H., Zink, D., Cremer, T., and Cremer, C. (2001). Morphology and dynamics of chromosome territories in living cells. Biochim. Biophys. Acta - Rev. Cancer *1551*.

Elgar, G., and Vavouri, T. (2008). Tuning in to the signals: noncoding sequence conservation in vertebrate genomes. Trends Genet. *24*, 344–352.

Encode, T., and Consortium, P. (2011). A User's Guide to the Encyclopedia of DNA Elements (ENCODE). PLoS Biol. *9*, e1001046.

Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature *473*, 43–49.

Fehrmann, R.S.N., Jansen, R.C., Veldink, J.H., Westra, H.-J., Arends, D., Bonder, M.J., Fu, J., Deelen, P., Groen, H.J.M., Smolonska, A., et al. (2011). Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. PLoS Genet. *7*, e1002197.

Feng, S., Cokus, S.J., Schubert, V., Zhai, J., Pellegrini, M., and Jacobsen, S.E. (2014). Genome-wide Hi-C analyses in wild-type and mutants reveal high-resolution chromatin interactions in Arabidopsis. Mol. Cell *55*, 694–707.

Fudenberg, G., Getz, G., Meyerson, M., and Mirny, L.A. (2011). High order chromatin architecture shapes the landscape of chromosomal alterations in cancer. Nat. Biotechnol. *29*, 1109–1113.

Fullwood, M., and Ruan, Y. (2009). ChIP-based methods for the identification of long-range chromatin interactions. J. Cell. Biochem. *107*, 30–39.

Fullwood, M.J., Liu, M.H., Pan, Y.F., Liu, J., Xu, H., Mohamed, Y. Bin, Orlov, Y.L., Velkov, S., Ho, A., Mei, P.H., et al. (2009). An oestrogen-receptor-alpha-bound human chromatin interactome. Nature *462*, 58–64.

Geyer, P.K., Green, M.M., and Corces, V.G. (1990). Tissue-specific transcriptional enhancers may act in trans on the gene located in the homologous chromosome: the molecular basis of transvection in Drosophila. EMBO J. *9*, 2247–2256.

Gibcus, J.H., and Dekker, J. (2013). The hierarchy of the 3D genome. Mol. Cell *49*, 773–782.

Gilad, Y., Rifkin, S.A., and Pritchard, J.K. (2008). Revealing the architecture of gene regulation: the promise of eQTL studies. Trends Genet. *24*, 408–415.

Giresi, P.G., Kim, J., McDaniell, R.M., Iyer, V.R., and Lieb, J.D. (2007). FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. Genome Res. *17*, 877–885.

Gross, D.S., and Garrard, W.T. (1988). Nuclease Hypersensitive Sites in Chromatin. Annu. Rev. Biochem. *57*, 159–197.

Guarente, L., and Hoar, E. (1984). Upstream activation sites of the CYC1 gene of Saccharomyces cerevisiae are active when inverted but not when placed downstream of the “TATA box”. Proc. Natl. Acad. Sci. U. S. A. *81*, 7860–7864.

Gutierrez-Hartmann, A., Duval, D.L., and Bradford, A.P. (2007). ETS transcription factors in endocrine systems. Trends Endocrinol. Metab. TEM *18*, 150–158.

He, B., Chen, C., Teng, L., and Tan, K. (2014). Global view of enhancer-promoter interactome in human cells. Proc. Natl. Acad. Sci. U. S. A. *111*, E2191–E2199.

Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y., Ching, C.W., Hawkins, R.D., Barrera, L.O., Van Calcar, S., Qu, C., Ching, K. a, et al. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat. Genet. *39*, 311–318.

Heintzman, N.D., Hon, G.C., Hawkins, R.D., Kheradpour, P., Stark, A., Harp, L.F., Ye, Z., Lee, L.K., Stuart, R.K., Ching, C.W., et al. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature *459*, 108–112.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell *38*, 576–589.

Hoffman, M.M., Buske, O.J., Wang, J., Weng, Z., Bilmes, J.A., and Noble, W.S. (2012). Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nat. Methods *9*, 473–476.

Hoffman, M.M., Ernst, J., Wilder, S.P., Kundaje, A., Harris, R.S., Libbrecht, M., Giardine, B., Ellenbogen, P.M., Bilmes, J.A., Birney, E., et al. (2013). Integrative annotation of chromatin elements from ENCODE data. Nucleic Acids Res. *41*, 827–841.

