2022

Computational Methods to Analyze Next-generation Sequencing Data in Genomics and Metagenomics

Saidi Wang
University of Central Florida

COMPUTATIONAL METHODS TO ANALYZE NEXT-GENERATION SEQUENCING DATA IN GENOMICS AND METAGENOMICS

by

SAIDI WANG
B.S. Tianjin Polytechnic University, 2014
M.S. Tongji University, 2017

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in the Department of Computer Science
in the College of Engineering and Computer Science
at the University of Central Florida
Orlando, Florida

Summer Term
2022

Major Professor: Haiyan Hu

ABSTRACT

This thesis focuses on two important computational problems in genomics and metagenomics with the public available next-generation sequencing data. One is about gene regulation, for which we explore how distal regulatory elements may interact with the proximal regulatory elements. The other is about metagenomics, in which we study how to reconstruct bacterial strain genomes from shotgun reads. Studying gene regulation, especially distal gene regulation, is important because regulatory elements, including those in distal regulatory regions, orchestrate when, where and how much a gene is activated under every experimental condition. Their dysfunction results in various types of diseases. Moreover, the current study on distal gene regulation is still under development. The study of bacterial strains is also vital, as the bacterial strains are the main source of drug resistance, mixed infection, reinfection, etc. The study of novel bacterial strains is still in its infancy, with only one tool that can work with multiple metagenomic samples while has suboptimal performance. We identified hundreds of pairs of regulatory elements that are biologically sound and are likely to contribute to the interaction of distal and proximal regulatory regions. We demonstrated for the first time that ribosomal protein genes share common distal regulatory regions under the same experimental conditions and might be differentially regulated across different experimental conditions. In addition, we developed a novel approach called SMS to reconstruct novel bacterial strains from multiple shotgun metagenomic samples. Tested on 702 simulated and 195 experimental datasets, we showed that SMS has high accuracy in inferring the present strains, including the strain number, strain abundance, strain variations, etc. Compared with the two existing approaches, SMS shows much better performance. Our studies shed new light on genomics and generated novel tools in metagenomics.

ACKNOWLEDGMENTS

I want to thank my advisor, Dr. Haiyan Hu, for her patience and guidance throughout all these years. She taught me a lot not only in research but also in life. Without her help and encouragement, I would not finish this long journey.

I would also like to thank my co-advisor, Dr. Xiaoman Li, for his earnest and patient guidance in my research. I still remember how he helps me improve my English and professional knowledge. He is so kind and professional. Thanks again for your help.

I want to thank my other committee members: Dr. Cliff Zou and Dr. Gang Chen, for their time and guidance. They gave me professional advice and helped my Ph.D. study.

Finally, I would like to thank my parents for their continuous support. My parents are ordinary Chinese farmers. They worked hard to support me in studying in big cities and coming to United States for my Ph.D. study. During this journey, they pay far more than they can afford. I do not know how I will pay them back and I hope that they can be healthy and enjoy their life. Thanks again to my great father and mother.

This work is supported by the National Science Foundation [grands 1661414, 1356524, 1149955, and 2015838] and the National Institute of Health [grant R15GM12340].

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

1.1 Next-generation sequencing

Next-generation sequencing (NGS) [1] refers to the non-Sanger-sequencing-based high-throughput DNA sequencing technologies that have been commercially available since 2005. The entire human genome can be sequenced in one day with NGS technologies. By contrast, it took Sanger sequencing more than a decade to generate the first human draft genome in 2001 [1, 2]. The main difference between Sanger sequencing and NGS includes the sequencing capability, the sequenced fragment length, etc. The sequenced DNA fragments are called reads. While the Sanger method only sequences 96 or 392 reads of 1000 base pairs (bps) long or os in one instrumental run, NGS is massively parallel and can sequence millions of much shorter reads [3]. For instance, the NGS reads from Illumina in the early time are about 36 bps long and are about 150 bps long currently.

NGS have revolutionized genomics, epigenomics and metagenomics research [4]. Various forms of NGS-based platforms have been developed, including chromatin immunoprecipitation followed by massive parallel sequencing (ChIP-Seq) [5, 6], RNA-sequencing (RNA-seq) [7], high-throughput chromosome conformation capture (Hi-C) [8], etc. ChIP-seq can identify transcription factor binding sites (TFBSs) of regulatory proteins called transcription factors (TFs) on the genome scale. It can also be applied to identify genome-wide histone modification patterns and other regulatory proteins. RNA-seq is widely used to profile gene expression under different experimental conditions, with more accurate measurements than the previously used microarray-based technologies [9]. Hi-C is used to detect the physical closeness of two genomic regions in 3D, which indicates the potential physical interactions of two distant genomic regions [10].

In addition to the above most widely used NGS platforms, there are other types of NGS experiments such as PAR-CLIP (Photoactivatable-Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation) [11], CLASH (Cross-linking Ligation And Sequencing of Hybrids) [12], CAGE (Cap-Analysis Gene Expression) [13], DNase-seq (DNase I hypersensitive sites sequencing), ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing), etc. PAR-CLIP and CLASH are used to study the microRNA target sites [11, 12, 14-17]. CAGE is used to measure gene expression levels and transcription start sites of different transcripts, which is originally based on microarray while relies on NGS technologies since the last decade [13, 18, 19]. DNase-seq and ATAC-seq determine the genome-wide open chromatin regions, which describe all potential active elements, including active genes and regulatory elements under a specific experimental condition. These high-throughput approaches have generated a plethora of data and have revolutionzed our understanding of genomics, epigenomics, metagenomics, etc. [20, 21].

1.2 Gene regulation

Gene regulation is a cellular mechanism related to gene expression that controls the types of gene products and the amount of each gene product synthesized under a given experimental context [22, 23]. It begins with open chromatin and transcription initiation, followed by transcription, post-transcirption, translation, post-translational modifications, etc. [24]. At every step, gene expression is regulated precisely. For instance, chromatin structure is a key factor in gene regulation, where euchromatin and heterochromatin can interconvert through DNA methylation to modify histones and determines when and where the genome is active and ready for transcription initialization.

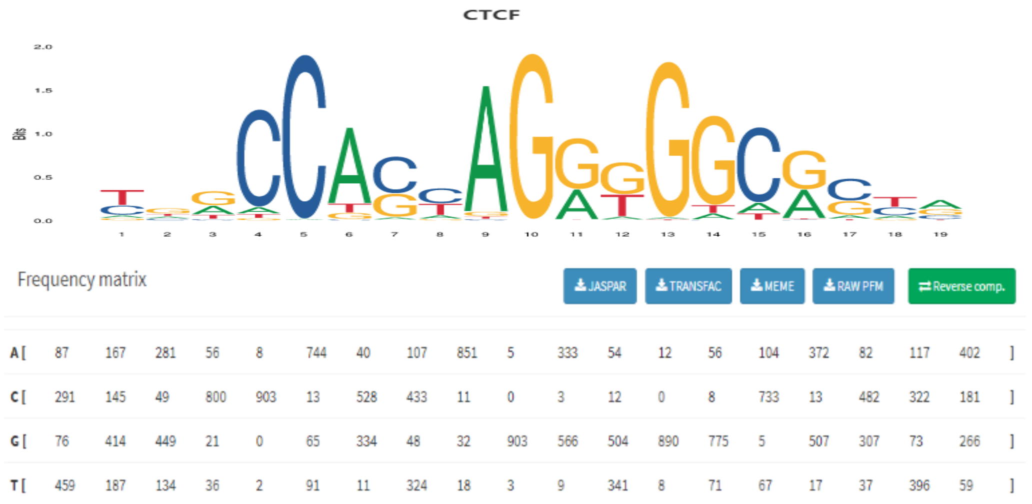
The regulation of gene expression has important implications for controlling developmental processes, responses to environmental stress, adaptation to new environmental conditions, etc. Almost every cell in an organism has the same set of genes in its DNA. Some genes in the genome are always expressed because their function is fundamental to the organism. On the contrary, other genes may only be expressed in specific cells, tissues, or organs. At the same time, the amount of expression for each gene is precisely controlled in individual cell types under specific experimental conditions.

Gene regulation modulates gene activities, which involve proximal regulatory elements in promoters and distal regulatory elements in enhancers [25-27]. Together with the regulatory proteins, the proximal and distal regulatory elements work together to turn on/off a gene in a cell. RNA polymerase is the enzyme responsible for transcription that polymerizes complementary nucleotides to synthesize mRNA molecules. TFs bind to their binding sites in promoters and enhancers to recruit RNA polymerase, which binds to specific DNA sequences called response elements in promoters. A promoter contains the basic regulatory elements of a gene, which explains the basal expression level of a gene. It is located near the gene, in upstream of the codon sequence. The size of the promoter can be 100-1000 bps. Enhancers are cis-acting elements involved in increasing the activity of specific promoters. There are short DNA sequences of about 50-1500 bps to which TFs called activators. Enhancers can be located up to several million bps from the promoters. They can also be in the upstream or downstream of the promoters. In spite of their large distances from the promoters, enhancers are spatially close to promoters, allowing interactions with RNA polymerases and basal TFs in promoters. Activators bound to enhancer regions subsequently bind to mediator complexes, which in turn recruit RNA polymerase and basal TFs to promoters. The orientation of the enhancer sequences does not affect their functionality in strengthening gene expression levels.

TFs play essential roles in gene regulation [6, 28, 29]. TFs are proteins that can bind to specific nucleotide sequences upstream of a gene and regulate the transcription of this gene. A motif is a pattern of the DNA segments bound by a TF, commonly represented by a position weight matrix or a consensus sequence (Figure 1-1). In high eukaryotes, multiple TFs often bind to their TFBSs within a short DNA region to work together to modulate the gene expression of their target genes. Such short regions with TFBSs of different TFs are called cis-regulatory modules. It is said that there are at least five to ten times more cis-regulatory modules than genes, which determine the temporal and spatial expression pattern of their target genes in high eukaryotes [30, 31].

In order to understand gene transcriptional regulation, it is important to identify and study enhancer-promoter (EP) interactions (EPIs) [10]. As described above, promoters are the upstream 1000 bps regions of gene transcription start sites, and enhancers are short genomic regions that can strengthen their target genes' transcriptional levels independent of their distance and orientation to the target genes [32]. Enhancers interact with promoters of their target genes, which will increase target genes' transcription and modulate their condition-specific expression [32-34]. In this thesis, we will study the EPIs and distal regulatory regions (also called enhancers) of ribosomal protein genes (RPGs), which will help us to better understand gene transcriptional regulation especially distal gene transcriptional regulation.

As pairs of interacting TFs have been shown to contribute to EPIs by binding to enhancers and promoters, it is important to investigate which TF pairs may contribute to EPIs. Since a motif is the pattern of the DNA segments bound by a TF, it is meaningful to study the motif pairs that contribute to EPIs. Until now, we still lack a clear view of how TF pairs interact, how the interaction of TF pairs will contribute to EPIs, which TF pairs impact EPIs, etc. It is thus important to identify TF-binding motif pairs in high-quality EPIs.



A [	87	167	281	56	8	744	40	107	851	5	333	54	12	56	104	372	82	117	402	]
C [	291	145	49	800	903	13	528	433	11	0	3	12	0	8	733	13	482	322	181	]
G [	76	414	449	21	0	65	334	48	32	903	566	504	890	775	5	507	307	73	266	]
T [	459	187	134	36	2	91	11	324	18	3	9	341	8	71	67	17	37	396	59	]

Figure 1-1: The CTCF motif from JASPAR.

Many resources are available to study gene regulation. TRANSFAC [28] provides data on eukaryotic TFs, their experimentally-determined TFBSs, consensus binding sequences, positional weight matrices, and regulated genes. JASPAR [29] is an open-access database containing manually curated, non-redundant TF motifs for TFs across six taxonomic groups. UniPROBE [35] database hosts data generated by universal protein binding microarray (PBM) technology on the in vitro DNA-binding specificities of TFs. MEME SUITE [36] provides a unified portal for online discovery and analysis of sequence motifs representing features such as DNA binding sites and protein interaction domains.

1.3 RPG (ribosomal proteins genes)

Human RPGs are house-keeping genes that code for the structural proteins in the ribosome, the machine that makes proteins in every organism. RPGs are well known for their coordinated expression, meaning that in a given species, their mRNA expression levels are highly correlated across various experimental conditions [37]. In order to understand the molecular basis of the RPG coordinated gene expression [38, 39], it will be crucial to study RPG transcriptional regulation and address the following two issues: 1) How RPG coordinated expression is controlled; 2) how distal regulatory elements may contribute to the coordinated gene expression of RPGs. Rarely is a study that explores the distal regulatory regions of RPGs. Studying the RPG distal regulatory regions will thus shed new light on our understanding of RPG coordinated transcriptional regulation.

Many studies have been carried out to understand how RPGs are coordinately regulated. Early experimental studies showed that several RPGs share TFBSs of a common TF and validated the regulatory roles of these TFBSs [40, 41]. Later, high-throughput experiments showed that TFs such as RAP1 and FHL1 bind promoters of almost all RPGs in yeast [42, 43]. With the genomes of human and other organisms available, computational studies became popular and demonstrated that there are TFBSs of the same TFs in promoters of almost all RPGs in a species [37, 44-46].

All the above studies focused on RPG promoter regions. Rarely is a study that explores the distal regulatory regions of RPGs. To fill this gap, Li et al. previously studied the putative regulatory regions within one megabase (Mbps) of the 80 human RPGs with the DNase I hypersensitive sites (DHSs) in 349 samples [33]. For the sake of simplicity, henceforth, we use a "sample" to refer to a cell line, a cell type, or a tissue under an experimental condition. They identified 217 putative regulatory regions of RPGs that are shared by the majority of the

349 samples. More than 86% of these shared regulatory regions were supported by the chromatin interaction data.

Although this previous study shed new light on human RPG transcriptional regulation, it is limited in the following aspects [33]. First, not all identified regions interacted with RPG promoters and thus they may not be RPG regulatory regions. Second, the previously identified regions are shared across the majority (>=85%) of the 349 samples and are limited in terms of studying sample-specific regulation of human RPGs. Third, these regulatory regions are limited to 1 Mbps neighborhood of RPGs, while regulatory regions may be more distal than 1 Mbps [47].

To understand human RPG distal regulation better, in this thesis, we defined sample-specific putative RPG regulatory regions directly from high-throughput chromatin interaction data in eleven samples [10, 48]. We identified about 22797 putative RPG regulatory regions, the majority of which were distal regions. More than 44% of these regions were only identified in one sample, implying that RPGs were likely differentially regulated in different samples. Interestingly, 2 to 77 RPGs shared a common regulatory region in a sample, and the same pairs of RPGs shared common regulatory regions across samples, which may partially explain their coordinated gene expression. By studying the overrepresented TF binding motifs in these regions in a sample, we identified common TF binding motifs shared by samples. Our study shed new light on the distal regulation of the human RPGs.

1.4 Metagenomics

Metagenomics is also called microbial environmental genomics. It directly extracts the DNA of all microorganisms in a sample, constructs a metagenomic library, and uses the research strategy of genomics to study the genetic composition and community functions of all microorganisms in the given sample. It is a strategy to study microbial diversity and discover new genes. Its main steps include Cloning the total DNA (also called metagenome) of all microorganisms in a specific environment and obtaining new physiologically active substances by means of building a metagenomic library and screening; or designing primers according to the rDNA database through Phylogenetic analysis obtained genetic diversity and molecular ecology information of microorganisms in this environment.

16S ribosomal RNA sequencing [49] and shotgun sequencing [50] are used in metagenomics to study both culturable and unculturable bacteria and archaea [51] with NGS technologies. 16S ribosomal RNA sequencing use PCR to target and amplify portions of the 16S rRNA gene. Then PCR amplicons from an individual sample could be pooled together and sequenced. Shotgun metagenomic sequencing is different from 16S ribosomal RNA sequencing, which only targets the 16S rRNA gene [52]. In shotgun metagenomic sequencing, microbial genomes of all microbial organisms in the sample are randomly fragmented. The generated fragments are then sequenced. The sequenced fragments called reads are then employed to infer the present species and strains and their abundance in this sample. Therefore, shotgun metagenomic sequencing is more unbiased than the 16S ribosomal RNA sequencing and is more widely applied in the current metagenomic studies [53-55].

Hundreds of computational methods are developed for high-level taxon analysis, leaving about a few dozen computational methods available to infer bacterial strains from shotgun metagenomic reads. Most of these computational tools for bacterial strain analyses are known-strain-based. For example, Pathoscope 2.0 [56] is based on the expectation-maximization algorithm to reassign reads to known species/strains. Sigma [57] is based on a known-strain reference genome and a user-defined database with the maximum likelihood estimation to predict strains. StrainSeeker [58] needs a known-strain reference genome, a k-mer database, and a guide tree to identify k-mers at nodes. Although this type of analysis is valuable to understanding bacterial strains, they are primarily used to discover known strains, not to reconstruct novel ones. In practice, many mutations are likely to accumulate in bacterial strains. Therefore, genomes of novel strains rather than known strains are expected to appear in the samples. In other words, known strain information may be limited in practice [59].

Dozens of computational methods are available to infer novel microbial strains from shotgun metagenomic reads. MixtureS [60] and StrainFinder [61] showed better performance for novel strain identifications previously [60]. MixtureS predicts novel bacterial strains in one sample based on binomial testing and has good performance. StrainFinder uses the multinomial distribution to predict novel strains in multiple samples. However, it cannot even reliably predict the strain number. MIDAS [62] analyzes novel bacterial strains based on marker genes with a database of reference genomes and can predict one strain at a time.

In summary, existing methods and tools often depend on known strains, barely can work on multiple samples, and are not reliable or not easy to use. It will be necessary to develop a more user-friendly tool to identify strains more accurately. In this thesis, we thus developed a novel approach called SMS to reconstruct bacterial strains from multiple shotgun metagenomic samples to address the above limitations.

CHAPTER 2: MOTIF PAIRS IN ENHANCER-PROMOTER INTERACTIONS

Previously published as Wang S, Hu H, Li X. A systematic study of motif pairs that may facilitate enhancer-promoter interactions. J Integr Bioinform. 2022 Feb 7;19(1).

2.1 A systematic study of motif pairs that may facilitate enhancer-promoter interactions

2.1.1 Introduction

Identifying enhancer-promoter (EP) interactions is important for the understanding of gene transcriptional regulation [10]. Enhancers are short genomic regions that can strengthen their target genes' transcriptional levels independent of their distance and orientation to the target genes [32]. They are in general several hundred bps long, can be hundreds to thousands of bps away from their target genes, and can be in the upstream or downstream of the target genes or in introns. By interacting with promoters of their target genes, enhancers increase target genes' transcription and modulate their condition-specific expression [32-34].

There are many studies that have attempted to identify EP interactions. Experimental approaches based on chromatin conformation capture techniques and their extensions have identified many EP interactions across several cell lines, cell types and tissues [10, 48, 63-68]. These experimental approaches nurtured our rudimentary understanding of EP interactions. However, they are either time-consuming or still costly because of the large number of EP interactions under an experimental condition and the required high-sequencing depth to comprehensively identify them on the genome-scale [10, 69]. Computational methods for EP interaction predictions are thus indispensable. These methods usually consider the distance, conservation, correlated activity between enhancers and promoters, etc., to identify EP interactions [70-79]. Although having shown success, they have a suboptimal performance on discovering EP interactions, especially condition-specific EP interactions [47, 74, 80-82]. It is thus necessary to further investigate the characteristics of EP interactions, which may significantly facilitate the improvement of the accuracy of the existing methods.