Hollenhorst, P.C., McIntosh, L.P., and Graves, B.J. (2011). Genomic and biochemical insights into the specificity of ETS transcription factors. Annu. Rev. Biochem. *80*, 437–471.

Hou, C., Li, L., Qin, Z.S., and Corces, V.G. (2012). Gene density, transcription, and insulators contribute to the partition of the Drosophila genome into physical domains. Mol. Cell *48*, 471–484.

Hwang, Y.-C., Zheng, Q., Gregory, B.D., and Wang, L.-S. (2013). High-throughput identification of long-range regulatory elements and their target promoters in the human genome. Nucleic Acids Res. *41*, 4835–4846.

Imakaev, M., Fudenberg, G., McCord, R.P., Naumova, N., Goloborodko, A., Lajoie, B.R., Dekker, J., and Mirny, L.A. (2012). Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat. Methods *9*, 999–1003.

Irizarry, R.A., Bolstad, B.M., Collin, F., Cope, L.M., Hobbs, B., and Speed, T.P. (2003a). Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res. *31*, e15.

Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., and Speed, T.P. (2003b). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostat. Oxford Engl. *4*, 249–264.

Jin, F., Li, Y., Dixon, J.R., Selvaraj, S., Ye, Z., Lee, A.Y., Yen, C.-A., Schmitt, A.D., Espinoza, C. a, and Ren, B. (2013). A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature *503*, 290–294.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A., and Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science *337*, 816–821.

Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. Science *316*, 1497–1502.

Kalhor, R., Tjong, H., Jayathilaka, N., Alber, F., and Chen, L. (2012). Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. Nat. Biotechnol. *30*, 90–98.

Kaplan, N., and Dekker, J. (2013). High-throughput genome scaffolding from in vivo DNA interaction frequency. Nat. Biotechnol. *31*, 1143–1147.

Karolchik, D., Barber, G.P., Casper, J., Clawson, H., Cline, M.S., Diekhans, M., Dreszer, T.R., Fujita, P. a, Guruvadoo, L., Haeussler, M., et al. (2014). The UCSC Genome Browser database: 2014 update. Nucleic Acids Res. *42*, D764–D770.

Kikuta, H., Laplante, M., Navratilova, P., Komisarczuk, A.Z., Engström, P.G., Fredman, D., Akalin, A., Caccamo, M., Sealy, I., Howe, K., et al. (2007). Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. Genome Res. *17*, 545–555.

Knight, P.A., and Ruiz, D. (2012). A fast algorithm for matrix balancing. IMA J. Numer. Anal. *33*, 1029–1047.

Laguna, T., Notario, L., Pippa, R., Fontela, M.G., Vázquez, B.N., Maicas, M., Aguilera-Montilla, N., Corbí, Á.L., Odero, M.D., and Lauzurica, P. (2015). New insights on the transcriptional regulation of CD69 gene through a potent enhancer located in the conserved non-coding sequence 2. Mol. Immunol. *66*, 171–179.

Lajoie, B.R., Dekker, J., and Kaplan, N. (2015). The Hitchhiker's guide to Hi-C analysis: practical guidelines. Methods *72*, 65–75.

Lan, X., Witt, H., Katsumura, K., Ye, Z., Wang, Q., Bresnick, E.H., Farnham, P.J., and Jin, V.X.

(2012). Integration of Hi-C and ChIP-seq data reveals distinct types of chromatin linkages. Nucleic Acids Res. 1–15.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Langer-Safer, P.R., Levine, M., and Ward, D.C. (1982). Immunological method for mapping genes on Drosophila polytene chromosomes. Proc. Natl. Acad. Sci. U. S. A. *79*, 4381–4385.

Le, T.B.K., Imakaev, M. V, Mirny, L.A., and Laub, M.T. (2013). High-resolution mapping of the spatial organization of a bacterial chromosome. Science *342*, 731–734.

Lettice, L.A., Heaney, S.J.H., Purdie, L.A., Li, L., de Beer, P., Oostra, B.A., Goode, D., Elgar, G., Hill, R.E., and de Graaff, E. (2003). A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Hum. Mol. Genet. *12*, 1725–1735.

Levine, M. (2010). Transcriptional Enhancers in Animal Development and Evolution. Curr. Biol. *20*, R754–R763.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

Li, G., Ruan, X., Auerbach, R.K., Sandhu, K.S., Zheng, M., Wang, P., Poh, H.M., Goh, Y., Lim, J., Zhang, J., et al. (2012). Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell *148*, 84–98.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Li, L., Wu, L.P., and Chandrasegaran, S. (1992). Functional domains in Fok I restriction endonuclease. Proc. Natl. Acad. Sci. *89*, 4275–4279.