There are several studies that pointed out a new venue to explore the characteristics of EP interactions, which suggested that the interaction of transcription factors (TFs) that bind an enhancer and TFs that bind a promoter of an EP pair may contribute to the interaction of this EP pair [32, 69, 77, 83-87]. For instance, it is well known that the TF and structural protein CTCF binds to a fraction of enhancers and promoters, which facilitates the physical interaction of enhancers and promoters in these EP pairs [88]. Another example, the ubiquitous TF YY1, binds to enhancers and promoters and contributes to EP interactions as well [89]. It is thus promising to systematically study the potential interactions of TFs that bind to enhancers and promoters and understand how such interactions may lead to the interaction of EP pairs. A computational study integrated chromatin immunoprecipitation followed by massive parallel sequencing (ChIP-seq) data and Hi-C data in two cell lines and predicted 565 interactions of DNA-binding proteins, including TFs [69]. This study was encouraging while limited with a small number of TFs in only two cell lines. To date, we still lack a clear view of the interaction of which TF pairs may render the specificity of the interaction of the enhancer and the promoter in an EP pair.

In order to address these problems, we systematically investigated the co-occurrence of potential TF binding motifs in enhancers and their corresponding interacting promoters (Material and Methods). A motif is a TF binding pattern, which is often represented by a position weight matrix [90, 91]. We identified 114 non-redundant motifs in interacting EP pairs that represented the binding patterns of potential TFs. We also identified 423 motif pairs that significantly co-occurred in interacting EP pairs. Interestingly, on average, more than 62% of these motif pairs in a cell line were shared across cell lines and were able to help to distinguish true interacting EP pairs from false ones. Our study provides a comprehensive list of motif pairs that may contribute to EP physical interactions and facilitate their predictions, which also creates meaningful hypotheses for experimental validation of EP interactions.

2.1.2 Material and Methods

2.1.2.1 Positive and negative EP pairs

The Hi-C contact matrices are downloaded in the following seven cell lines: GM12878, HMEC, HUVEC, IMR90, K562, KBM7 and NHEK, which were normalized with the Knight and Ruiz normalization vectors by Rao et al. [10]. Two genomic regions are interacted in a cell line (except GM12878) if the corresponding entry in the normalized contact matrix of this cell line was larger than 30. The interacting regions defined by this cutoff would include almost all pairs of interacting regions defined in IMR90 and K562 by independent studies [63, 64]. Because the Hi-C sequencing depth in GM12878 was one magnitude larger than that in all other cell lines (Table 2-1), to control false positives, a larger cutoff 150 is used in in GM12878. This larger cutoff resulted in a similar number of selected pairs of interacting regions in GM212878 [10]. In this way, positive pairs of interacting regions are got. Note that we could use the looplists defined by Rao et al. as positive pairs of interacting regions [10]. However, the number of looplists was small, which resulted in an even smaller number of positive EP pairs that could not be used to discover interacting TF pairs below.

Table 2-1: Basic information about the data used in the paper.

cell line	sequencing depth (million)	#enhancers	#promoters	#enhancer length	#promoter length	#positive EP pairs	#3rd type negative EP pairs
GM12878	15112.0	2731	2171	372	1100	3688	28458
HMEC	1068.0	1761	1713	370	1100	2157	10719
HUVEC	892.8	751	650	382	1100	835	4966
IMR90	1683.1	2344	2137	381	1100	3226	8859
K562	1366.2	2096	1942	367	1100	2972	9666
KBM7	1247.9	6278	5970	320	1100	7862	56787
NHEK	1347.5	1160	1018	372	1100	1313	5022

In order to obtain positive EP pairs in a cell line, the above positive pairs of genomic regions are overlapped with the corresponding “active” enhancers and “active” promoters (Figure 2-1A). An active enhancer was one of the 32284 enhancers defined by FANTOM [92] that overlapped with the H3K27ac ChIP-seq peaks [93] in the corresponding cell line. To our knowledge, FANTOM enhancers were the largest collection of mammalian enhancers with direct experimental evidence. With the transcription start sites (TSSs) defined in GENCODE, we defined 57820 promoters, each of which was the genomic region from the upstream 1000 bps to the downstream of 100 bps the TSS of a GENCODE gene. An active promoter was then defined with these GENCODE promoters and the ENCODE RNA-seq data as previously [47, 76]. In this way, every positive EP pair had its enhancer overlapping with one genomic region and its promoter overlapping with the other genomic region of a positive pair of genomic regions, and the distance between the active enhancer and the active promoter was within 2.5 kilobase pairs to 2 megabase pairs. The majority of the positive EP pairs were likely to be true positives, despite false positives and false negatives.

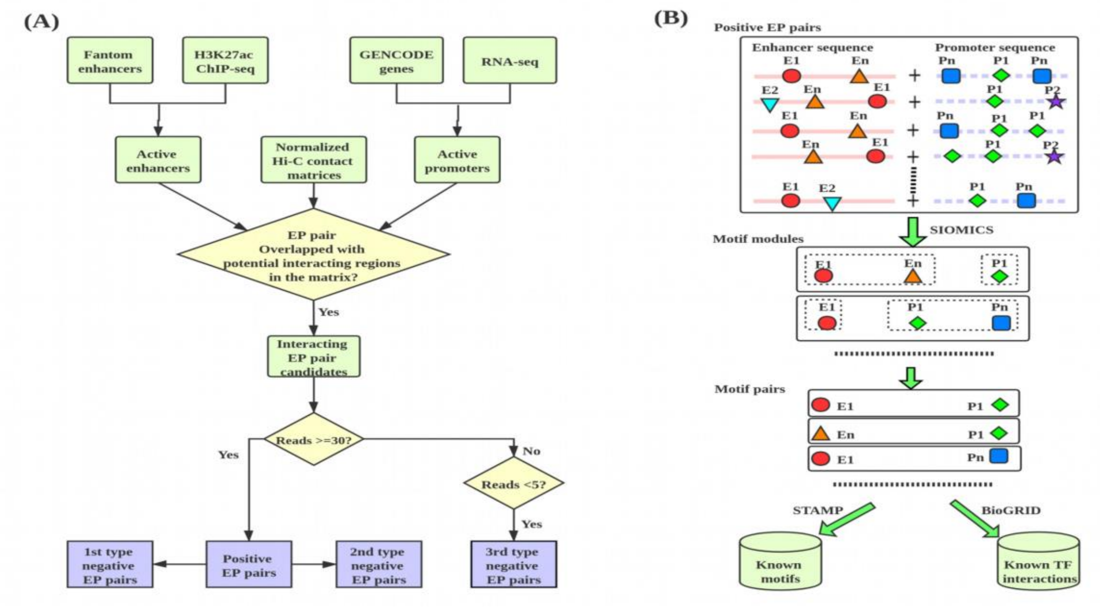


Figure 2-1: EP pairs and motif pairs (A). The procedure to obtain positive and negative EP pairs. (B). The pipeline to study motif pairs in positive EP pairs.

To assess how well the predicted motif pairs facilitate the identification of true interacting EP pairs, we generated three types of negative EP pairs (Figure 2-1A). The first type was the permuted version of the positive ones, in which the enhancer and the promoter sequence of a negative EP pair was a random permutation of the enhancer and the promoter sequence in the corresponding positive EP pair, respectively. The second type of negative EP pairs was generated by replacing the enhancers in positive EP pairs with randomly chosen genomic regions. These random genomic regions had a similar length distribution and a similar distance distribution to promoters as the enhancers in positive EP pairs. The third type was defined from the normalized Hi-C contact matrices with the cutoff 5, similar to the positive EP pairs (Figure 2-1A). In brief, if a pair of genomic regions had fewer than 5 supported normalized Hi-C reads, we called this pair of regions a negative pair of genomic regions. We then overlapped the negative pairs of genomic regions with the active FANTOM enhancers and active GENCODE promoters to obtain negative EP pairs. The first two types of negative EP pairs were used to assess whether the predicted motif pairs could distinguish non-EP pairs from positive EP pairs, while the third type was used to determine whether they could separate the interacting pairs from non-interacting pairs.

2.1.2.2 Non-redundant known motifs

We collected known TF binding motifs from JASPAR and CIS-BP databases [90, 94]. We compared every pair of motifs from these two sources with the tool STAMP [95]. As previously [96, 97], if two motifs had a STAMP similarity E-value smaller than 1E-05, we claimed they were similar. We obtained 649 non-redundant known motifs from the two sources by keeping only one motif in each group of similar motifs.

2.1.2.3 Discovery of motif pairs

To study motif pairs that may facilitate EP interactions, we obtained the DNA sequence of the enhancer and promoter in every positive EP pair in each cell line. We then concatenated an enhancer sequence with its corresponding promoter sequence, if this enhancer and this promoter formed a positive EP pair (Figure 2-1B). The obtained sequences were repeat masked by repeatmasker with the default parameters (https://www.repeatmasker.org/) so that patterns due to the overrepresentation of repeats in input sequences would not be identified as TF binding motifs.

Next, we applied the tool SIOMICS [96, 98] to these repeat-free concatenated sequences in every cell line to identify motif modules. A motif module is a group of motifs whose binding sites significantly co-occur in input sequences. Biologically, a motif module mimics the group of motifs for one TF and its cofactor TFs, where this TF and its cofactor TFs bind to sequences to regulate a common group of genes. SIOMICS considers multiple co-occurring sequence patterns to identify motif modules and motifs, which significantly reduces false positive predictions compared with the strategy to predict individual TF motifs separately [91]. Moreover, it can de novo predict motifs, which thus does not depend on the limited number of known motifs available. We ran SIOMICS on the repeat-masked sequences in each cell line separately, with the default parameters except s = 30 and n = 1500, which meant that we intended to identify up to 1500 motifs in a cell line and all motifs in a motif module must co-occur in at least 30 input sequences. The choice of 1500 is because there are about 1500 sequencing-specific binding TFs in the human genome [99]. SIOMICS assesses the statistical significance of the co-occurrence of every group of motifs by the Poisson clumping heuristic [100] and outputs the significant motif groups as motif modules (corrected p-value<0.01).

We then compared the predicted motifs in motif modules with the above non-redundant known motifs. A predicted motif was similar to a known motif if the STAMP E-value smaller than 1E-5. The TF(s) corresponding to this known motif was considered to be the TF(s) bound to this predicted motif. We also compared motifs predicted in different cell lines and claimed that two predicted motifs were the same if their STAMP E-value was smaller than 1E-8. This more stringent cutoff 1E-8 was used here because the same motifs predicted by the same tool should be more similar to each other than the motifs from different tools/sources [101, 102].

Alternatively, we studied motif modules with known TF motifs. With the above 649 known motifs, we scanned the same EP sequences with FIMO [103] in every cell line separately to obtain initial putative binding sites of known motifs. We then studied the co-occurrence of known motifs with these binding sites by ChIPModule [104]. ChIPModule is similar to SIOMICS except that it considers the co-occurrence of the binding sites of known motifs to predict motif modules, which minimizes the false positive predictions in these binding sites defined by FIMO [104].

Finally, with the predicted motif modules, we obtained all pairs of motifs in every motif module. We then kept the pairs with one motif occurring in promoters and the other motif occurring in enhancers of positive EP pairs. The occurrence of a motif in enhancers and promoters was defined by SIOMICS. In other words, we filtered pairs that co-occurred in only enhancers or only promoters. These remaining pairs were the final motif pairs we considered, the TFs of which may be likely to interact and contribute to the interaction of the EP pairs (Figure 2-1B).

2.1.2.4 Homogeneous motif pairs

We consider a motif pair composed of the same motif as a homogeneous motif pair if this motif significantly co-occurs in both enhancers and promoters of positive EP pairs. We apply two approaches to measure the significance of such a co-occurrence of the same motif to identify homogeneous motif pairs. In one way, assume there are *N* positive EP pairs, and *n* has such a motif in both enhancers and promoters (based on FIMO scan). Assume the average promoter and enhancer length is l_1 and l_2 in this cell line, respectively. Also, assume this motif occurs *x* times in these *N* EP pairs. We calculate the p-value as *pbinom(n, N ,p),* where *pbinom(n, N ,p),* where $p = \frac{x}{N*(l_1+l_2)}$, $pbinom(x,n,p) = \sum_{i=x}^{n} \frac{n!}{i!(n-i)!} p^i (1-p)^{n-i}$. If this p-value is smaller than $0.01/K$, where *K* is the number of the predicted motifs in this cell line, we claim that this motif forms a homogeneous motif pair. In the other way, assume this motif occurs in *x* of the *N* enhancers and *y* of the *N* promoters based on the FIMO scan. We calculate the p-value with the same formula but different $p = \frac{x*y}{N*N}$. If this p-value is smaller than $0.01/K$, we claim this motif is significant.

2.1.2.5 Enrichment analysis of the predicted EP motif pairs

We compared the predicted EP motif pairs with known motif pairs of interacting TFs. We collected directly and indirectly interacting TF pairs from BioGRID [105]. The direct TF interactions meant that two TFs physically interacted with each other. The indirect ones referred to pairs of TFs without direct interaction but directly interacting with a common third protein. There were 6820 pairs of direct and 120,277 pairs of indirect known TF interactions in BioGRID, which involved 1520 and 1207 TFs, respectively. We then assessed the statistical significance of the enrichment of the motif pairs of known interacting TFs in the predicted motif pairs in every cell line by the hypergeometric testing. In brief, assume there were N TFs and M pairs of TFs in BioGrid, among which there were *m* pairs that involved *n* TFs in the predicted EP motif pairs in a cell line. We calculated the p-value of enrichment of motif pairs of known interacting TFs as $phyper\left(m, \frac{n(n-1)}{2}, M, \frac{N(N-1)}{2}\right)$, where $phyper(x_1, y_1, x_2, y_2) = \sum_{k=x_1}^{min\,(y_1,x_2)} \frac{y_1!(y_2-y_1)!x_2!(y_2-x_2)!}{y_2!k!(x_2-k)!(y_1-k)!(y_2-x_2-y_1+k)!}$. We also compared the predicted EP motif pairs with those predicted in a previous study, which predicted 298 pairs of TF interactions involved 61 TFs in GM12878 and 46 pairs of TF interactions involved 22 TFs in K562 [69].

2.1.2.6 Enhancer and promoter enriched motifs

We studied whether a predicted motif preferred to occur in enhancers or promoters. We assessed the statistical significance of a preference in two ways by the binomial testing, similarly to what we did in the analysis of the homogenous motif pairs. That is, we calculated the significance by considering the number of sequences only or both the number and the length of sequences.

2.1.2.7 Machine learning methods to distinguish positive from negative EP pairs

We studied how well the predicted motif pairs distinguish positive from negative EP pairs. We described each EP pair with a *4n+1* vector, where 4 entries were for each of the n motif pairs and one entry was for the positive or negative status. The four entries for a motif pair were the occurrence number of its motifs (based on FIMO) in the enhancer and promoter, respectively.

We applied the following four methods (https://scikit-learn.org/stable/), random forests, least absolute shrinkage and selection operator (lasso), decision tree, and support vector machines [106-109], to distinguish positive from negative EP pairs . We did 10-fold cross-validation to measure the performance of different methods. The four methods had similar F1 scores in separating positives from negatives. Because lasso selects a subset of the predicted motif pairs while achieved similar performance, we presented our study with lasso in this study.

2.1.3 Results

2.1.3.1 The predicted motif pairs were likely to be biologically meaningful

We identified 434 motif pairs in interacting EP pairs in seven cell lines (Table 2-2). For every motif pair, at least one motif occurred in enhancers, and the other motif occurred in promoters of significantly many interacting EP pairs. These motif pairs were from the predicted motif modules, each of which contained 2 to 5 motifs. As mentioned above, a motif module is a statically significant group of co-occurring motifs, which represents the motif combination of a TF and its cofactors [110]. The predicted motifs, motif pairs, motif modules, and other information are available at https://doi.org/10.6084/m9.figshare.14192000.

Table 2-2: The predicted motif pairs in seven cell lines.

Cell line (billion)	#enhancers	#promoters	#EP pairs	#predicted motifs	#predicted motif pairs
GM12878(15.1)	2731	2171	3688	51(76.47%)	233(66.52%, 0.86%, 1.23E-14)
HMEC(1.1)	1761	1713	2157	33(87.88%)	88(59.09%, 2.27%, 0)
HUVEC(0.9)	751	650	835	8(100.0%)	5(60.0%, 0, 0)
IMR90(1.7)	2344	2137	3226	53(71.7%)	116(59.48%, 7.76%, 0)
K562(1.3)	2096	1942	2972	48(83.33%)	144(56.25%, 6.25%, 3.33E-16)
KBM7(1.2)	6278	5970	7862	78(53.85%)	264(42.8%, 8.33%, 1.25E-14)
NHEK(1.3)	1160	1018	1313	18(88.89%)	28(89.29%, 7.14%, 4.44E-16)

The sequencing depth is under each cell line name in the first column, in the unit of billion. The percentage in the second last column is the percent of motifs in a cell line identified in other cell lines. The four numbers in the last column are the number of the predicted motif pairs, the percentage of the predicted motif pairs in a cell line identified in other cell lines, the percentage of random motif pairs in a cell line identified in other cell lines, and the p-value of the number of the predicted motif pairs in a cell line identified in other cell lines, respectively.

The identified motif pairs were likely to be biologically meaningful, because we did not discover any motif pair when we carried out the same procedure in random sequences (the first type of negative EP pairs). We generated the corresponding number of random sequences as the original input for each of the seven cell lines by randomly permuting the nucleotides in each original sequence. We could not identify any motif in any cell line by applying the same procedure to these random sequences in each cell line. We thus could not identify any motif pair, implying the biological significance of the identified motif pairs.

The predicted motifs also corroborated the biological significance of the identified motif pairs. Motif pairs were composed of pairs of motifs predicted in the corresponding cell line. We noticed that on average, we independently discovered, more than 80% of the predicted motifs in different cell lines. The re-discovered motifs in multiple cell lines were not due to the shared EP pairs. We removed the shared EP pairs between every pair of GM12878, IMR90 and KBM7, which had the largest number of EP pairs, we could still find about 75% of the predicted motifs shared between every pair of the three cell lines. The independent discovery of the majority of motifs in other cell lines supported that these motifs were likely biological meaningful, which corroborated the function of the predicted motif pairs. Moreover, we also noticed that, on average, more than 55% of motifs in a cell line were similar to the annotated known motifs [90], further supporting the biological significance of the identified motif pairs in different cell lines.

The conservation of the identified motif pairs supported their biological significance as well. On average, more than 62% of motif pairs in one cell line were independently identified in other cell lines (Table 2-2). By randomly choosing the same number of motif pairs in each of the seven cell lines, we never had more than 10% random motif pairs discovered in other cell lines (p-value<1.25E-14). After removing similar motif pairs (both pairs of motifs had STAMP E-value<1E-08), we obtained 423 non-redundant motif pairs in seven cell lines. The conservation of the identified motif pairs suggests that these motif pairs were likely to be biologically meaningful.

2.1.3.2 The predicted motif pairs were enriched with motif pairs of interacting TFs

In addition to the above evidence that supported the predicted motif pairs, we noticed that the TFs binding to these motif pairs is likely to interact. We obtained the TFs that may bind to a motif by comparing the predicted motifs with known motifs. In this way, we obtained the predicted TF pairs for the corresponding predicted motif pairs. We then compared the predicted TF pairs with the known interacting TF pairs extracted from BioGRID [105] (Material and Methods). We found that the predicted motif pairs were significantly enriched with those of interacting TFs in BioGRID.

In brief, in every cell line, we obtained TF pairs corresponding to the predicted motif pairs. Multiple TFs may bind to the same motif. We thus considered the TFs for a predicted motif in two ways: one was to include all TFs with their motifs similar to a predicted motif as the TFs of this predicted motif, and the other was to consider only the TF with the most similar motif as the TF of a predicted motif (STAMP E-value <1E-05 in both cases). In this way, we obtained two sets of TF pairs for the predicted motif pairs in every cell line (Figure 2-2 and Table 2-3). We then compared each of the two sets of TF pairs with the interacting TF pairs in BioGRID. The interacting TF pairs in BioGRID interacted directly or indirectly through a third common protein (Material and Methods). We found that the predicted interacting TF pairs were significantly enriched with the known interacting TF pairs in BioGRID in almost every cell line by hypergeometric testing (Figure 2-2 and Table 2-3).

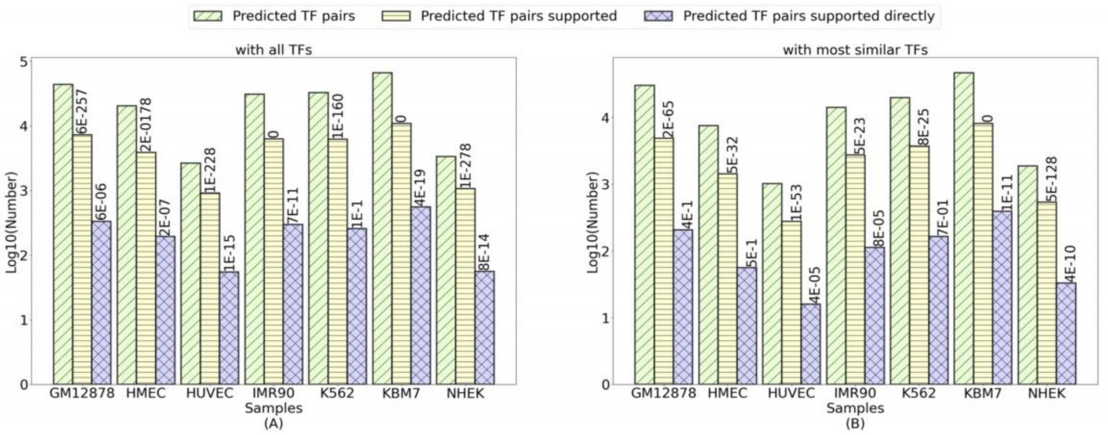


Figure 2-2: The predicted motif pairs are enriched with known interacting TF pairs.

Table 2-3: Comparison of the predicted TF interaction with known ones in BioGRID.

cell line	#predicted TFs supported by BioGRID	#predicted direct TFs pairs supported by BioGRID	P-value of #predicted direct TF pairs supported by BioGRID	#predicted TF pairs supported by BioGRID	P-value of # TF pairs supported by BioGRID
GM12878	299 (265)	335 (208)	5.7E-06 (4.4E-01)	7252 (4868)	6.4E-257 (1.5E-65)
HMEC	213 (139)	194 (56)	2.4E-07 (5.0E-01)	3902 (1430)	6.1E-178 (4.7E-32)
HUVEC	73 (44)	55 (16)	1.13E-15 (4.1E-05)	912 (277)	8.66E-228 (1.3E-53)
IMR90	262 (205)	297 (113)	7.0E-11 (8.2E-01)	6303 (2753)	0 (5.3E-23)
K562	285 (241)	258 (164)	1.0E-01 (6.9E-01)	6207 (3733)	1.2E-160 (8.1E-25)
KBM7	359 (307)	557 (392)	4.4E-19 (1.19E-11)	10876 (8179)	0 (0)
NHEK	78 (57)	56 (33)	7.5E-14 (4.0E-10)	1071 (536)	1.2E-278 (4.5E-128)

In each entry, the information in order is the result based on all TFs for each predicted motif with STAMP cutoff 1E-05, and the result based on the most similar TF for each predicted motif with STAMP cutoff 1E-05. The p-value in 4th and 6th column is calculated based on hypergeometric testing. There are 1520 TFs in BioGRID, which are supported by GO. And there are 6820 TF pairs in BioGRID based on the 1520 TFs

Previously, Zhang et al. studied the ChIP-seq data and Hi-C data and computationally predicted the interactions of 61 TFs in GM12878 and 22 TFs in K562 [69]. We compared the predicted TF pairs in this study with theirs. There were 27 and 10 TFs in GM12878 and K562, respectively, shared by Zhang et al.'s study and this study. The two studies did not share all TFs, because certain TFs do not have sequence-specific binding motifs or do not have a known motif. There were 55 interactions in GM12878 and 4 interactions in K562 involving these shared TFs identified in Zhang et al.'s study. We identified 46 of the 55 interactions in GM12878 and 4 of 4 interactions in K562 (p-value 4.0E-27 and 0, respectively).

We investigated why we did not predict the remaining 9 interactions in GM12878. We found that we predict at least 8 of these 9 TF interactions. The motif pairs corresponding to these 8 TF pairs did not satisfy the motif similarity cutoff 1E-05 when we compared the predicted motifs with the known motifs. We also examined the motif pairs that were composed of known motifs and predicted in GM12878. We could identify all of the 55 TF pairs in GM12878, including all TF pairs of the missing 9 TF pairs. Moreover, we similarly compared the TF interactions predicted by Zhang et al. with the BioGRID. Zhang et al. predicted much fewer interactions, and the enrichment p-values of their predictions were much larger (Table 2-4).

Table 2-4: Comparison between Zhang et al.'s study with BioGRID

cell line	#predicted TFs supported by BioGRID	#predicted direct TFs pairs supported by BioGRID	P-value of # predicted direct TF pairs supported by BioGRID	#predicted TF pairs supported by BioGRID	P-value of # TF pairs supported by BioGRID
GM12878	61	25	5.8E-05	155	1
K562	22	4	1.3E-2	34	3.0E-2

2.1.3.3 The predicted motif pairs were supported by EN-CODEC annotation

We compared the predicted motif pairs in GM12878 and K562 with the EN-CODEC annotation [111]. EN-CODEC did not provide motif pairs or TF pairs. Instead, it annotated TFs that bind to enhancers and promoters of individual gene based on TF-specific ChIP-seq data in GM12878 and K562. Its enhancers were defined computationally based on 10 histone markers and integrated with additional experimental evidence. The enhancers were connected to their target genes by computational methods and filtered with experimental data such as Hi-C data. Because of the computational nature of the predicted enhancers and EP pairs in EN-CODEC, together with the fact that the binding of the cofactors instead of a TF under consideration may result in the discovery of the binding of this TF instead of its cofactors in ChIP-seq, the TF-gene relation annotated in EN-CODEC may have both false positives and false negatives.

From the annotation, we defined an EN-CODEC TF pair as a pair of TFs with known motifs, in which one TF bound to enhancers and the other TF bound to promoters of the same genes for more than 30 genes. The cutoff 30 was for the consistency purpose, as each of our predicted motif pairs above occurred in at least 30 EP pairs. We considered only TFs with known motifs, because we could only compare the predicted motif pairs with TF pairs of known motifs. In this way, we obtained 1379 and 4390 EN-CODEC TF pairs in GM12878 and K562, respectively, which consisted 67 TFs in GM12878 and 109 TFs in K562 (Table 2-5). We predicted motifs for 31 of the 67 TFs in GM12878 and 57 of the 109 TFs in K562. For motif pairs composed of these predicted motifs, more than 77% of motif pairs in GM12878 and all motif pairs in K562 were supported by the EN-CODEC TF pairs, indicating a high precision of our predicted motif pairs. On the other hand, fewer than 12% of EN-CODEC TF pairs were supported by our motif pairs.

The much lower percentage of EN-CODEC TF pairs were supported, likely due to the large percentage of false positive EN-CODEC TF pairs we obtained above. Here we had only 67 TFs in GM12878 and 109 TFs in K562, while we had 1379 TF pairs in GM12878 and 4390 TF pairs in K560 (Table 2-5). In other words, more than 62% and 74% of all possible TF pairs regulated more than 30 genes, which was highly unlikely, indicating that the cofactors in ChIP-seq data may have biased the defined the TF-gene relation in EN-CODEC. In fact, Zhang et al. integrated the same Hi-C and TF-specific ChIP-seq data in GM12878 and K562 and obtained much fewer TF pairs. Moreover, we could only identify 77 motif pairs in GM12878 and 490 motif pairs in K562 with known motifs of these TFs by the aforementioned ChIPModule analyses we did, suggesting that the majority of the defined EN-CODEC TF pairs did not occur in EP pairs of enough genes to be statistically significant. In other words, although we may have missed certain motif pairs that contribute to EP interactions, at least more than 77% of the predicted motif pairs were likely biologically meaningful.

Table 2-5: EP motif pair comparison with EN-CODEC.

Cell line	method	% predicted motif pairs shared with EN-CODEC	% TF pairs in EN-CODEC identified
GM12878	Based on all TFs	64/75=85.33%	87/1379=6.31%
	Based on unique TFs	51/66=77.27%	50/1379=3.63%
K562	Based on all TFs	25/25=100.00%	490/4390=11.16%
	Based on unique TFs	22/22=100.00%	237/4390=5.40%

'Based on all TFs' is the result based on all TFs with their motifs similar to each predicted motif (STAMP E-value<1E-05). 'Based on unique TF' is the result based on the TF with its motif most similar to each predicted motif (STAMP E-value<1E-05).

2.1.3.4 The predicted motif pairs can help to distinguish positive EP pairs from negative ones

Since the predicted motif pairs were likely to be biologically meaningful, we tested whether they could help to distinguish positive EP pairs from negative ones (Material and Methods). We found that the predicted motif pairs separated the positive EP pairs from the first two types of negative EP pairs well and reasonably distinguish the positive EP pairs from the third type of negative EP pairs (Table 2-6). All had the F1 score larger than 0.66.

Table 2-6: The accuracy of motif pairs in distinguishing positive EP pairs from three types of negative EP pairs based on lasso.

Cell line	1st type	2nd type	3rd type	#selected motif pairs	%selected motif pairs shared
GM12878	(0.91,0.92,0.92)	(0.86,0.76,0.80)	(0.69,0.87,0.77)	(78, 96, 70)	43/70=61.43%
HMEC	(0.90,0.90,0.90)	(0.85,0.72,0.78)	(0.52,0.99,0.68)	(66, 58, 36)	26/36=72.22%
HUVEC	(0.83,0.88,0.85)	(0.67,0.79,0.70)	(0.51,1.00,0.67)	(5, 5, 5)	5/5=100.00%
IMR90	(0.91,0.92,0.92)	(0.91,0.79,0.84)	(0.50,0.99,0.67)	(56, 86, 43)	18/43=41.86%
K562	(0.91,0.90,0.91)	(0.87,0.75,0.81)	(0.50,0.97,0.66)	(71, 102, 53)	25/53=47.17%
KBM7	(0.91,0.89,0.9)	(0.90,0.72,0.80)	(0.59,0.90,0.71)	(107,108,53)	16/17=94.12%
NHEK	(0.89,0.90,0.89)	(0.65,0.78,0.70)	(0.51,0.99,0.67)	(23, 24, 17)	16/40=40.00%
Average	(0.89,0.90,0.90)	(0.82,0.76,0.78)	(0.55,0.96,0.69)	(58,68,40)	55.46%

The three numbers from the 2nd column to the 4th column are the precision, recall and F1 score. The second last column is the number of motif pairs selected by lasso in distinguishing positives from negatives for the three types of negatives in order. The last column shows the percentage of the selected motif pairs based on the third type of negatives by lasso in multiple cell lines.

We tried to determine how well the identified motif pairs could differentiate the positive EP pairs from the first two types of negative ones (Material and Methods). These two types of negative ones were “false” EP pairs. We found that the predicted motif pairs told the positive EP pairs apart from the first type of negative EP pairs with an average precision of 0.89, and a recall of 0.90 in individual cell lines in 10-fold cross validation. Similarly, on average, the predicted motif pairs distinguished the positive EP pairs from the second type of negative EP pairs with an average precision of 0.82, and a recall of 0.76 in the 10-fold cross-validation (Table 2-6).

We also studied how well the predicted motif pairs separated the positive EP pairs from the third type of negative EP pairs. In the 10-fold cross validation, the precision in all cell lines was from 0.50 to 0.69, while the recall was from 0.87 to 1 (Table 2-6). The much-reduced precision was likely because the number of negative EP pairs was much larger than that of positive EP pairs. We also noticed that the F1 score was decreasing from the first type of negatives to the third type of negatives, suggesting that it was more difficult to distinguish positive EP pairs from the third type of negatives than that from the first type of negatives. However, the F1 score was still above 0.66, indicating that the predicted motif pairs could facilitate to distinguish the true EP interactions from the false ones. In total, lasso selected 5 to 70 motif pairs in a cell line, which corresponded to 147 non-redundant motif pairs (Table 2-6). There were 30 motif pairs selected independently in at least two different cell lines.

We studied whether the predicted motif pairs in one cell line could distinguish the positive EP pairs from the third type of negative EP pairs in another cell line. The identified motif pairs in one cell line had similar precision and recall to distinguish the positive EP pairs from the third type of negative EP pairs in every other cell line to the predicted motif pairs from the corresponding cell line. This suggested that a large proportion of the predicted motif pairs in one cell line were likely to be conserved in another cell line. In other words, the predicted motif pairs represented conserved mechanisms across cell lines. We noticed that different cell lines shared not only the majority of the predicted motif pairs but also the majority of selected motif pairs used to distinguish positive EP pairs from the third type of negative EP pairs (Table 2-6).

2.1.3.5 The selected motif pairs are likely to contribute to EP interactions

We studied whether the selected motif pairs contribute to EP interactions. Starting from the above 147 selected motif pairs, we identified pairs of TFs with their motifs similar to the selected motif pairs (STAMP E-value<1E-5). We could identify TF pairs for 72 of the above 147 selected motif pairs and 19 of the 30 motif pairs selected in multiple cell lines. For 64 of the 72 selected motif pairs and 18 of the 19 selected conserved motif pairs, their corresponding pairs of TFs interact in BioGRID. For at least 45 of the 72 pairs and 14 of the 19 pairs are shown to contribute to EP interactions in literature, among which 40 of the 72 pairs and 14 of the 19 pairs are supported by both BioGRID and literature. We provided two examples of the TF pairs corresponding to these selected motif pairs in the following.

An example of a novel motif pair selected is for the TF pair GATA1-ZNF423. GATA1 is known to bind to distal regions and physically interacts with ZFPM1 in the beta-major globin promoter [112]. Similar to ZFPM1, ZNF521, a paralog of ZNF423 that shares 65% of homology with ZNF423, is known to have a functional NuRD sequence at the N-terminal [113, 114]. Moreover, ZNF521 modulates erythroid cell differentiation through direct binding with GATA1 [115]. It is thus evident that GATA1-ZNF423 interaction is likely to facilitate EP interaction, which may be through the GATA1 interaction with the NuRD sequence at the N-terminal of ZNF423.

Here is another novel motif pair that may facilitate EP interactions. This selected motif pair is for the TF pair EBF1-ZNF143. In vertebrates, the EBF1 is demonstrated to have the role of controlling the higher-order chromatin structure [116]. ZNF143 is known to preferentially occupy anchors of chromatin interactions connecting enhancers and promoters [117]. Moreover, EBF1, ZNF143, and RAD21 have a three-way interaction in GM12878 [116]. It is thus likely that the interaction of EBF1-ZNF143 may contribute to EP interactions [76].

2.1.4 Conclusion

We de novo identified 423 motif pairs in interacting EP pairs. These motif pairs were likely to be biologically meaningful because they were statistically significant, conserved across cell lines, enriched with motif pairs of known interacting TFs, and so on. We also demonstrated that the predicted motif pairs could help to distinguish positive EP pairs from negative ones. We provided the predicted motifs, motif pairs, and other related information about these motifs and motif pairs at https://doi.org/10.6084/m9.figshare.14192000.

We identified 1183 motif pairs in interacting EP pairs with known motifs as well (Table 2-7). We found that most of the identified motif pairs based on known motifs were similar to those de novo predicted ones in the corresponding cell lines. For instance, in KBM7, 94% of the identified motif pairs based on known motifs were similar to the de novo predicted motif pairs. A small fraction of the motif pairs based on known motifs were not discovered in the de novo predicted motif pairs, likely due to the STAMP E-value cutoff 1E-05 used.

Table 2-7: Predicted motif pairs from known motifs.

cell line	#enhancers	#promoters	# EP pairs	#predicted motifs (%shared)	#predicted motif modules	#predicted motif pairs
GM12878	2371	2171	3688	50 (96.0%)	10425	444 (95.3%, 12.2%, 2.1E-14, 89.4%)
HMEC	1761	1713	2157	31 (100.0%)	736	98 (100.0%, 14.3%, 3.6E-15, 86.7%)
HUVEC	751	650	835	10 (100.0%)	41	13 (100.0%, 30.8%, 2.2E-07, 46.2%)
IMR90	2344	2137	3226	48 (97.9%)	9770	477 (86.6%, 10.5%, 3.2E-15, 89.5%)
K562	2096	1942	2972	48 (97.9%)	7480	376 (91.8%, 10.9%, 0, 88.0%)
KBM7	6278	5970	7862	74 (76.6%)	28701	989 (48.6%, 13.7%, 4.8E-14, 94.0%)
NHEK	1160	1018	1313	12 (100%)	118	24 (100.0%, 16.7%, 1.4E-15, 79.2%)

The percentage in the “#predicted motifs” column is the percent of motifs in a cell line identified in other cell lines. The four numbers in the last column are the number of the predicted motif pairs, the percent of the predicted motif pairs in a cell line identified in other cell lines, the percentage of random motif pairs in a cell line identified in other cell lines, and the p-value of the number of the predicted motif pairs in a cell line identified in other cell lines, and the percentage of motif pairs found in the de novo predicted motif pairs independent of known motifs in the paper, respectively.

We noticed that more than 55% of predicted motifs were similar to known motifs in one cell line. We also observed that more than 80% of the predicted motifs in one cell line were usually identified in other cell lines. In addition, we studied whether the predicted motifs preferred to occur in enhancers and promoters (Table 2-8). Without considering the sequence length difference between enhancers and promoters, almost all motifs preferred to occur in promoters in all cell lines. When we considered the sequence length difference, where on average the promoters were three times longer than the enhancers, there was barely any motif preferring promoters to enhancers. Therefore, the majority of motifs occurred in both enhancers and promoters, with more frequent occurrence of their binding sites in enhancers.

Table 2-8: Almost all motifs (SIOMICS) are likely to occur in both enhancers and promoters.

cell line	# predicted motifs	with length			without length		
		%Enhancer motif	%Promoter motif	%Enhancer and Promoter motif	% Enhancer motif	% Promoter motif	% Enhancer and Promoter motif
GM12878	241	51.5%	2.1%	46.4%	0.4%	99.2%	0.4%
HMEC	83	45.8%	0%	54.2%	2.4%	96.4%	1.2%
HUVEC	14	14.3%	7.1%	78.6%	0.0%	100.0%	0.0%
IMR90	190	61.1%	1.5%	37.4%	0.0%	99.5%	0.5%
K562	180	49.4%	2.2%	48.3%	0.5%	98.9%	0.6%
KBM7	428	57.0%	1.2%	41.82%	0.7%	98.8%	0.5%
NHEK	29	31.0%	13.8%	55.2%	0.0%	100.0%	0.0%

We also checked whether there were homogeneous motif pairs that have the same motifs significantly occurring in both enhancers and promoters, such as the aforementioned CTCF-CTCF motif pair and the YY1-YY1 motif pair (Material and Methods). If we considered the sequence lengths, 78.6% to 93.1% motifs could form homogenous motif pairs that significantly co-occurred in positive EP pairs, including the CTCF-CTCF motif pair in six of the seven cell lines and the YY1-YY1 motif pairs in five of the seven cell line. Even if we did not consider the sequence length, we still could identify 13, 5, and 158 motifs that could form homogeneous motif pairs in GM12878, HMEC and KBM7. In this case, CTCF was still found in HMEC. If we lower the STAMP E-value cutoff when comparing the predicted motifs with known motifs, the predicted motifs similar to CTCF and YY1 were found in both GM12878 and KBM7. We provided two lists of homogeneous motif pairs based on the two different considerations at https://doi.org/10.6084/m9.figshare.14192000 for future validation studies.