Li, L., Lyu, X., Hou, C., Takenaka, N., Nguyen, H.Q., Ong, C.-T., Cubeñas-Potts, C., Hu, M., Lei, E.P., Bosco, G., et al. (2015). Widespread rearrangement of 3D chromatin organization underlies polycomb-mediated stress-induced silencing. Mol. Cell *58*, 216–231.

Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science *326*, 289–293.

Lin, C., Valladares, O., Childress, D.M., Klevak, E., Gel-, E.T., Hwang, Y.-C., Tsai, E.A., Schellenberg, G.D., Wang, L., and Geller, E.T. (2013). DRAW+SneakPeek: Analysis Workflow and Quality Metric

Management for DNA-Seq Experiments. Bioinformatics btt422 – .

Lomvardas, S., Barnea, G., Pisapia, D.J., Mendelsohn, M., Kirkland, J., and Axel, R. (2006). Interchromosomal interactions and olfactory receptor choice. Cell *126*, 403–413.

Maksimenko, O., and Georgiev, P. (2014). Mechanisms and proteins involved in long-distance interactions. Front. Genet. *5*, 28.

Mali, P., Yang, L., Esvelt, K.M., Aach, J., Guell, M., DiCarlo, J.E., Norville, J.E., and Church, G.M. (2013). RNA-guided human genome engineering via Cas9. Science *339*, 823–826.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. Nature *461*, 747–753.

Marie-Nelly, H., Marbouty, M., Cournac, A., Liti, G., Fischer, G., Zimmer, C., and Koszul, R. (2014). Filling annotation gaps in yeast genomes using genome-wide contact maps. Bioinformatics *30*, 2105–2113.

Marshall, P., Chartrand, N., and Worton, R.G. (2001). The mouse dystrophin enhancer is regulated by MyoD, E-box-binding factors, and by the serum response factor. J. Biol. Chem. *276*, 20719–20726.

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

Mastrangelo, I.A., Courey, A.J., Wall, J.S., Jackson, S.P., and Hough, P. V (1991). DNA looping and Sp1 multimer links: a mechanism for transcriptional synergism and enhancement. Proc. Natl. Acad. Sci. U. S. A. *88*, 5670–5674.

McKnight, S., and Kingsbury, R. (1982). Transcriptional Control Signals of a Eukaryotic Protein-Coding Gene. Sci. (New York, NY) *217*, 316–324.

McNabb, D.S., Reed, R., and Marciniak, R.A. (2005). Dual luciferase assay system for rapid assessment of gene expression in Saccharomyces cerevisiae. Eukaryot. Cell *4*, 1539–1549.

Mizuguchi, T., Fudenberg, G., Mehta, S., Belton, J.-M., Taneja, N., Folco, H.D., FitzGerald, P., Dekker, J., Mirny, L., Barrowman, J., et al. (2014). Cohesin-dependent globules and heterochromatin shape 3D genome architecture in S. pombe. Nature *516*, 432–435.

Mojica, F.J., Díez-Villaseñor, C., Soria, E., and Juez, G. (2000). Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria. Mol. Microbiol. *36*, 244–246.

Nagano, T., Lubling, Y., Stevens, T.J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E.D., Tanay, A., and Fraser, P. (2013). Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature.

Nath, J., and Johnson, K.L. (2000). A review of fluorescence in situ hybridization (FISH): current status and future prospects. Biotech. Histochem. *75*, 54–78.

Nechanitzky, R., Akbas, D., Scherer, S., György, I., Hoyler, T., Ramamoorthy, S., Diefenbach, A., and Grosschedl, R. (2013). Transcription factor EBF1 is essential for the maintenance of B cell identity and prevention of alternative fates in committed cells. Nat. Immunol. *14*, 867–875.

Nolis, I.K., McKay, D.J., Mantouvalou, E., Lomvardas, S., Merika, M., and Thanos, D. (2009). Transcription factors mediate long-range enhancer-promoter interactions. Proc. Natl. Acad. Sci. U. S. A. *106*, 20222–20227.

Pavletich, N., and Pabo, C. (1991). Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. Science (80-. ). *252*, 809–817.

Phillips, J.E., and Corces, V.G. (2009). CTCF: Master Weaver of the Genome. Cell *137*, 1194–1211.

Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R., and Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. *20*, 110–121.

Pope, B.D., Ryba, T., Dileep, V., Yue, F., Wu, W., Denas, O., Vera, D.L., Wang, Y., Hansen, R.S., Canfield, T.K., et al. (2014). Topologically associating domains are stable units of replication-timing regulation. Nature *515*, 402–405.

Pruitt, K.D., Tatusova, T., and Maglott, D.R. (2005). NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. *33*, D501–D504.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Raney, B.J., Cline, M.S., Rosenbloom, K.R., Dreszer, T.R., Learned, K., Barber, G.P., Meyer, L.R., Sloan, C.A., Malladi, V.S., Roskin, K.M., et al. (2011). ENCODE whole-genome data in the UCSC genome browser (2011 update). Nucleic Acids Res. *39*, D871–D875.

Rao, S.S.P., Huntley, M.H., Durand, N.C., and Stamenova, E.K. (2014). A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell *159*, 1665–1680.

Rödelsperger, C., Guo, G., Kolanczyk, M., Pletschacher, A., Köhler, S., Bauer, S., Schulz, M.H., and Robinson, P.N. (2011). Integrative analysis of genomic, functional and protein interaction data

predicts long-range enhancer-target gene interactions. Nucleic Acids Res. *39*, 2492–2502.

Roh, T., Wei, G., Farrell, C.M., and Zhao, K. (2007). Genome-wide prediction of conserved and nonconserved enhancers by histone acetylation patterns. Genome Res. *17*, 74–81.

Rosenbloom, K.R., Dreszer, T.R., Long, J.C., Malladi, V.S., Sloan, C.A., Raney, B.J., Cline, M.S., Karolchik, D., Barber, G.P., Clawson, H., et al. (2011). ENCODE whole-genome data in the UCSC Genome Browser: update 2012. Nucleic Acids Res. *40*, D912–D917.

Ryba, T., Hiratani, I., Lu, J., Itoh, M., Kulik, M., Zhang, J., Schulz, T.C., Robins, A.J., Dalton, S., and Gilbert, D.M. (2010). Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. Genome Res. *20*, 761–770.

Sabo, P.J., Humbert, R., Hawrylycz, M., Wallace, J.C., Dorschner, M.O., McArthur, M., and Stamatoyannopoulos, J.A. (2004). Genome-wide identification of DNaseI hypersensitive sites using active chromatin sequence libraries. Proc. Natl. Acad. Sci. U. S. A. *101*, 4537–4542.

Sanyal, A., Lajoie, B.R., Jain, G., and Dekker, J. (2012). The long-range interaction landscape of gene promoters. Nature *489*, 109–113.

Schoenfelder, S., Clay, I., and Fraser, P. (2010). The transcriptional interactome: gene expression in 3D. Curr. Opin. Genet. Dev. *20*, 127–133.

Schug, J., Schuller, W.-P., Kappen, C., Salbaum, J.M., Bucan, M., and Stoeckert, C.J. (2005). Promoter features related to tissue specificity as measured by Shannon entropy. Genome Biol. *6*, R33.

Selvaraj, S., R Dixon, J., Bansal, V., and Ren, B. (2013). Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. Nat. Biotechnol. *31*, 1111–1118.

Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A., and Cavalli, G. (2012). Three-dimensional folding and functional organization principles of the Drosophila genome. Cell *148*, 458–472.

Shannon, C. (1948). A Mathematical Theory of Communication. Bell Syst. Tech. J. *27*, 379–423.

Shen, Y., Yue, F., McCleary, D.F., Ye, Z., Edsall, L., Kuan, S., Wagner, U., Dixon, J., Lee, L., Lobanenkov, V. V, et al. (2012). A map of the cis-regulatory sequences in the mouse genome. Nature *488*, 116–120.

Shibata, Y., Sheffield, N.C., Fedrigo, O., Babbitt, C.C., Wortham, M., Tewari, A.K., London, D., Song, L., Lee, B.-K., Iyer, V.R., et al. (2012). Extensive evolutionary changes in regulatory element activity during human origins are associated with altered gene expression and positive selection. PLoS Genet. *8*, e1002789.

Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., van Steensel, B., and de Laat, W. (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). Nat. Genet. *38*, 1348–1354.

Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). BioGRID: a general repository for interaction datasets. Nucleic Acids Res. *34*, D535–D539.