CHAPTER 3: DISTAL REGULATORY REGIONS OF HUMAN RIBOSOMAL PROTEINS GENES

Previously published as Wang S, Hu H, Li X. Shared distal regulatory regions may contribute to the coordinated expression of human ribosomal protein genes. Genomics. 2020 Jul;112(4):2886-2893.

3.1 Shared distal regulatory regions may contribute to the coordinated expression of human ribosomal proteins genes

3.1.1 Introduction

It is important to study the transcriptional regulation of ribosomal protein genes (RPGs) [38, 118]. RPGs are house-keeping genes that code for the structural proteins in the ribosome, the machine that makes proteins in every organism. In addition to their ribosome-related function, RPGs have also been involved in other functions and their dysfunction may result in various diseases [119, 120]. As a set of essential genes and one type of the most abundantly expressed genes [121], RPGs are well known for their coordinated expression, meaning that in a given species, their mRNA expression levels are highly correlated across various experimental conditions [37]. To study RPG transcriptional regulation is thus fundamentally important, not only for our understanding of the molecular basis of their functions, but also for deciphering the general principles of gene transcriptional regulation especially coordinated gene regulation [38, 39].

Many studies have been carried out to understand how RPGs are coordinately regulated. Early experimental studies showed that several RPGs share transcription factor (TF) binding sites (TFBSs) of a common TF and validated the regulatory roles of these TFBSs [40, 41] . Later, high-throughput experiments showed that TFs such as RAP1 and FHL1 bind to their TFBSs in promoters of almost all RPGs in yeast [42, 43]. With the genomes of human and other organisms available, computational studies became popular and demonstrated that there are TFBSs of common TFs spread in promoters of almost all RPGs in a species [37, 44-46].

All above studies focused on RPG promoter regions. Rarely is a study that explores the distal regulatory regions of RPGs. Here and in the following, promoters were defined as previously [33, 78] as the upstream 1000 base pairs (bps) to the downstream 100 bps of RPG transcriptional start sites (TSSs); and distal regions were defined as genomic regions that were at least 2500 bps away from the annotated genes. To fill this gap, we previously studied the putative regulatory regions within one megabase (Mbps) of the 80 human RPGs with the DNase I hypersensitive sites (DHSs) in 349 samples [33]. For the sake of simplicity, henceforth, we used a “sample” to refer to a cell line, a cell type, or a tissue under an experimental condition. We identified 217 putative regulatory regions of RPGs that are shared by the majority of the 349 samples. More than 86% of these shared regulatory regions were supported by the chromatin interaction data .

Although our previous study shed new light on human RPG transcriptional regulation, it is limited in the following aspects [33]. First, not all identified regions interacted with RPG promoters and thus they may not be RPG regulatory regions. Second, the previously identified regions are shared across the majority (>=85%) of the 349 samples and are limited in terms of studying sample-specific regulation of human RPGs . Third, these regulatory regions are limited to 1 Mbps neighborhood of RPGs, while regulatory regions may be more distal than 1 Mbps[47].

To understand human RPG distal regulation better, in this study, we defined sample-specific putative RPG regulatory regions directly from high-throughput chromatin interaction data in eleven samples [10, 48]. We identified about 22797 putative RPG regulatory regions, the majority of which were distal regions. More than 44% of these regions were only identified in one sample, implying that RPGs were likely differentially regulated in different samples. Interestingly, 2 to 77 RPGs shared a common regulatory region in a sample and the same pairs of RPGs shared common regulatory regions across samples, which may partially explain their coordinated gene expression. By studying the overrepresented TF binding motifs in these regions in a sample, we identified common TF binding motifs shared by samples. Our study shed new light on the distal regulation of the human RPGs.

3.1.2 Material and Methods

3.1.2.1 Human RPGs and high-throughput chromatin interaction data

We obtained the coordinates of the 80 human RPGs from the *National Center for Biotechnology Information*. We compared the obtained RPG coordinates with those at the RPG database (http://ribosome.med.miyazaki-u.ac.jp/) and found that they were consistent.

We obtained chromatin interaction data from two studies (Table 3-1). One was the Hi-C data in seven cell lines (GM12878, IMR90, HMEC, KBM7, HUVEC, NHEK, K562) from Rao et al. [10] . Rao et al. defined high-confidence interacting pairs of genomic regions called looplists, the number of which was too small to be used here. We thus downloaded their normalized contact matrix for each of the above seven samples from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE63525. Rao et al. generated these contact matrices by the Knight and Ruiz normalization vectors. Knight and Ruiz normalization vectors [10]. They provided the normalized number of Hi-C reads that supported the interaction of the two corresponding genomic regions. We considered the pairs of genomic regions with at least 30 supporting normalized Hi-C reads as interacting pairs of regions in this study. Here 30 was the largest cutoff that enabled the inclusion of more than 99% of the defined interacting regions by other studies in two common cell lines, IMR90 and K562 [63, 64]. Note that these pairs of interacting regions can be from different chromosomes, although the majority of them are intra-chromosome interactions. We obtained the corresponding DHS data for each of the seven samples from the ENCODE project [93] (https://www.encodeproject.org/search/?type=Experiment).

Table 3-1: The identified RPG regulatory regions in the two datasets.

Data Source	Sample	#RPGs involved (simulation)	#Direct regions (simulation)	#Indirect regions (simulation)	Mean length	Medium length	Mean distance (Mbps)	Medium distance (Mbps)	#Unique reads (million)
Rao (16588, 15148)	GM12878	79 (68)	2226 (741)	14315 (7924)	5000	5000	48.3	35.7	15112.0
	HMEC	47 (17)	85 (25)	305 (79)	5000	5000	45.3	26.3	1068.0
	HUVEC	41 (10)	68 (14)	270 (56)	5000	5000	40.7	32.7	892.8
	IMR90	78 (52)	347 (164)	1634 (821)	5000	5000	47.9	30.7	1683.1
	K562	75 (40)	351 (102)	1549 (679)	5000	5000	53.5	38.1	1366.2
	KBM7	59 (31)	112 (52)	684 (302)	5000	5000	45.1	19.0	1247.9
	NHEK	49 (23)	102 (35)	461 (162)	5000	5000	37.8	18.8	1347.5
Javierre (6209,3522)	nB	73(9)	528 (26)	2515 (232)	5175	3914	51.8	37.1	2127.3
	tCD8	68(9)	580 (29)	2600 (308)	5105	3789	45.0	20.5	1849.2
	FoeT	66(7)	484 (17)	2418 (201)	5098	3790	43.4	24.3	2728.4
	tCD4	68(10)	586 (29)	3553 (361)	5338	4040	50.2	39.3	2227.4

The other dataset was a promoter capture Hi-C dataset in seventeen primary cell types, where relatively more abundant data were available in eight of the seventeen cell types [48]. These eight cell types were aCD4, nB, EP, tB, tCD8, FoeT, naCD4, tCD4. The interactions between genomic regions were defined in the original study [48]. All pairs of interaction regions were from the same chromosomes. We were able to download the corresponding DHS data for the following four cell types: aCD4, nB, tCD8, FoeT from https://www.encodeproject.org/search/?type=Experiment.

3.1.2.2 Direct and indirect RPG regulatory regions and enhancers

With a chromatin interaction dataset and the corresponding DHS data in a sample, we obtained direct and indirect regulatory regions of RPGs in this sample (Figure 3-1). A direct region in a sample is a region overlapping with at least one DHS region and interacting with another region that overlaps with RPG promoters . The overlap of two genomic regions is done by the bedtools (https://bedtools.readthedocs.io/en/latest/content/tools/intersect.html) with the following command: bedtools intersect -a a.bed -b b.bed -wao > out.bed. The interaction and DHS regions used are defined in the corresponding sample. Similarly, an indirect region is a region overlapping with at least one DHS region and interacting with another region that overlaps with a direct region or an indirect region. Note that a RPG may have multiple direct and indirect regions, an indirect region may interact with another indirect region of the same RPG, and a direct region of a RPG may be an indirect region of another RPG in the same sample.

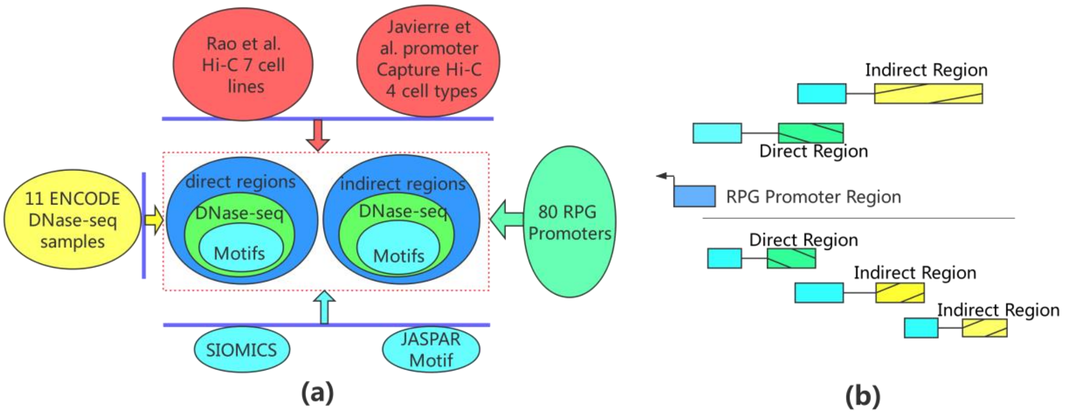


Figure 3-1: The identification of the putative RPG regulatory regions (a) Different sources of interaction data were used to infer RPG regulatory regions and the enriched TF binding motifs in these regions; (b) An example of direct and indirect regulatory regions of a RPG.

Each direct or indirect region was about 5000 bps long, which depended on the Hi-C resolution and was on average much longer than known regulatory regions [110, 122]. To predict TF binding motifs in these regions, we considered the minimum sub-regions within a regulatory region that contained all overlapping DHSs in this region. When the minimum regions were shorter than 800 bps, we extended them equally on both sides of these regions so that the regions were at least 800 bps. The reason to extend short region was that the length of the known mammalian regulatory regions are normally in this range based on previous studies and the DHS data may not be perfect [110, 122]. We then obtained the DNA sequences for these processed regions.

3.1.2.3 Motif analyses in promoters and enhancers

For a given set of sequences, such as the set of sequences from all potential RPG regulatory regions in a sample, we predicted the overrepresented motifs in these sequences by the SIOMICS tool [98, 123]. SIOMICS considers the co-occurrence and overrepresentation of various combinations of patterns (initialized with 8-mers, 8 bps long DNA segments) in the input sequences to identify motifs through an effective tree structure and algorithm, which showed good performance in previous studies [103, 123]. The combination of motifs output from SIOMICS are called motif modules, which represent groups of motifs and their cofactor motifs. We considered motif modules in input sequences, as in high eukaryotes, it is the TFBSs of different TFs in a short region to form cis-regulatory modules to control the gene expression patterns [122].

We compared the predicted motifs with the motifs in the JASPAR database [124]. The JASPAR database is widely used for its manually annotated TF motifs. We claimed a predicted motif was similar to a known motif in JASPAR if it had a STAMP [95] similarity E-value smaller than 1e-5, a cutoff used in previous studies [97, 125].

3.1.2.4 Other analyses

We downloaded the normalized gene expression data in 79 different tissues from UCSC(GNF Expression Atlas 2)[126], which is widely used to study gene transcriptional regulation[122, 127]. For every pair of human RPGs, we calculated their Spearman's correlation coefficient. We then compared the correlation of RPG pairs with a common distal regulatory region and the correlation of RPG pairs without a common distal regulatory region by the Wilcoxon test[128].

3.1.3 Results

3.1.3.1 About 22797 regions may regulate human RPGs

We studied the direct and indirect regulatory regions of RPGs in eleven samples based on the high-throughput chromatin interaction data [10, 48] and the DHSs in the corresponding samples [93] (Material and Methods) (Figure 3-1). We used the interaction data from two studies, because both had multiple samples with a high sequencing depth. The sequence depth is the ratio of the sum of the length of all uniquely mapped Hi-C reads in a sample to the length of the human genome. In a sample, a direct region of a RPG is a region that physically interact with this RPG promoter based on the corresponding chromatin interaction data and overlap with at least a DHS region in this sample, and an indirect region of a RPG is a region that indirectly interact with this RPG promoter and overlap with at least a DHS region (Material and Methods). In total, we identified about 22797 putative regulatory regions that interacted with RPG promoters in different samples. The majority of these regions were distal regions (Table 3-2) . The number of the putative regions varied across samples. The details were in the following.

Table 3-2: The distance of the identified regulatory regions to RPGs.

Data Source	Cell Line or Type	Mean (Mbp)	Min (bp)	Medium(bp)	Max (Mbp)	Average (Mbp)	# regions in Upstream	#regions in Downstream	#regions in introns
Rao all regions	GM12878	1.3	47	131424	223.9	2.8	4584	4269	93
	IMR90	2.6	47	20464	216.5		472	538	87
	HMEC	3.0	47	5728	149.0		67	84	53
	KBM7	5.0	47	5536	136.2		95	90	67
	HUVEC	5.1	47	7841	111.2		57	74	46
	NHEK	1.7	47	7637	148.2		88	131	58
	K562	1.2	47	17425	188.7		535	518	91
Rao_direct regions	GM12878	0.5	47	70536	134.7	1.2	1150	1043	41
	IMR90	0.7	47	10464	120.0		150	162	36
	HMEC	1.8	47	5421	149.0		31	36	18
	KBM7	1.7	47	5189	60.6		40	41	27
	HUVEC	2.1	47	5728	44.3		26	27	12
	NHEK	1.0	47	5536	69.1		34	40	25
	K562	0.6	47	9811	137.2		148	163	42
Javierre all regions	nB	7.8	28	294687	237.9	9.7	1093	1403	79
	tCD8	9.1	28	406751	234.4		1241	1430	78
	FoeT	3.8	28	433201	112.8		1031	1359	71
	tCD4	18.4	28	1070975	237.3		1658	2058	74
Javierre_direct regions	nB	5.3	3296	325844	237.5	5	265	265	0
	tCD8	5.1	3296	341043	234.4		265	319	0
	FoeT	1.9	3296	391885	65.8		215	270	0
	tCD4	7.7	4138	446695	234.4		276	317	0

In seven samples from Rao et al., we identified 16588 potential RPG regulatory regions. The number of regions in one sample varied from 338 to 16541, depending on the sequencing depth and the nature of the samples (Figure 3-2A, C, E). For instance, in GM12878, there were 2226 direct regions and 14315 indirect regions identified, which was at least eight times of the direct and indirect regions identified in other samples. This was because GM12878 had a sequencing depth about nine to seventeen times of that in other samples [10]. In general, with a larger sequencing depth in a sample, there are more potential RPG regulatory regions identified in this sample (Figure 3-3). However, this is not always true. For instance, KBM7 had a lower sequencing depth than NHEK, while it had more direct and indirect RPG regulatory regions than NHEK. The different number of direct and indirect regions in samples with similar sequencing depth, such as that in K562, KBM7, and NHEK, indicates the sample-specific characteristics (Figure 3-2C). On average, we identified 470 direct and 2745 indirect regions in a sample excluding GM12878.

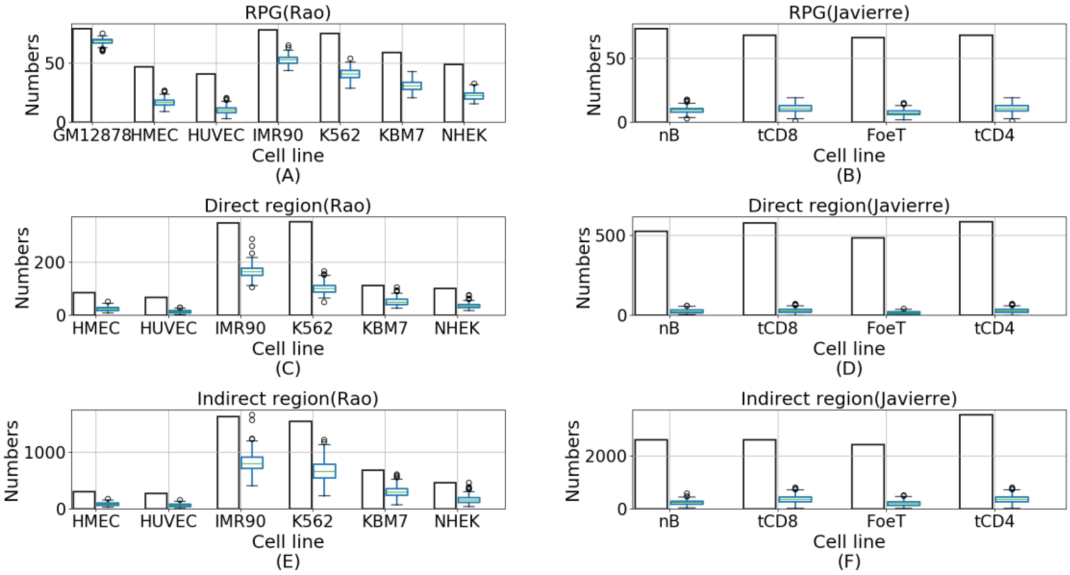


Figure 3-2: The identified putative RPG distal regulatory regions. (**A**) & (**B**) The number of RPGs with identified regulatory regions in a sample; (**C**) & (**D**) The number of identified direct regions in a sample; (**E**) & (**F**) The number of identified indirect regions in a sample. In each section, the box plot is from 200 simulated sets of 80 random genomic regions. There are 2226 direct and 14315 indirect regions identified (741 direct and 7924 indirect regions

identified for random regions) in GM12878, which are not shown in (C) and (E), as they are much larger than the corresponding numbers in other samples.

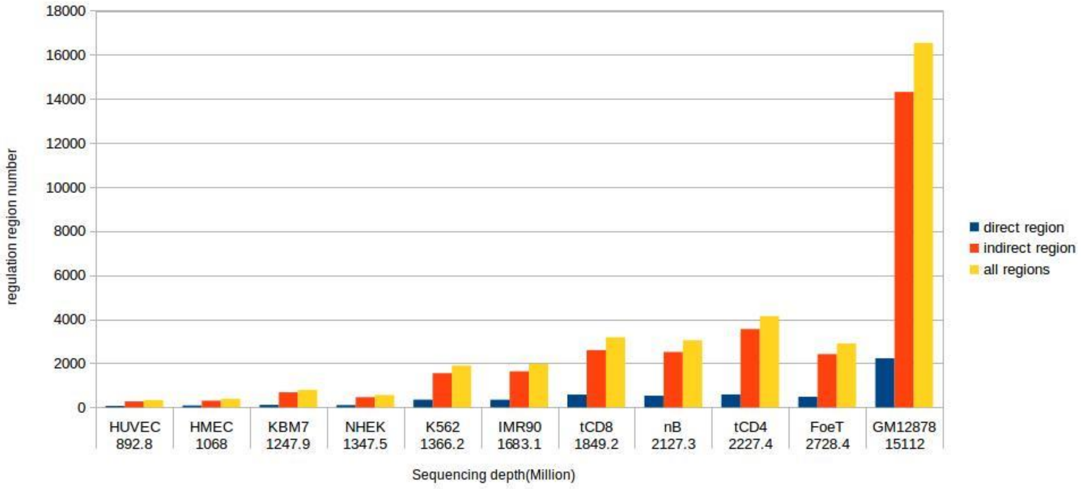


Figure 3-3: The sequencing depth and the number of regulatory regions identified across samples. In both Rao et al.'s samples and Javierre et al.'s samples, the higher sequencing depth does not always mean a larger number of RPG regulatory regions identified.

To assess the statistical significance of the identified regulatory regions, we randomly chosen 80 genomic regions, each of which was the same length as the RPG promoters. We then applied the same procedure to identify direct and indirect regions in each sample for these 80 random regions. We repeated this procedure 200 times with 200 groups of 80 random regions. We identified much fewer direct and indirect regions that interacted with the 80 random regions (Figure 3-2A, 2C, 2E). For instance, in K562, we had 102 direct regions and 679 indirect regions for random regions on average, while there were 351 direct and 1549 indirect regions for the 80 RPGs. This suggested that compared with random genomic regions, RPGs had significant more potential regulatory regions.

Similarly, we identified in total 6209 regions that were likely to regulate RPGs in four samples from Javierre et al [48]. Javierre et al. studied seventeen samples while only four samples had the corresponding DHS data and had enough sequencing depth to have putative regulatory regions for at least 50 RPGs (Figure 3-2B). The number of regulatory regions in a sample varied from 2902 to 4139, depending on the samples instead of the sequencing depth (Figures 3-2D and 3-2F). For instance, the sample FoeT had the largest sequencing depth, while the number of regions identified in FoeT was the smallest. In these four samples, the number of RPG regulatory regions identified was larger than that in all samples from Rao et al. except GM12878. On average, in each sample, we identified 545 direct and 2792 indirect regions, respectively (Figure 3-2D and 3-2F). Compared with randomly chosen genomic regions, on average, there were 25 direct and 275 indirect regions for the 80 random regions in 200 simulations. Interestingly, despite the higher sequencing depth and more RPG regulatory regions identified in Javierre et al.'s samples, the number of RPGs with identified regulatory regions was smaller in Javierre et al.'s samples compared with that in Rao et al.'s samples, which may be due to the bias of the capture Hi-C experiments in identifying chromatin interactions, the unsaturated sequencing depth, sample-specific RPG regulatory regions, etc.

The above direct and indirect regions in a sample were obtained by overlapping the corresponding interacting regions defined by Hi-C with the RPG promoters (Material and Methods). Since the interacting regions were defined at about 5000 bps resolution [10, 48], we loosed the criteria of overlapping of two regions. We claimed two regions overlapping if they were within *x* bps to each other, for *x* to be 1000, 2000, or 5000 bps, respectively. For a given *x*, we defined direct and indirect regions of RPGs similarly as illustrated in Figure 1. We found that the number of the defined RPG direct and indirect regions was similar as that with x equal to 0. This suggested that the RPG direct and indirect regions defined were robust and were not greatly affected by the overlapping criteria. It also indicated that these regions were not close to each other. In fact, the mean and median distance of adjacent regions was 299,917 bps and 10,000 bps, respectively, in Rao et al.'s samples; and 93,913 bps and 3,919 bps, respectively, in Javierre et al.'s samples.

We also studied the distances between the identified regions and their corresponding RPGs (Figure 3-4). In Rao et al.'s data, 55.5% (9210/16588) of these regions were distal regions. The distance between a region and the corresponding RPG had a mean of 2.8 Mbps and a median of 28007 bps. Similarly, in Javierre et al.'s data, 98.9% (6140/6209) of these regions were distal regions. The distance between a region and the corresponding RPG had a mean of 9.7 Mbps and a median of 551403 bps. Since almost all human RPGs have neighboring protein-coding genes within 1 Mbps [33], this suggested that RPGs were not the closest genes to many of these regions.

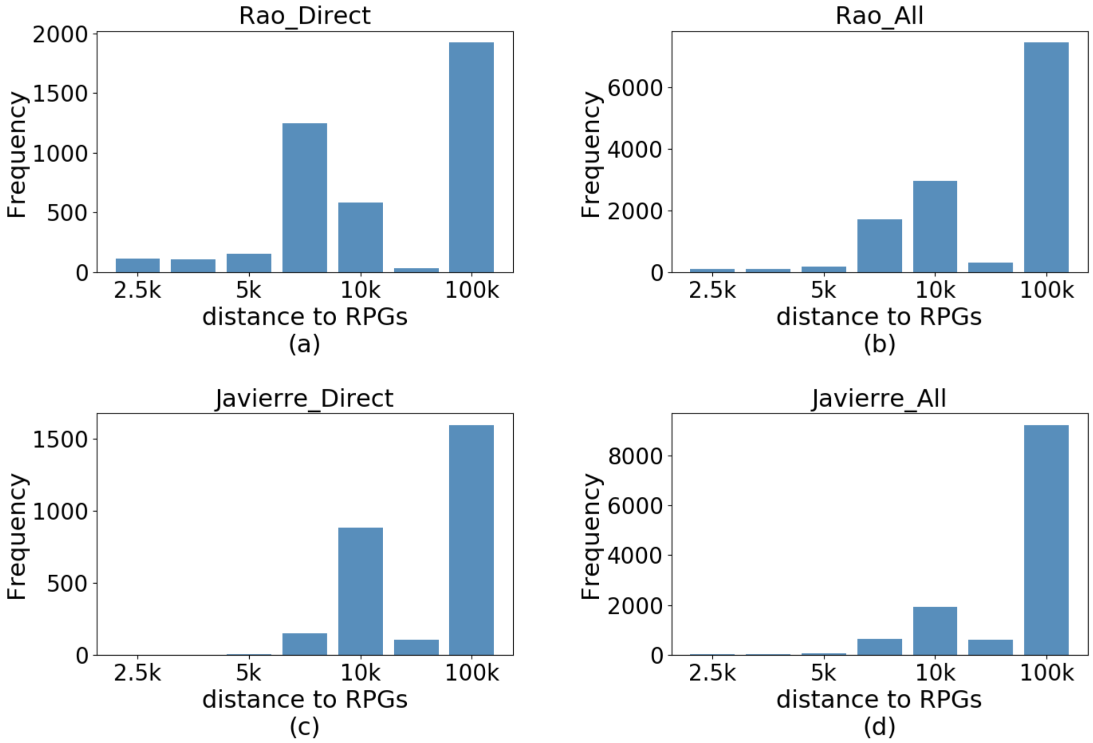


Figure 3-4: The distance between a regulatory region and the corresponding RPG. We divided the distances into seven bins, such as the bin >=2.5k but <5k, where k means kilobase pairs. (**a**) and (**b**) Direct regions and all regulatory regions from Rao et al.'s data, respectively; (**c**) and (**d**) Direct regions and all regulatory regions from Javierre et al.'s data, respectively. 3.1.3.2 The identified putative RPG regulatory regions varied dramatically across samples

We compared the identified RPG regulatory regions in different samples (Table 3-1). We found that the majority of them were not the same and not even overlapping across samples. This suggests that RPGs are likely regulated by different distal regions under different experimental conditions, which is consistent with our previous study [33].

More than 91% (15148) of the 16588 regions in Rao et al.'s data were not shared across samples. Excluding GM12878, which had much higher sequencing depth than other samples, ~80% (2891) of the 3598 regions were identified in only one of the remaining samples. This percentage became smaller for Javierre et al.'s data, where more than 56% (3522) of the 6209 regions were identified in only one sample. A large proportion of the regulatory regions were sample-specific, which were unlikely caused by the difference of the sequencing depth. This was because in all seven samples except GM12878 in Rao et al.'s data and in all four samples in Javierre et al.'s data, the sequencing depth was similar, while the number of identified regulatory regions was very different. Moreover, although GM12878 had a much higher sequencing depth, more than 49.7% of regions identified in other six samples in Rao et al.'s data were not identified in GM12878. It thus implied that RPGs were likely to be regulated differently across different samples.

To assess the statistical significance of the shared regions across samples, we studied the shared interacting regions by the aforementioned 200 sets of 80 random regions. We found that in these 200 simulations, the random regions always had fewer potential regulatory regions but higher percentages of unshared potential regulatory regions across samples (Table 3-1). For instance, there were 3598 regions identified for the 80 RPGs in all seven samples except GM12878 from Rao et al., 80.1% of which did not overlap with any identified region in other five samples. Correspondingly, on average, there were 1837 regions identified for the 80 random regions in these six samples, 90.4% of which did not overlap with any identified region in other five samples. Moreover, the random regions always had lower percentages of regions shared by different number of samples than the 80 RPGs. For instance, in Javierre et al.'s data, there were 11.9% of regions were shared by all four samples for RPGs, compared with the average 6.8% of regions shared by four samples for the 80 random regions (Table

3-3). These observations are consistent with the fact that RPGs and their regulation are more conserved across samples than random regions.

Table 3-3: The comparison of regulatory regions across samples.

	Number of regions	%Regions not shared	%Regions shared by 2 samples	%Regions shared by 3 samples	%Regions shared by 4 samples	%Regions shared by 5 samples	%Regions shared by >=6 samples
Rao	16588 (9400)	91.3% (95.4%)	4.9% (3.2%)	2.2% (1%)	0.6% (0.2%)	0.5% (0.1%)	0.5% (0.1%)
Rao without GM12878	3598 (1837)	80.3% (90.4%)	11.8% (6.6%)	3.2% (1.7%)	2.1% (0.7%)	1.2% (0.4%)	1.4% (0.2%)
Javierre	6209 (672)	56.72% (61.0%)	19.87% (20.8%)	11.53% (11.2%)	11.87% (7.0%)	NA	NA

The number in the parentheses are for the sets of 80 random regions.

To further understand the conservation of these regions across samples, we studied how direct regions were shared across samples (Figure 3-4). The direct regions were those physically interacting with RPG promoters and detected by the Hi-C experiments (Figure 3-1). We found that a large proportion of the direct regions in a sample did not overlap with the direct regions in another samples in both Rao et al.'s and Javierre et al.'s data, suggesting that RPGs are likely to have different regulatory regions across samples. Moreover, fewer than 20% of GM12878 direct regions were found in other samples, which was likely to due to its much larger sequencing depth. In addition, the direct regions in the seven samples by Rao et al. and those in the four samples by Javierre et al. were quite different, indicating the intrinsic difference between the seven cell lines and the four cell types.

All these observations together suggested RPGs are likely to have different regulatory regions across samples. Otherwise, we should have seen that a much larger portion of direct regions shared across samples. For instance, if 90% of the RPG regulatory regions were conserved across samples, we should have seen that two samples shared at least 80% of their regulatory regions. However, this was not case. For instance, there were more than 37% and 42% of HMEC direct regions did not overlap with KBM7 direct regions and HUVEC direct regions, respectively (Table 3-4). In fact, GM12878 had a much higher sequencing depth than other samples, which should include almost all direct regions in other samples, while close to 20% of HUVEC direct regions and more than 56% of direct regions in the four samples considered by Javierre et al. were not identified in GM12878.

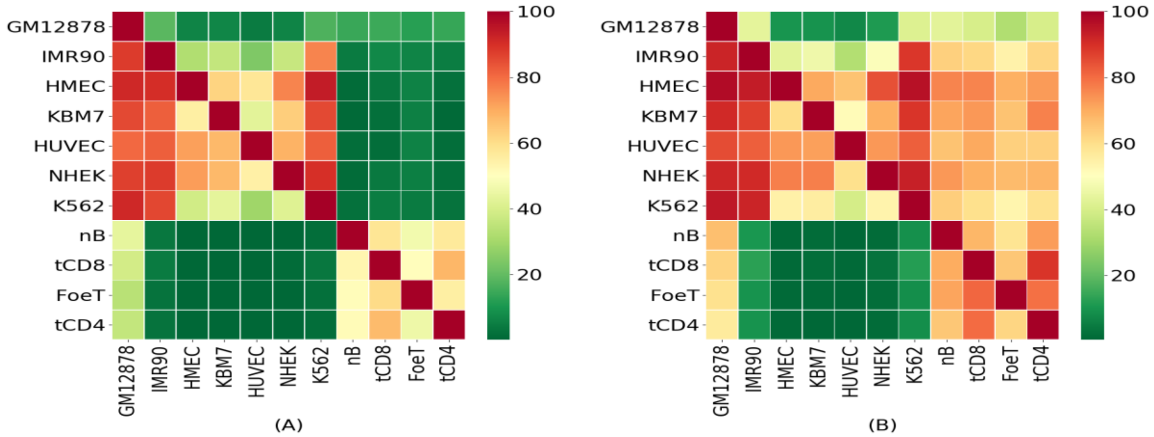


Figure 3-5: The comparisons of regulatory regions across samples. The percentage of direct regions in a sample (row) overlap with (**A**) the direct regions and (**B**) all regulatory regions in another sample (column) is represented by the heatmap.

We also compared the direct regions in a sample with all regulatory regions in another sample (Figure 3-5B). There were more direct regions in a sample identified in another sample, when we considered all regulatory regions instead of direct regulatory regions. However, the increment was small, only a handful of percentage, indicating that the majority of direction regions in one sample were still direct regions in another sample. Although most direct regions were shared across samples when we considered all regulatory regions, there were still a fraction of direct regions not shared, which were likely due to sample-specific regulatory regions. For instance, at a much larger sequencing depth in GM12878, there were still about 15% of direct regions in HUVEC were not identified in GM12878 (Table 3-5).

Table 3-4: The direct regions shared across samples in Rao et al.'s data and Javierre et al.'s data.

	GM12878	IMR90	HMEC	KBM7	HUVEC	NHEK	K562	nB	tCD8	FoeT	tCD4
GM12878	100.0%	18.6%	6.0%	6.6%	4.4%	6.3%	17.0%	14.4%	14.9%	12.1%	14.1%
IMR90	88.2%	100%	32.3%	36.0%	24.2%	36.6%	76.4%	4.3%	6.9%	6.3%	4.6%
HMEC	91.8%	90.6%	100%	62.4%	57.7%	76.5%	94.1%	18%	53%	88%	35%
KBM7	85.7%	82.1%	55.4%	100%	42.9%	63.4%	85.7%	0.9%	1.8%	5.4%	0.9%
HUVEC	80.9%	82.4%	72.1%	67.7%	100%	69.1%	82.4%	1.5%	1.5%	5.9%	1.5%
NHEK	87.2%	88.2%	72.6%	67.7%	54.9%	100%	90.2%	1%	3.9%	4.9%	2%
K562	91.5%	86.3%	38.8%	43.3%	29.1%	41.6%	100%	2.3%	4.8%	4.3%	2.6%
nB	43.8%	3.4%	0.2%	0.2%	0.2%	0.2%	1.7%	100%	58.1%	46.8%	57%
tCD8	39%	4.1%	0.5%	0.3%	0.2%	1%	3.6%	52.9%	100%	50.7%	68.1%
FoeT	34.1%	2.7%	0.4%	0.4%	0.4%	0.4%	2.1%	51%	60.8%	100%	55.2%
tCD4	35.8%	2.6%	0.3%	0.2%	0.2%	0.3%	1.5%	51.4%	67.4%	45.6%	100%

In each entry (*i, j*) , the number is the percentage of the direct regions in the *i*-th sample overlapping with the direct regions in the *j*-th sample.

Table 3-5: The direct regions shared across samples (compared with all RPG regulatory regions identified) in Rao et al.'s data and Javierre et al.'s data.

	GM12878	IMR90	HMEC	KBM7	HUVEC	NHEK	K562	nB	tCD8	FoeT	tCD4
GM12878	100%	43.4%	10.2%	9.8%	7.2%	11.2%	41.7%	43.1%	40.2%	32.7%	40.1%
IMR90	92.5%	100%	43.2%	46.1%	32.9%	49.3%	88.8%	63.1%	61.7%	54.5%	62%
HMEC	97.7%	94.1%	100%	70.6%	65.9%	84.7%	96.5%	76.5%	76.5%	69.4%	72.9%
KBM7	91.1%	87.5%	60.7%	100%	51.8%	69.6%	89.3%	71.4%	73.2%	66.1%	76.8%
HUVEC	85.3%	82.4%	73.5%	72.1%	100%	73.5%	82.4%	66.2%	70.6%	63.2%	63.2%
NHEK	92.2%	91.2%	77.5%	77.5%	59.8%	100%	93.1%	73.5%	69.6%	67.7%	68.6%
K562	94.9%	92.6%	54.7%	55.8%	39.6%	54.1%	100%	63.8%	59.8%	53.6%	59.3%
nB	66.7%	10.4%	1.1%	1.5%	0.6%	1.7%	8.9%	100%	68.4%	58.7%	72.5%
tCD8	62.6%	12.8%	1.2%	2.1%	1.2%	3.3%	12.1%	70.3%	100%	65.3%	89.3%
FoeT	59.3%	9.7%	1%	1%	0.8%	2.1%	8.5%	71.3%	81%	100%	79.8%
tCD4	56.8%	9.7%	1%	0.9%	0.9%	1.7%	8.4%	65.2%	80.4%	62.1%	100%

In each entry (i, j), the number is the percentage of the direct regions in the i-th sample overlapping with all regions in the j-th sample.

We also studied how indirect regions were shared across samples (Table 3-6). The indirect regions were not as conserved as direct regions. In other words, there were an even higher percentage of indirect regions that were not shared by two samples. Moreover, there were much more indirect regions that were not conserved across samples.

Table 3-6: The indirect regions shared across samples in Rao et al.'s data and Javierre et al.'s data.

	GM12878	IMR90	HMEC	KBM7	HUVEC	NHEK	K562	nB	tCD8	FoeT	tCD4
GM12878	100%	8.4%	2%	2.5%	1.6%	2.4%	8.4%	12.6%	11.3%	10.2%	12.1%
IMR90	61%	100%	12.6%	13.5%	9.5%	14.8%	42.2%	28.6%	27.5%	23.3%	28.4%
HMEC	70.8%	61%	100%	37.7%	38%	45.6%	62.3%	45.3%	43.9%	41.3%	41.6%
KBM7	39.2%	27.3%	16.2%	100%	4.3%	19.4%	27.3%	22.1%	21.9%	20.3%	23.1%
HUVEC	61.1%	50.4%	40.4%	37%	100%	38.2%	50.7%	37.8%	37.8%	2.2%	35.2%
NHEK	56.2%	46.2%	29.7%	30.2%	23.4%	100%	44.9%	35.6%	33.4%	30.8%	32.1%
K562	70.6%	52.2%	17.3%	17.2%	12.9%	18.1%	100%	33.5%	31.6%	26.2%	30.6%
nB	48.5%	12.8%	4.3%	5.1%	3.4%	5%	12.8%	100%	49.9%	40.4%	52.2%
tCD8	40.5%	11.9%	3.9%	5%	3.2%	4.6%	11.4%	48.3%	100%	45.3%	68%
FoeT	36.7%	10.1%	3.3%	4.3%	2.4%	3.9%	9.5%	42%	48.7%	100%	52.8%
tCD4	32.9%	9%	2.6%	3.7%	2.2%	3.3%	8.1%	36.9%	49.8%	35.9%	100%

In each entry (i, j), the number is the percentage of the indirect regions in the i-th sample overlapping with the indirect regions in the j-th sample.

3.1.3.3 RPGs shared distal regulatory regions to form putative co-regulated gene clusters

With many regulatory regions identified only in one sample, we attempted to understand how RPGs are coordinately regulated. We hypothesized that in a sample, there may exist a region, which physically interacted with multiple regions that targeted various RPGs. In this way, such a region controls all RPGs and thus may contribute to their coordinate transcriptional regulation. We had no success in finding such a region in any sample. However, we did notice that one region may regulate multiple RPGs in every sample.