Stranger, B.E., Nica, A.C., Forrest, M.S., Dimas, A., Bird, C.P., Beazley, C., Ingle, C.E., Dunning, M., Flicek, P., Koller, D., et al. (2007). Population genomics of human gene expression. Nat. Genet. *39*, 1217–1224.

Struhl, K. (1984). Genetic properties and chromatin structure of the yeast gal regulatory element: an enhancer-like sequence. Proc. Natl. Acad. Sci. U. S. A. *81*, 7865–7869.

Tanizawa, H., Iwasaki, O., Tanaka, A., Capizzi, J.R., Wickramasinghe, P., Lee, M., Fu, Z., and Noma, K.-I. (2010). Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation. Nucleic Acids Res. *38*, 8164–8177.

Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B., et al. (2012). The accessible chromatin landscape of the human genome. Nature *489*, 75–82.

Varoquaux, N., Liachko, I., Ay, F., Burton, J.N., Shendure, J., Dunham, M.J., Vert, J.-P., and Noble, W.S. (2015). Accurate identification of centromere locations in yeast genomes using Hi-C. Nucleic Acids Res. *43*, 5331–5339.

Visel, A., Blow, M.J., Li, Z., Zhang, T., Akiyama, J. a, Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., et al. (2009). ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature *457*, 854–858.

Wang, C., Liu, C., Roqueiro, D., Grimm, D., Schwab, R., Becker, C., Lanz, C., and Weigel, D. (2015). Genome-wide analysis of local chromatin packing in Arabidopsis thaliana. Genome Res. *25*, 246–256.

Wang, J., Zhuang, J., Iyer, S., Lin, X.-Y., Greven, M.C., Kim, B.-H., Moore, J., Pierce, B.G., Dong, X., Virgil, D., et al. (2013). Factorbook.org: a Wiki-based database for transcription factor-binding data generated by the ENCODE consortium. Nucleic Acids Res. *41*, D171–D176.

van de Werken, H.J.G., Landan, G., Holwerda, S.J.B., Hoichman, M., Klous, P., Chachik, R., Splinter, E., Valdes-Quezada, C., Oz, Y., Bouwman, B.A.M., et al. (2012). Robust 4C-seq data analysis to screen for regulatory DNA interactions. Nat. Methods *9*, 969–972.

Whalen, S., Truty, R.M., and Pollard, K.S. (2015). Protein binding and methylation on looping chromatin accurately predict distal regulatory interactions (Cold Spring Harbor Labs Journals).

Wilber, A., Nienhuis, A.W., and Persons, D.A. (2011). Transcriptional regulation of fetal to adult hemoglobin switching: new therapeutic opportunities. Blood *117*, 3945–3953.

Williamson, I., Berlivet, S., Eskeland, R., Boyle, S., Illingworth, R.S., Paquette, D., Dostie, J., and Bickmore, W. a (2014). Spatial genome organization: contrasting views from chromosome conformation capture and fluorescence in situ hybridization. Genes Dev. *28*, 2778–2791.

Wu, C., Wong, Y.C., and Elgin, S.C. (1979). The chromatin structure of specific genes: II. Disruption of chromatin structure during gene activity. Cell *16*, 807–814.

Yaffe, E., and Tanay, A. (2011). Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nat. Genet. *43*, 1059–1065.

Young, M.D., Willson, T.A., Wakefield, M.J., Trounson, E., Hilton, D.J., Blewitt, M.E., Oshlack, A., and Majewski, I.J. (2011). ChIP-seq analysis reveals distinct H3K27me3 profiles that correlate with transcriptional activity. Nucleic Acids Res. *39*, 1–13.

Zhang, Y., Wong, C.-H., Birnbaum, R.Y., Li, G., Favaro, R., Ngan, C.Y., Lim, J., Tai, E., Poh, H.M., Wong, E., et al. (2013). Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations. Nature *504*, 306–310.

Zhao, Z., Tavoosidana, G., Sjölinder, M., Göndör, A., Mariano, P., Wang, S., Kanduri, C., Lezcano, M., Sandhu, K.S., Singh, U., et al. (2006). Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. Nat. Genet. *38*, 1341–1347.

Zheng, Q., Ryvkin, P., Li, F., Dragomir, I., Valladares, O., Yang, J., Cao, K., Wang, L.-S., and Gregory, B.D. (2010). Genome-Wide Double-Stranded RNA Sequencing Reveals the Functional Significance of Base-Paired RNAs in Arabidopsis. PLoS Genet. *6*, e1001141.