We started to identify pairs of RPGs that had at least a pair of their regulatory regions overlapped. In each sample, there was at least one pair of RPGs that had their regulatory regions overlapped (Table 3-8). In other words, these pairs of RPGs shared common regulatory regions in a sample. There were 890 such pairs in GM12878 that involved 77 of the 80 RPGs (except RPS4Y, RPL34 and RPL36A), which was much larger than that in other samples, most likely due to its much larger sequencing depth. In samples other than GM12878, on average, we identified five pairs of RPGs sharing regulatory regions that involved 30 RPGs. The regulatory regions shared by different RPGs may partially explain their coordinated transcriptional regulation.

We tried to understand what characteristics these pairs of RPGs sharing regulatory regions may have. We checked whether these pairs were from the same ribosomal unit. We found that most pairs contained one RPG from the small unit and the other RPG from the large unit. We checked whether these pairs were from the same chromosomes or have a higher sequencing similarity and did not observe such a relationship. We also studied whether these RPG pairs may have more correlated expression (Material and Methods). Indeed, these RPG pairs had significantly larger gene expression correlation across different human tissues than the RPG pairs that did not share any regulatory region (Mann-Whitney test p-value<2E-7). We checked whether these pairs were conserved across samples as well. We found that except those from GM12878, they were indeed quite conserved across samples (Table 3-7). For instance, 100% of the identified RPG pairs in HMEC, HUVEC, KBM7, NHEK and tCD8 were also identified in other samples. As to the 80 random regions, in 200 simulation runs, we barely had any pair of random regions sharing regulatory regions across samples (Table 3-7). The RPG pairs in GM12878 were often not identified in other samples, which was likely due to the much smaller sequencing depth in other samples.

Table 3-7: Clusters of RPGs shared their regulatory regions.

Data source	Sample	#Pairs (#RPGs involved)	%Shared RPG pairs (%shared random pairs) across samples	Loose clusters		Strict clusters	
				#Clusters (#RPGs involved)	Minimum(Maximum) #RPGs in a cluster	# Clusters (#RPGs involved)	Minimum(Maximum) #RPGs in a Cluster
Rao	GM12878	890 (77)	1.91%(1.03%)	1 (77)	77 (77)	820 (77)	2 (14)
	HMEC	1 (2)	100%(0)	1 (2)	2 (2)	1 (2)	2 (2)
	HUVEC	1 (2)	100%(0)	1 (2)	2 (2)	1 (2)	2 (2)
	IMR90	13 (19)	61.54%(0.12%)	6 (19)	2 (8)	13 (19)	2 (2)
	K562	9 (12)	66.67%(0.13%)	3 (12)	2 (8)	9 (12)	2 (2)
	KBM7	4 (8)	100%(0)	2 (8)	2 (2)	4 (8)	2 (2)
	NHEK	2 (4)	100%(0)	2 (4)	2 (2)	2 (4)	2 (2)
Javierre	nB	16 (18)	62.50%(0)	6 (18)	2 (7)	8 (18)	2 (3)
	tCD8	21 (23)	100%(0)	8 (23)	2 (6)	11 (23)	2 (3)
	FoeT	24 (23)	62.50%(0)	8 (23)	2 (5)	10 (23)	2(5)
	tCD4	22 (27)	95.45%(0)	11 (27)	2 (4)	11 (27)	2 (4)

With the above pairs of RPGs in a sample, we grouped them into clusters of RPGs in two ways (Table 3-7). One was the strict way, in which we required that every pair of RPGs in a resulted cluster shared at least one regulatory region. We called the resulted clusters strict clusters. The other was the loose way, where a RPG was added into a cluster if this RPG shared a regulatory region with at least one RPG in that cluster, with the pairs of RPGs identified above as the initial clusters. We called the resulted final clusters by the second way loose clusters. We obtained 1 to 820 strict clusters and 1 to 11 loose clusters in a sample. The strict clusters in a sample contained 2 to 77 RPGs, with the 77 RPGs in different clusters. Similarly, the loose clusters in a sample contained 2 to 77 RPGs, where the 77 RPGs could be in one cluster such as a cluster in GM12878. In terms of 80 random regions, in 200 simulations, except in GM12878, they barely formed clusters in a sample (Table 3-8). Even when they formed clusters, the number of regions involved was much smaller. Most importantly, the pairs of random regions sharing a regulatory region in a sample rarely shared any regulatory region in another sample. In other words, the observed shared regulatory regions by pairs or groups of RPGs may explain the coordinated regulation of RPGs, as their regulatory regions were connected instead of independent.

Table 3-8: Average simulation result of the Cliques (200 runs)

Data source	Sample	#Pairs (#RPG involved)	#Shared pairs across samples	Loose clusters		Strict clusters	
				#Clusters (#RPGs involved)	Minimum (Maximum)#RPGs in a cluster	# Clusters (#RPGs involved)	Minimum (Maximum)#RPGs in a Cluster
Rao	GM12878	195	1.03%	1(64)	42(63)	893(4)	1(7)
	HMEC	0	0	1(1)	1(1)	1(1)	1(1)
	HUVEC	0	0	1(1)	1(1)	1(1)	1(1)
	IMR90	2	0.12%	6(18)	1(13)	18(17)	1(2)
	K562	4	0.13%	2(15)	4(14)	18(15)	1(2)
	KBM7	1	0	4(7)	1(4)	7(7)	1(1)
	NHEK	0	0	3(4)	1(2)	4(4)	1(1)
Javierre	nB	0	0	2(2)	1(1)	2(2)	1(1)
	tCD8	0	0	3(4)	1(1)	4(4)	1(1)
	FoeT	0	0	3(4)	1(1)	4(4)	1(1)
	tCD4	0	0	4(4)	1(1)	4(4)	1(1)

3.1.3.4 RPGs shared common regulatory motifs across samples

To understand why RPGs have coordinated expression patterns, we also studied the putative regulatory motifs in the above RPG regulatory regions. We only considered the DHSs within these regions in the corresponding samples for the motif analysis, as these regions were open for TFs to bind. The average length of these DHS regions was 150 bps, shorter than that of known regulatory regions, which was mostly several hundred bps but can up to a couple of thousand bps [110, 122, 129, 130]. We thus extended each region equally from its two ends if this region was shorter than 800 bps so that the extended regions were at least 800 bps. We then identified motifs in these extended regions by de novo motif discovery [98, 123], as the number of known motifs was still limited [94, 104, 124]. We found that about two dozen motifs were shared by different samples.

By de novo motif discovery (Material and Methods), we identified 68 to 1118 motifs in different samples (Table 3-9). The number of motifs identified in a sample correlated well with the number of RPG regulatory regions identified in this sample, with GM12878 having the largest number of motifs and HUVEC having the smallest number of motifs. To assess the statistical significance of the identified motifs, we permuted the input genomic sequences and identified motifs in the permuted sequences in each sample. We identified at least eight times fewer motifs in the random sequences in every sample (Table 3-9), suggesting that the identified motifs in RPG regulatory regions were statistically significant and likely meaningful.

To assess the biological meaning of the predicted motifs, we further compared the predicted motifs with the known motifs in the JASPAR database [124]. In a sample, 38.28% to 54.36% of motifs were similar to known motifs (STAMP E-value < 1E-5 [95]). Moreover, we compared the motifs predicted in different samples. There were 98.53% to 100% of motifs identified in a sample that were also independently predicted in at least another sample. Note that the majority of the regions in two samples were different, suggesting that these motifs were likely biologically meaningful. In addition, we compared the predicted motifs with known RPG regulating motifs. We collected fourteen motifs that were reported to regulate RPGs in literature [33]. We found that on average, 53.25% of these RPG-regulating motifs were identified in a sample. Note that these RPG-regulating motifs were previously identified in RPG promoter regions, and now we identified them in the RPG distal regions as well. In total, almost all motifs predicted in a sample were either similar to known motifs, or independently identified in other samples, or similar to known RPG-regulating motifs.

Table 3-9: Motif discovery in the putative regulatory regions of RPGs.

Data Source	Sample	# predicted motifs (random)	%motifs similar to JASPAR motif	%motifs similar to motifs in other samples	%known RPG-regulating motifs identified	%motifs supported
Rao	GM12878	1118 (64)	38.28%	99.55%	71.43%	99.55%
	IMR90	371 (16)	42.05%	100.0%	57.14%	100.0%
	HMEC	103 (2)	41.75%	100.0%	35.71%	100.0%
	KBM7	149 (3)	54.36%	100.0%	35.71%	100.0%
	HUVEC	68 (8)	50.0%	98.53%	28.57%	98.53%
	NHEK	189 (12)	41.8%	100.0%	42.86%	100.0%
	K562	362 (14)	46.96%	100.0%	50%	100.0%
Javierre	nB	487 (23)	50.51%	99.59%	64.29%	99.59%
	tCD8	528 (26)	45.83%	99.81%	71.43%	99.81%
	FoeT	552 (48)	46.56%	100.0%	57.14%	100.0%
	tCD4	607 (12)	46.46%	99.67%	71.43%	99.67%

Despite of the existence of different motifs in different samples, we were able to identify 48 motifs shared by at least four samples between Rao et al.'s data and Javierre et al's data, including the CTCF motif. We identified 99 motifs shared by at least four samples from Rao et al. and 131 motifs by the four samples from Javierre et al. Interestingly, 48 motifs were shared by the 99 motifs from Rao et al. and the 131 motifs from Javierre et al, demonstrating that there were common regulatory mechanisms among RPGs in spite of the different putative regulatory regions and regulatory motifs. Among these 48 motifs, 24 of them were known motifs and 11 of them were known to regulate RPGs.

3.1.4 Discussion

We studied the putative regulatory regions of human RPGs in eleven samples. We identified about 22797 regions that directly or indirectly interacted with RPG promoters, the majority of which were distal regions. There were a large fraction of regulatory regions that were different in different samples. Interestingly, about 1% to 91% direct regions in a sample were often identified to interact with RPG promoter directly in other samples. Moreover, different RPGs may share common regulatory regions and form a co-regulated gene group. Such co-regulated gene groups were conserved across samples. In addition, in different samples, there were common regulatory motifs identified. All these observations may explain why human RPGs are coordinately regulated even though they have different regulatory regions and are regulated differently across samples.

We identified 16588 regulatory regions that likely regulate RPGs from Rao et al.'s data. However, this number may be over-estimated, given the much higher sequencing depth in GM12878 and the imperfect cutoff 30 to define chromatin interaction from the normalized Hi-C contact matrices in GM12878. With this said, it is no doubt that there should be thousands of distal regions that may regulate RPGs. In fact, if we considered the other six samples from Rao et al., there were 9210 different distal regions identified. If we considered the four samples from Javierre et al., there were 6140 different distal regions. Note that the Javierre et al.'s interaction data were defined by the original study [48]. Since we only considered a handful of samples, there may be even more distal regions, given the fact that the majority of regions identified in a sample were not identified in a new sample.

Previously, we identified 217 RPG regulatory regions based on DHS data in 349 samples [33]. Compared with the regions identified here, 95.9% of the 217 RPG regulatory regions were identified in the seven samples from Rao et al., while only 1.9% of the regions identified in these seven samples here were also identified by the previous study. Similarly, 74.8% of the 217 regions were identified in the four samples from Javierre et al. that accounted for about 3.4% of the identified regions in Javierre et al.'s samples. These numbers suggested that the previously identified regions were limited by considering regions shared by the majority samples. It also implied that RPGs are likely regulated differently in different samples.

Although the identified RPG regulatory regions here physically interact with RPG promoters in the corresponding samples, they were still putative RPG regulatory regions. This was because we did not know whether these direct or indirect interactions changed the RPG expression levels. Future studies may explore in this direction to define more accurate RPG regulatory regions. With this said, these regions represented our current understanding of RPG distal transcriptional regulation. Moreover, these regions shed new light on our understanding of the coordinated regulation of human RPGs.

We noticed that 77 of the 80 human RPGs were in a loose cluster in GM12878 (Table 3-2). Because of the much larger sequencing depth in GM12878, we are not sure whether this is true in other samples, if the sequencing depth in other samples is increased. It will be valuable to test this in the future. If it is true, this cluster may significantly contribute to RPG coordinated regulation. Even if it is not true, it is clear that there are several dozen RPGs in different samples sharing regulatory regions, which facilitates their coordinated activities. It is worth pointing out that the pairs of RPGs sharing regulatory regions in a sample were also observed in a different sample, suggesting that such a sharing mechanism is conserved.

We identified different numbers of motifs across samples. This is not surprising, since the number of regulatory regions is quite different across samples. However, we noticed that there are about a dozen motifs shared by different samples from different studies, suggesting that these shared motifs may indeed RPG-specific and they may contribute to the RPG coordinated regulation as well. It is worth investigating whether these shared motifs, especially the novel ones, are bona fide RPG-regulating motifs.

We noticed a surprising difference between the Hi-C data from Rao et al. and the promoter capture Hi-C data from Javierre et al. The sequencing depth was slightly larger in samples from Javierre et al. than those from Rao et al. except in GM12878. There were indeed more regions identified in the corresponding samples from Javierre et al. Surprisingly, there were slightly fewer RPGs with identified regions from Javierre et al. than those from Rao et al. We are not sure this is because the promoter capture Hi-C is biased, there is something different among the samples in the two studies, or something else.

CHAPTER 4: STRAIN GENOME RECONSTRUCTION

4.1 SMS: a novel approach for microbial strain genome reconstruction in multiple samples

4.1.1 Introduction

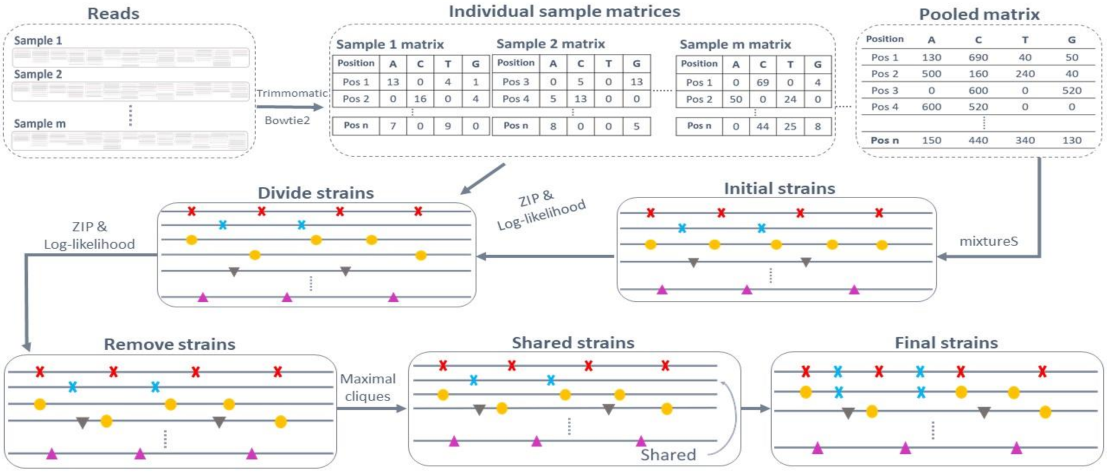
Bacteria are ubiquitous and play crucial roles in human health [131-133]. Multiple strains of a bacterial species usually coexist in an environmental niche. These strain genomes of the same species are different from each other, with small variations such as single nucleotide polymorphisms (SNPs), different gene contents, and/or different plasmid genes [134]. Such a difference results in different fitness to survive or react to stimuli, which is often the cause of drug resistance, mixed infection, etc. [135, 136]. It is thus important to study bacterial strains and reconstruct their genomes.

Dozens of computational methods are available to infer bacterial strains from shotgun metagenomic reads [56-58, 60-62, 137-146]. Most of them rely on prior knowledge of known strains. These methods have successfully identified known strains while cannot be applied to study new strains that commonly exist. A handful of methods that do not depend on known strains are thus developed, which can be divided into two groups [61, 62, 140, 142, 143, 147]. One group defines strain variations and strains based on species-specific marker genes, which can significantly speed up the process to analyze a large number of species in a microbiome while depending on the quality and quantity of the marker genes [62, 147]. The other group considers the SNPs across the entire reference genomes of a species instead of only the marker gene regions, which can delineate the strain genomes in detail and are important for the study of individual pathogen species [60, 61, 139, 141]. These methods have shed new light on bacterial strains in environmental samples. However, their performance is still suboptimal in terms of the predicted strain number and abundance. For instance, a recent method, StrainFinder, did not have good accuracy in predicting strain SNPs and strain abundance, even provided with the correct strain number [60].

To accurately identify strains in shotgun metagenomic samples, we developed a novel method called SMS (Strains in Multiple Samples). Starting from a species genome, SMS de novo reconstructs its strain genomes from shtogun reads in multiple samples. It models the coverage of every strain in individual samples by zero-inflated Poisson (ZIP) distributions and classifies SNPs with adatively inferred centers, which enables it to identify low-coverage strains and predict strains with high accuracy. Tested on 702 simulated and 195 experimental datasets, SMS accurately predicted the strain number, abundance, and SNPs. Compared with two recent approaches, SMS showed much better performance.

4.1.2 Methods

SMS reconstructs bacterial strain genomes with a reference genome and raw reads in multiple shotgun metagenomic samples (Figure 4-1). The basic assumption is that different SNPs from the same strain follow a common ZIP distribution in a sample, and SNPs from different strains follow different ZIP distributions in individual samples. Assume there are R strains of a species of interest in m samples. Starting from the cleaned raw reads, SMS defines SNPs based on the reads mapped to the reference. It then determines the initial strains and their abundance with the pooled sample, the combined *m* samples. Next, SMS refines the initial strains and their abundance based on the SNP coverage patterns across samples. The rationale is that unique SNPs from the same strain will have more similar coverage patterns across samples than SNPs from different strains. Finally, SMS outputs the predicted strains and their abundance. The details are in the following sections.



Sample 1 matrix

Position	A	C	T	G
Pos 1	13	0	4	1
Pos 2	0	16	0	4
⋮				
Pos n	7	0	9	0

Sample 2 matrix

Position	A	C	T	G
Pos 3	0	5	0	13
Pos 4	5	13	0	0
⋮				
Pos n	8	0	0	5

Sample m matrix

Position	A	C	T	G
Pos 1	0	69	0	4
Pos 2	50	0	24	0
⋮				
Pos n	0	44	25	8

Pooled matrix

Position	A	C	T	G
Pos 1	130	690	40	50
Pos 2	500	160	240	40
Pos 3	0	600	0	520
Pos 4	600	520	0	0
⋮				
Pos n	150	440	340	130

Figure 4-1: The SMS workflow.

4.1.2.1 Identification of potential SNPs

With reads from *m* samples, SMS trims reads and filters low-quality reads with the tool trimmomatic. SMS then maps the cleaned reads to the reference genome by bowtie2 [148]. In every sample, SMS obtains a 4 by *n* sample-specific matrix composed of the frequencies of A, C, G, and T in the mapped reads at each of the *n* reference genome positions. Similarly, SMS acquires a pooled matrix of 4 by *n* for the pooled sample, the sum of the *m* sample-specific matrices. SMS then determines the n' potential polymorphic positions based on these *m*+1 matrices. A reference genome position is potentially polymorphic if the following criteria are satisfied: 1). It has a coverage larger than 10% of the pooled coverage. The coverage of a genome (position or SNP) is calculated as the average number of reads mapped to this genome (position or SNP); 2). It has at least two nucleotides with coverage no smaller than 5% of the pooled coverage. Note that when the reference nucleotide at a position has fewer than 5% of the pooled coverage, the reference nucleotide is replaced with the most frequent nucleotide at this position; 3). Each of its two most frequent nucleotides must occur in at least 5% of the m samples. Finally, SMS considers all *n1* nucleotides with coverage larger than 5% of the genome coverage at these positions as potential SNPs, where $n' \leq n1 \leq 3n'$.

4.1.2.2 Prediction of the strain number and abundance

With the *n1* potential SNPs, SMS infers the strain number and abundance in four steps.

First, SMS obtains an initial number of strains and their SNPs. SMS applies mixtureS to the above *n1* SNPs with the pooled sample and outputs the predicted strains and their abundance. MixtureS reconstructs the strain genomes from shotgun reads in one sample and has shown good performance previously [60]. In this way, the strains with different pooled coverage are separated into R strains. R is automatically inferred.

Second, SMS refines the predicted strains so that almost all SNPs in an actual strain are assigned to one predicted strain. Since the coverage of SNPs from the same strain are expected to follow the same ZIP distributions in individual samples, the coverage vectors of two SNPs from the same strain are more similar than those of two SNPs from different strains.

Here the coverage vector of a SNP is a vector composed of its coverage in the *m* samples. The similarity measurement of two vectors is described in the next section. Based on this observation, SMS iteratively regroups the *n1* SNPs into *R* groups so that SNPs from the same group have more similar vectors. Starting from the predicted *R* strains by mixtureS, the majority of SNPs in each of which are likely from the same strain, SMS represents each strain by a *m* by 1 coverage vector, the average of the coverage vectors of the SNPs currently assigned to this strain. SMS then re-assigns each of the *n1* SNPs to the strain with the most similar coverage vector to the coverage vector of this SNP. With the re-assigned SNPs, the coverage vectors of the strains are recalculated. This process is repeated a given number of times or until the assigned SNPs to each strain do not change. In this way, the coverage vector of each predicted strain and the assignment of the *n1* SNPs become more and more accurate, with almost all SNPs from an actual strain grouped together.

Third, SMS investigates whether there are more or fewer than *R* strains. SMS divides each strain into two strains, one strain at a time. To determine whether a strain should be divided, SMS models each strain in a sample by a ZIP distribution, estimates the parameters of the ZIP distributions, and calculates the likelihood ratio of observing the SNPs in this strain across the *m* samples to that in two divided strains. The details of the ZIP parameter estimation and the likelihood testing are in the following sections. A strain is divided only when its division significantly increases the likelihood (Chi-square test p-value<0.001). If a strain is divided, SMS considers whether the two new divided strains can be further divided similarly. This process is repeated until no strain can be further divided. With all possible divisions that significantly increase the likelihood, SMS obtains the updated *R* strains and repeats Step two to reassign the *n1* SNPs to these *R* strains again. SMS then considers removing each strain, one strain at a time. The process is similar to dividing a strain based on the ZIP parameter estimation and the likelihood test.

Finally, SMS removes the predicted strains that are majorly composed of shared SNPs by multiple strains and reassigns their SNPs to the corresponding strains. To remove a strain, SMS identifies its consistent strains. Strain one is a consistent strain of strain two if every entry in the coverage vector of strain one is no large than the corresponding entry in the coverage vector of strain two plus a small cutoff. Similarly, multiple strains together are consistent with strain two if the sum of the corresponding entries in their coverage vectors is no large than the corresponding entry in the coverage vector of strain two plus the same cutoff. With the consistent strains of a strain, SMS constructs a graph, with each consistent strain as a node and edges connecting pairs of strains that are together still consistent with this strain. SMS then identifies the largest cliques in this graph with the corresponding groups of strains together consistent with this strain. With a clique identified, SMS removes this strain and reassigns its SNPs to all consistent strains in this clique. In this way, SMS finalizes the predicted strains and their SNPs. The abundance of every strain is calculated as the average coverage of the SNPs unique to this strain.

4.1.2.3 The similarity of two coverage vectors

SMS calculates the similarity of two coverage vectors $(a_1, a_2, \dots, a_m)$ and $(b_1, b_2, \dots, b_m)$ by a pre-defined regression formula: 79.25*d*+ 43.06(*c*+*c*^{3})-0.04/(0.0025+*d*), where *d* is the distance between the two vectors, and *c* is their Kendall rank correlation. This formula was constructed based on a set of 18 pre-simulated training datasets. SMS chooses this similarity measurement, because it shows better performance than others, including correlation, Euclendian distance, relative entropy, etc.

4.1.2.4 ZIP model of a strain in a sample

SMS models the coverage of the SNPs from the *p-th* strain in the *q-th* sample by a ZIP distribution $ZIP(x, \pi_{pq}, \lambda_{pq})$ when the *p-th* strain occurs in the *q-th* sample, where

$$ZIP(x,\pi,\lambda) = \begin{cases} \pi + (1-\pi) * \exp(-\lambda), for\ x = 0 \\ \dfrac{(1-\pi) * \lambda^x}{x!} * \exp(-\lambda), for\ x = 1, 2, 3, \dots \end{cases} \quad (1)$$

Assume we have an *n1* by *m* matrix, $X = (x_{ij})$, which store the coverage of the above *n1* SNPs in the *m* samples. Assume $Z = (z_{ir})$ is the indicator to tell whether the *i-th* SNP belongs to the *r-th* strain, where $\sum_{r=1}^{R} z_{ir} = 1$ for all *i* from 1 to *n1* and z_{ir} can be only 0 or 1. Assume Y= (y_{jr}) is the indicator to show whether the *r-th* strain occurs in the *j-th* sample, where y_{jr} can be only 0 or 1. If at least one SNP from a strain has a non-zero coverage in a sample, we tentatively claim that this strain occurs in this sample. When $y_{jr} = 1$, we also define $b_{jr} = \sum_{i=1}^{n1} z_{ir} I_{x_{ij}=0}$, $n_{jr} = \sum_{i=1}^{n1} z_{ir}$, and $a_{jr} = \sum_{i=1}^{n1} z_{ir} x_{ij} / n_{jr}$.

To estimate the parameters in the ZIP, for a given strain that occurs in a given sample, say the *r-th* strain in the *j-th* sample (i.e., y_{jr}=1), SMS initializes $\lambda_{jr} = \frac{s_{jr}^2 + a_{jr}^2}{a_{jr}} - 1$, $\pi_{jr} = \frac{s_{jr}^2 - a_{jr}}{s_{jr}^2 + a_{jr}^2 - a_{jr}}$, with $s_{jr}^2 = \frac{\sum_{i=1}^{n1} z_{ir}(x_{ij} - a_{jr})^2}{\sum_{i=1}^{n1} z_{ir} - 1}$. SMS then uses the following iteration method to obtain the maximal likelihood estimation of π_{jr} and λ_{jr}: first replaces π_{jr} by $\pi_{jr} = \frac{n_{jr}(\lambda_{jr} - a_{jr})e^{-\lambda_{jr}}}{\lambda_{jr} b_{jr} - n_{jr}(\lambda_{jr} - a_{jr})\left(1 - e^{-\lambda_{jr}}\right)}$ in the equation $\frac{n_{jr} a_{jr}}{\lambda_{jr}} - \frac{(1 - \pi_{jr}) b_{jr}}{\pi_{jr} + (1 - \pi_{jr}) e^{-\lambda_{jr}}} = 0$ to obtain an equation of λ_{jr}, then solves this equation by the Newton's iteration method. Everywhere in this process, if π_{jr}=0, you will directly estimate λ_{jr}=a_{jr}.

4.1.2.5 Log likelihood test

Given *R* strains, the full likelihood of observation the frequencies of these *n1* SNPs in the *m* samples is

$$L(X, Z|\pi, \lambda) = \prod_{i=1}^{n1} \prod_{j=1}^{m} (\sum_{r=1}^{R} z_{ir} y_{jr} ZIP(x_{ij}, \pi_{jr}, \lambda_{jr})). \qquad (2)$$

When SMS splits one strain into two or removes one strain, the likelihood can be similarly calculated. To assess the significance of changing the current *R* strains, we calculate the ratio of the likelihood after changing (split or remove) to the likelihood before changing. The ratio approximately follows a Chi-square distribution with the degree of freedom equal to the difference of the parameters in the two models. If the Chi-square test p-value is smaller than a pre-defined cutoff, SMS correspondingly modifies the current *R* strains.

4.1.2.6 Simulated and experimental datasets

We simulated 702 datasets (Table 4-1). In each dataset, a reference genome was chosen, 2 to 4 strains were simulated, and 5 to 35 samples were generated. For each reference genome, their four strains were generated by randomly choosing 0.01% of the genome positions and then randomly substituted the reference nucleotide with another nucleotide. The read coverage of a reference genome in a dataset was one of the following five coverage, 50x, 100x, 150x, 200x, and 300x. The number of strains and their relative abundance in a dataset were specified by one of the following five configurations: 10:20:30:40, 10:25:25:40, 10:30:60, 15:30:55, and 30:70. For a dataset, with the chosen configuration and the number of samples, a subset of samples were randomly chosen for each strain and the coverage of this strain in one of the samples was then randomly determined so that the pooled coverage of this strain was the same as what was specified in the configuration. With the coverage of strains in a sample, paired reads of 100 base pairs long were randomly generated using dwgsim (https://github.com/nh13/DWGSIM).

Table 4-1: Simulated datasets.

Types	Strain Configuration	#strains	coverage
Type1	10:20:30:40	4	100x
	10:30:60	3	100x
	15:30:55	3	100x
	30:70	2	100x
	10:25:25:40	4	100x
	10:25:25:40	4	150x
	10:25:25:40	4	200x
	10:25:25:40	4	300x
Type2	10:20:30:40	4	100x
	10:25:25:40	4	100x
	10:25:25:40	4	150x
	10:25:25:40	4	200x
	10:25:25:40	4	300x
Type3	10:20:30:40	4	100x
	10:25:25:40	4	100x
	10:25:25:40	4	150x
	10:25:25:40	4	200x
	10:25:25:40	4	300x
Type4	10:30:60	3	100x
	15:30:55	3	100x
	10:30:60	3	150x
	10:30:60	3	200x
	10:30:60	3	300x
	15:30:55	3	150x
	15:30:55	3	200x
	15:30:55	3	300x

There are 27 datasets generated in each row, each of which corresponds to one of the three different bacterial species (species 1: *Bartonella clarridgeiae* NC_014932, species 2: *Enterococcus casseli flavus* NC_020995, and species 3: *Methanobrevibacter smithii* NC_009515) and one of the nine different sample numbers (5, 8, 10, 12, 15, 20, 25, 30, 35). The number and order of strains in each dataset are specified by the strain configuration. For instance, the configuration 10:20:30:40 in the first row specifies that there are four strains, with the coverage of the 1st, 2nd, 3rd and 4th strain as 10X, 20X, 30X and 40X, respectively. There is no shared SNPs among strains in Type 1 datasets. In each type 2 dataset, the first two strains share 30% of their SNPs, the first three strains share 20% of their SNPs, and the fourth strain shares no SNP with other strains. In each type 3 dataset, the first two strains share 30% of their SNPs, the first three strains share 10% of their SNPs, and there is no SNP shared between the first three strains and the last strain. In each type 4 dataset, the first two strains share 30% of their SNPs, and the third strain share no SNP with other strains.

We tested SMS on 195 experimental datasets [136]. Each dataset is known to have two *Mycobacterium tuberculosis* strains with predicted abundance. The abundance is inferred from two different computational methods. The actual SNPs in each strain are unknown.

4.1.2.7 Comparison with existing methods

We compared SMS with mixtureS and StrainFinder in a desktop computer with the Intel Core i9-9900KF CPU (16 cores@3.6GHz) and 32 gigabytes memory. We used the following commands to run the three tools respectively:

```
SMS: python SMS/running.py --output_name %s  --genome_len %s --genome_name %s --genome_file_loc %s --bam_loc_file %s --res_dir %s
```

```
MixtureS: python mixtureS/mixture_model.py --sample_name %s  --genome_len %s --genome_name %s --genome_file_loc %s --bam_file %s --res_dir %s
```

```
StrainFinder: python StrainFinder/StrainFinder.py --aln %s -N %s --max_reps 10 --dtol 1 --ntol 2 --max_time 3600 --coverage --em_out %s --out_out %s --log %s --n_keep %s --force_update --merge_out –msg
```

4.1.3 Results

4.1.3.1 SMS correctly predicted the strain numbers

We studied the number of strains predicted in 702 simulated datasets (Table 4-1). There were 5 to 35 samples and 2 to 4 strains in every dataset, with the pooled coverage of strains from 100X to 300X. The pooled coverage was the sum of the coverage of all strains of a species in all samples. The number of strains and their relative abundance are specified by one of the following five configurations in each dataset: 10:20:30:40, 10:25:25:40; 10:30:60, 15:30:55, and 30:70. For instance, for a dataset with the configuration 10:20:30:40, the proportion of reads from the four strains was 10%, 20%, 30% and 40%, respectively.

Overall, SMS predicted the correct strain numbers in all but five datasets. Interestingly, SMS did not predict the correct strain number in at least one dataset for each of the three randomly selected species, implying that its performance was not species-specific. In each of the five datasets, a pair of strains shared 30% of their SNPs. In four of the five datasets, three strains were sharing 20% of their SNPs. These shared SNPs may have confused SMS when the coverage was 100X. As expected, when the coverage was increased, SMS predicted the correct strain number in each of the five corresponding datasets. These analyses suggested that SMS can accurately predict the strain number, even when the pooled coverage was 100X and there were only five samples in a dataset. Moreover, the predicted strain number was even more accurate with a larger pooled coverage (200X coverage for perfect prediction here).

4.1.3.2 SMS reliably estimated the strain abundance

We investigated how well SMS predicted the strain abundance. No matter whether the strain number was correctly predicted, the predicted strain abundance agreed well with the known strain abundance (Figure 4-2). This agreement did not depend on the sample number, the pooled coverage, the strain number, etc.

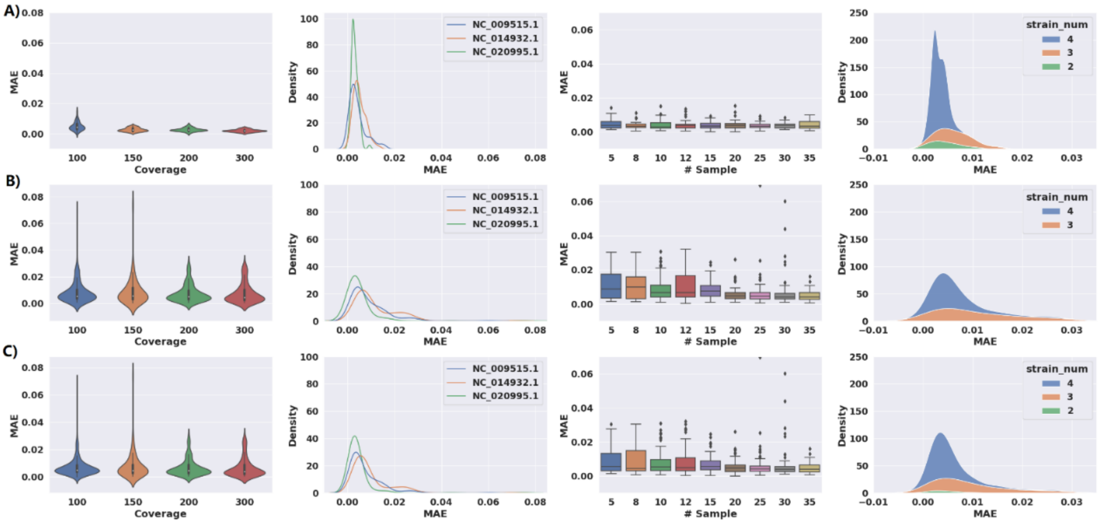


Figure 4-2: The predicted strain abundance. A) Unshared datasets; B) Shared datasets; and C) All datasets. MAE is the average Maximal Absolute Difference of the predicted abundance and the corresponding true abundance across datasets.

In the 697 datasets SMS correctly predicted the strain number, the predicted strain abundance was within 97.31% of the true abundance. The mean and median ratio of the predicted abundance to the true abundance were 0.99 and 1.00, respectively. Even in the five datasets with the incorrectly predicted strain number, the predicted strain abundance was similar to the true abundance. For instance, SMS predicted four strains in three datasets with three strains. In two datasets, two strains had the predicted abundance about 0.08 and 0.29, respectively, which were close to the corresponding true abundance 0.10 and 0.30. The two remaining predicted abundance were about 0.42 and 0.21, which differed from the third true abundance, 0.60. In the third dataset, one strain was predicted with an abundance of 0.31, close to the true abundance of 0.30. The wrong prediction of the strain number and strain abundance was likely due to the third strain's uneven and relatively limited coverage. After increasing the coverage, SMS predicted the correct strain number and more similar abundance.

The accuracy was in general improved with more samples and a larger pooled coverage in a dataset (Figure 4-2). For instance, when the sample number was larger, the median of the predicted abundance was closer to the true abundance, and the variation of the maximal absolute difference (MAE) between the predicted abundance and the true abundance was smaller. The accuracy was not affected much by different species or the number of strains in a dataset (Figure 4-2). For instance, the MAE was within a similar range and with a similar mean/median when there were different numbers of strains. The small variations suggested that the predicted abundance by SMS was robust to different bacterial genomes, different number of strains, etc.

4.1.3.3 SMS faithfully determined the SNPs

Existing methods mainly focus on the predicted strain number and only occasionally consider their abundance. Rarely do they mention the accuracy of the predicted strain SNPs. With the simulated datasets, we systematically evaluated the predicted SNPs. We found that SMS has a precision of 0.97 and a recall of 0.96 to predict strain SNPs.

We studied the datasets without shared SNPs among strains. In all 216 datasets, on average, SMS had a precision of 0.98 and a recall of 0.98. For a given species with a specified pooled coverage, the precision and recall were higher on datasets with more samples in general. Similarly, they were generally higher on datasets with a larger pooled coverage when the species and the sample number were fixed. For instance, for the reference species genome NC_009515.1 and the sample number 20, the precison increased from 0.98 to 0.99 and the recall increased from 0.97 to 0.99 when the pooled coverage increased from 100X to 300X.

We also studied the predicted strains on datasets with shared SNPs among strains. We again focused on the two most challenging configurations: 10:20:30:40 and 10:25:25:40. They were challenging because the shared SNPs among strains may have similar coverage across samples with SNPs unique to other strains. For instance, the shared SNPs between the first two strains in the configuration 10:20:30:40 had a relative abundance of 30%, the same as the relative abundance of the third strain. Even with such complexity, SMS on average had a precision of 0.97 and a recall of 0.96 on all datasets (Supplementary Tables S7 and S8). The performance suggested that SMS could reconstruct the complicated evolutionary trajectories of strains with shotgun sequencing reads.

4.1.3.4 SMS performed well on experimental datasets

We tested SMS on 195 experimental datasets (Table 4-2). We chose these datasets because their strain numbers were known. The strain abundance was also predicted previously [136]. Note that the datasets from the Critical Assessment of Metagenome Interpretation challenge did not provide the strain number, strain abundance and SNPs unique to strains, thus not suitable for the strain genome reconstruction here [149].

Table 4-2: 195 TB abundance result from three tools.

sample	# of verified method in refernce paper	major strain proportion	MixtureS	SMS	StrainFinder
ERR036194	single	1	2(0.349)	2(0.071)	2(0.459)
ERR036233	both	0.72	4(0.113)	2(0.124)	2(0.215)
ERR036248	both	0.88	2(0.01)	2(0.045)	2(0.114)
ERR037469	both	0.63	2(0.144)	2(0.001)	2(0.115)
ERR037547	both	0.85	2(0.042)	2(0.025)	2(0.035)
ERR126641	both	0.84	4(0.137)	2(0.025)	2(0.063)
ERR126642	both	0.8	2(0.088)	2(0.161)	2(0.055)
ERR161024	both	0.86	3(0.082)	2(0.033)	2(0.003)
ERR161026	both	0.85	3(0.106)	2(0.182)	2(0.066)
ERR161027	both	0.82	3(0.06)	2(0.081)	2(0.046)
ERR161034	both	0.65	3(0.086)	2(0.332)	2(0.067)
ERR161039	both	0.63	4(0.157)	2(0.352)	2(0.084)
ERR161049	both	0.87	3(0.099)	2(0.209)	2(0.246)
ERR161050	both	0.73	5(0.218)	2(0.214)	2(0.214)
ERR161055	both	0.89	2(0.005)	2(0.022)	2(0.387)
ERR161071	both	0.84	4(0.148)	2(0.318)	2(0.008)
ERR161077	both	0.78	5(0.238)	2(0.116)	2(0.206)
ERR161078	both	0.58	4(0.123)	2(0.054)	2(0.038)
ERR161081	both	0.88	2(0.124)	2(0.105)	2(0.317)
ERR161084	both	0.87	3(0.103)	2(0.349)	2(0.047)
ERR161088	both	0.88	3(0.11)	2(0.175)	1(0.12)
ERR161090	both	0.91	2(0.039)	2(0.062)	2(0.406)
ERR161091	both	0.89	2(0.003)	2(0.08)	2(0.336)
ERR161097	both	0.85	2(0.025)	2(0.131)	2(0.057)
ERR161120	both	0.86	2(0.104)	2(0.128)	2(0.177)
ERR161122	both	0.87	3(0.099)	2(0.187)	2(0.239)
ERR161123	both	0.81	3(0.056)	2(0.166)	2(0.042)
ERR161170	both	0.89	2(0.228)	2(0.094)	2(0.352)
ERR161173	both	0.9	2(0.265)	2(0.086)	2(0.378)
ERR161176	both	0.88	3(0.1)	2(0.101)	2(0.375)
ERR161184	both	0.87	3(0.196)	2(0.114)	2(0.349)

sample	# of verified method in refernce paper	major strain proportion	MixtureS	SMS	StrainFinder
ERR161194	both	0.86	3(0.101)	2(0.07)	2(0.121)
ERR161195	both	0.88	2(0.01)	2(0.104)	2(0.377)
ERR163932	both	0.91	3(0.097)	2(0.201)	2(0.401)
ERR163940	both	0.87	2(0.04)	2(0.109)	2(0.364)
ERR163942	both	0.87	3(0.094)	2(0.078)	2(0.144)
ERR163943	both	0.83	3(0.087)	2(0.011)	2(0.101)
ERR163947	both	0.5	3(0.122)	2(0.211)	2(0.04)
ERR163954	both	0.92	2(0.155)	2(0.208)	2(0.417)
ERR163971	both	0.89	2(0.007)	2(0.021)	2(0.388)
ERR163986	both	0.9	2(0.018)	2(0.398)	2(0.394)
ERR163996	both	0.88	3(0.098)	2(0.042)	2(0.377)
ERR164007	both	0.88	2(0.01)	2(0.334)	2(0.167)
ERR164021	both	0.7	3(0.043)	2(0.001)	2(0.187)
ERR176446	both	0.88	4(0.152)	2(0.1)	2(0.379)
ERR176458	both	0.82	4(0.16)	2(0.29)	2(0.316)
ERR176460	both	0.88	2(0.001)	2(0.27)	2(0.373)
ERR176461	both	0.8	2(0.06)	2(0.189)	2(0.298)
ERR176514	single	1	2(0.282)	2(0.048)	2(0.48)
ERR176521	both	0.89	2(0.132)	2(0.182)	2(0.385)
ERR176533	both	0.9	2(0.001)	2(0.036)	2(0.275)
ERR176549	both	0.72	2(0.053)	2(0.071)	2(0.213)
ERR176556	both	0.86	2(0.041)	2(0.032)	2(0.113)
ERR176557	both	0.86	2(0.058)	2(0.081)	2(0.013)
ERR176600	both	0.89	2(0.014)	2(0.096)	2(0.386)
ERR176604	both	0.89	3(0.111)	2(0.384)	2(0.369)
ERR176610	both	0.9	2(0.026)	2(0.089)	2(0.397)
ERR176611	both	0.88	3(0.1)	2(0.106)	2(0.376)
ERR176616	both	0.63	5(0.191)	2(0.333)	2(0.125)
ERR176620	both	0.54	3(0.096)	2(0.378)	2(0.015)
ERR176621	both	0.9	2(0.002)	2(0.039)	2(0.273)
ERR176631	both	0.89	3(0.099)	2(0.089)	2(0.388)
ERR176650	both	0.87	4(0.152)	2(0.225)	2(0.363)
ERR176652	both	0.82	5(0.245)	2(0.097)	2(0.304)
ERR176653	both	0.8	4(0.178)	2(0.154)	2(0.29)
ERR176655	both	0.91	2(0.208)	2(0.251)	2(0.403)
ERR176664	both	0.88	3(0.229)	2(0.166)	2(0.378)
ERR176668	both	0.92	2(0.276)	2(0.038)	2(0.395)
ERR176672	both	0.89	2(0.237)	2(0.201)	2(0.371)

sample	# of verified method in refernce paper	major strain proportion	MixtureS	SMS	StrainFinder
ERR176681	both	0.89	2(0.187)	2(0.068)	2(0.388)
ERR176688	both	0.9	2(0.253)	2(0.077)	2(0.381)
ERR176701	both	0.83	4(0.165)	2(0.159)	2(0.327)
ERR176703	single	1	2(0.348)	2(0.494)	2(0.493)
ERR176706	both	0.89	2(0.178)	2(0.073)	2(0.362)
ERR176709	both	0.65	5(0.171)	2(0.042)	2(0.142)
ERR176713	both	0.89	4(0.194)	2(0.265)	2(0.345)
ERR176723	both	0.88	3(0.103)	2(0.043)	2(0.361)
ERR176725	both	0.88	4(0.152)	2(0.181)	2(0.37)
ERR176734	both	0.91	2(0.017)	2(0.063)	2(0.375)
ERR176738	both	0.88	4(0.193)	2(0.184)	2(0.379)
ERR176746	both	0.87	2(0.176)	2(0.053)	2(0.359)
ERR176748	both	0.87	2(0.031)	2(0.119)	2(0.09)
ERR176749	both	0.81	2(0.035)	2(0.175)	2(0.023)
ERR176755	both	0.89	2(0.179)	2(0.055)	2(0.375)
ERR176785	single	1	2(0.32)	2(0.172)	2(0.479)
ERR176793	both	0.54	5(0.198)	2(0.146)	2(0.004)
ERR176796	both	0.88	2(0.118)	2(0.072)	2(0.374)
ERR176802	both	0.89	2(0.217)	2(0.187)	2(0.347)
ERR176807	both	0.87	3(0.098)	2(0.164)	2(0.367)
ERR176809	both	0.89	2(0.202)	2(0.064)	2(0.387)
ERR176810	both	0.91	2(0.214)	2(0.072)	2(0.409)
ERR176813	both	0.88	3(0.106)	2(0.364)	2(0.377)
ERR181686	both	0.87	3(0.094)	2(0.183)	2(0.079)
ERR181688	both	0.83	2(0.041)	2(0.153)	2(0.249)
ERR181689	both	0.85	2(0.044)	2(0.148)	2(0.069)
ERR181695	both	0.87	2(0.013)	2(0.277)	2(0.368)
ERR181705	both	0.84	2(0.044)	2(0.231)	2(0.028)
ERR181708	both	0.87	3(0.092)	2(0.027)	2(0.321)
ERR181749	both	0.87	2(0.209)	2(0.352)	2(0.285)
ERR181750	both	0.89	2(0.078)	2(0.045)	2(0.359)
ERR181752	both	0.85	2(0.007)	2(0.116)	2(0.317)
ERR181753	both	0.87	2(0.004)	2(0.363)	1(0.13)
ERR181782	both	0.86	3(0.085)	2(0.114)	2(0.042)
ERR181784	both	0.82	3(0.055)	2(0.131)	2(0.047)
ERR181785	both	0.8	2(0.086)	2(0.267)	2(0.027)
ERR181810	both	0.9	2(0.23)	2(0.007)	2(0.375)
ERR181811	both	0.54	5(0.191)	2(0.413)	2(0.0)

sample	# of verified method in refernce paper	major strain proportion	MixtureS	SMS	StrainFinder
ERR181813	both	0.62	4(0.121)	2(0.072)	2(0.084)
ERR181827	both	0.88	3(0.096)	2(0.2)	2(0.374)
ERR181828	both	0.88	3(0.092)	2(0.194)	2(0.319)
ERR181838	both	0.88	2(0.004)	2(0.108)	2(0.378)
ERR181845	both	0.9	2(0.035)	2(0.051)	2(0.397)
ERR181849	both	0.88	3(0.095)	2(0.111)	2(0.361)
ERR181866	both	0.89	3(0.104)	2(0.101)	1(0.11)
ERR181870	both	0.9	2(0.024)	2(0.085)	2(0.398)
ERR181876	both	0.91	2(0.1913)	2(0.078)	2(0.2058)
ERR181878	both	0.81	2(0.101)	2(0.178)	2(0.293)
ERR181880	both	0.86	2(0.011)	2(0.124)	2(0.341)
ERR181881	both	0.84	4(0.146)	2(0.292)	2(0.129)
ERR181909	both	0.89	2(0.037)	2(0.092)	2(0.378)
ERR181913	both	0.89	2(0.179)	2(0.097)	2(0.362)
ERR181923	both	0.88	2(0.015)	2(0.073)	1(0.12)
ERR181933	both	0.91	2(0.043)	2(0.003)	2(0.409)
ERR181937	both	0.87	2(0.012)	2(0.059)	2(0.349)
ERR181945	both	0.87	4(0.145)	2(0.067)	2(0.364)
ERR181953	both	0.87	2(0.013)	2(0.092)	2(0.368)
ERR181974	both	0.85	3(0.094)	2(0.146)	2(0.219)
ERR181977	both	0.8	4(0.143)	2(0.096)	2(0.056)
ERR181983	both	0.9	3(0.094)	2(0.399)	2(0.373)
ERR182015	both	0.85	4(0.143)	2(0.342)	2(0.176)
ERR182026	both	0.84	4(0.151)	2(0.01)	2(0.064)
ERR182027	both	0.87	4(0.138)	2(0.208)	2(0.167)
ERR182041	both	0.88	2(0.011)	2(0.065)	2(0.368)
ERR182049	both	0.89	3(0.098)	2(0.091)	2(0.366)
ERR190340	both	0.63	5(0.225)	2(0.356)	2(0.127)
ERR190342	both	0.86	3(0.09)	2(0.185)	2(0.165)
ERR190343	both	0.8	3(0.045)	2(0.215)	2(0.029)
ERR190379	both	0.77	4(0.111)	2(0.239)	2(0.2)
ERR190388	both	0.91	3(0.268)	2(0.319)	1(0.09)
ERR211990	both	0.89	3(0.101)	2(0.084)	2(0.343)
ERR212002	both	0.86	3(0.106)	2(0.065)	2(0.323)
ERR212004	both	0.86	4(0.157)	2(0.331)	2(0.336)
ERR212041	both	0.85	2(0.162)	2(0.165)	2(0.337)
ERR212058	both	0.88	3(0.11)	2(0.09)	2(0.349)
ERR212059	both	0.86	3(0.11)	2(0.356)	2(0.356)

sample	# of verified method in refernce paper	major strain proportion	MixtureS	SMS	StrainFinder
ERR212069	both	0.86	3(0.092)	2(0.0)	2(0.312)
ERR212086	both	0.88	2(0.021)	2(0.179)	2(0.129)
ERR212098	both	0.84	3(0.073)	2(0.109)	2(0.066)
ERR212100	both	0.84	3(0.073)	2(0.309)	2(0.027)
ERR212101	both	0.85	3(0.076)	2(0.174)	2(0.009)
ERR212107	both	0.87	3(0.092)	2(0.112)	2(0.211)
ERR212112	both	0.89	4(0.191)	2(0.184)	2(0.357)
ERR212134	both	0.85	2(0.007)	2(0.143)	2(0.328)
ERR212161	both	0.87	3(0.102)	2(0.356)	2(0.336)
ERR212165	both	0.85	3(0.09)	2(0.01)	2(0.266)
ERR216899	both	0.88	4(0.212)	2(0.033)	2(0.367)
ERR216913	both	0.88	2(0.004)	2(0.37)	2(0.371)
ERR216914	both	0.75	5(0.235)	2(0.236)	2(0.225)
ERR216917	both	0.89	4(0.201)	2(0.083)	2(0.381)
ERR216932	both	0.89	3(0.233)	2(0.084)	2(0.387)
ERR216933	both	0.9	2(0.243)	2(0.084)	2(0.341)
ERR216942	both	0.87	2(0.019)	2(0.353)	2(0.366)
ERR216952	both	0.91	4(0.228)	2(0.007)	1(0.09)
ERR216956	both	0.89	3(0.153)	2(0.197)	2(0.386)
ERR216961	both	0.93	3(0.217)	2(0.051)	2(0.392)
ERR216966	both	0.88	3(0.131)	2(0.006)	2(0.316)
ERR216967	both	0.88	3(0.13)	2(0.215)	2(0.376)
ERR216971	both	0.69	6(0.211)	2(0.033)	2(0.187)
ERR216974	both	0.89	3(0.145)	2(0.186)	2(0.223)
ERR216977	both	0.89	3(0.167)	2(0.021)	2(0.387)
ERR216983	both	0.89	3(0.139)	2(0.322)	2(0.386)
ERR216984	both	0.88	3(0.123)	2(0.218)	2(0.374)
ERR216989	both	0.87	3(0.151)	2(0.099)	2(0.238)
ERR221524	both	0.88	2(0.005)	2(0.097)	2(0.336)
ERR221534	single	1	2(0.102)	2(0.47)	2(0.465)
ERR221536	both	0.87	3(0.101)	2(0.081)	2(0.249)
ERR221538	both	0.88	4(0.16)	2(0.351)	2(0.368)
ERR221539	both	0.82	3(0.073)	2(0.284)	2(0.105)
ERR221561	both	0.69	4(0.115)	2(0.132)	2(0.189)
ERR221567	both	0.87	3(0.107)	2(0.369)	2(0.369)
ERR221592	single	1	2(0.25)	2(0.047)	2(0.478)
ERR221611	both	0.87	3(0.102)	2(0.201)	2(0.357)
ERR245716	single	1	2(0.279)	2(0.099)	2(0.29)

sample	# of verified method in refernce paper	major strain proportion	MixtureS	SMS	StrainFinder
ERR245754	both	0.57	3(0.093)	2(0.064)	2(0.007)
ERR245758	both	0.79	2(0.078)	2(0.122)	2(0.038)
ERR245795	both	0.65	2(0.112)	2(0.196)	2(0.147)
ERR245797	both	0.82	2(0.07)	2(0.085)	2(0.054)
ERR323044	both	0.71	2(0.052)	2(0.272)	2(0.208)
ERR323054	both	0.66	5(0.192)	2(0.323)	2(0.152)
ERR323056	single	1	4(0.348)	2(0.481)	2(0.484)
ERR323082	both	0.71	4(0.114)	2(0.206)	2(0.162)
ERR473322	both	0.77	4(0.195)	2(0.213)	2(0.237)
ERR473340	single	1	5(0.38)	2(0.043)	2(0.437)
ERR473359	both	0.5	3(0.063)	2(0.234)	2(0.017)
ERR773806	both	0.91	4(0.3)	2(0.19)	2(0.353)

SMS identified two strains in each of these 195 datasets, which agreed well with the previous study [136]. This study showed that there were at least 11 heterozygous sites in each of these 195 datasets. Interestingly, SMS showed that the two strains in different datasets were the same, which was consistent with the fact that these datasets were from clinical samples collected from the same region. Moreover, SMS distinguished strains with similar abundance in these datasets. For instance, in the dataset ERR323056, there were 69 heterozygous sites observed in reads [136]. SMS predicted two strains with a relative abundance of 0.52 and 0.48. The previous study based on the SNP frequency identified only one strain, likely due to their similar abundance. Since the strain abundance was unknown, we compared the predicted abundance by SMS and by the previous study. The difference of the predicted strain abundance to the predicted abundance previously had a mean and median of 0.16 and 0.12 respectively, if we considered only the 186 datasets where the previous study correctly predicted the strain number.

4.1.3.5 SMS reconstructed strain genomes better than existing methods

We compared SMS with mixtureS [60] and StrainFinder [61]. mixtureS and StrainFinder showed better performance for novel strain identifications previously [60]. Since mixtureS works on one sample, we ran it on the pooled sample in each dataset. Because StrainFinder is unable to determine the strain numbers, we specified the known strain numbers in the corresponding datasets.

We compared the strain number, abundance and SNPs predicted by the three methods. SMS performed much better than others (Table 4-3). For instance, for simulated datasets with no shared SNPs among strains, SMS predicted the correct strain number in all 216 datasets while mixtureS correctly predicted the strain number in 98 datasets. On average, the predicted SNPs by SMS had a precision of 0.97 and a recall of 0.98, larger than those of mixtureS and StrainFinder. Moreover, the predicted strain abundance by SMS had an average MAE of 0.004, compared with 0.08 by mixtureS and 0.07 by StrainFinder.

Table 4-3: The performance of the three tools.

Dataset		SMS			mixtureS			StrainFinder	
		# (%) of datasets	Precision, Recall, F1	MAE	# (%) of datasets	Precision, Recall, F1	MAE	Precision, Recall, F1	MAE
702 simulated datasets	Unshared	216 (100%)	0.97, 0.98, 0.98	0.004	98 (45.37%)	0.81, 0.83, 0.80	0.08	0.66, 0.56, 0.53	0.07
	Shared	481 (98.97%)	0.97, 0.96, 0.96	0.008	184 37.86%	0.83, 0.58, 0.63	0.07	0.68, 0.56, 0.56	0.06
	All	697 (99.29%)	0.97, 0.96,0.96	0.007	282 40.17%	0.82, 0.66, 0.68	0.07	0.68, 0.56, 0.55	0.06
195 experimental datasets		195 (100%)	NA	0.16	146 74.87%	NA	0.12	NA	0.26

The three columns for each tool are the number (percentage) of datasets where the tool predicted the correct strain number; the precision, recall and F1 score of the predicted strain SNPs; and the average MAE of the predicted strain abundance.

We also studied the running time of different methods (Table 4-4). SMS took a little more time to run than mixtureS. However, the difference was not so evident. For all tools, the time cost mainly depended on the number of strains and the number of SNPs, instead of the dataset sizes.

Table 4-4: Tool running time on nine simulated datasets.

configuration_#samples_pooled coverage_species index	data size (MB)	# Reads	MixtureS (second)	StrainFinder (second)	SMS (second)
10:20:30:40_5_100_1	470	926579	69	1064	71
10:20:30:40_8_100_1	470	926581	40	1341	42
10:20:30:40_10_100_1	469	926579	24	1188	26
10:20:30:40_12_100_1	469	926581	33	1030	35
10:20:30:40_15_100_1	469	926579	28	1212	31
10:20:30:40_20_100_1	469	926580	40	1281	42
10:20:30:40_25_100_1	469	926578	83	1190	86
10:20:30:40_30_100_1	469	926578	25	1085	28
10:20:30:40_35_100_1	469	926585	18	1194	22

SMS reconstructs bacterial strain genomes with multiple shotgun samples. It considers the coverage variation of individual strains across samples to distinguish strains. As demonstrated in simulated and experimental datasets, SMS is able to separate strains with similar abundance. The capability to separate strains with similar abundance is in general improved with more samples and larger pooled coverage.

4.1.4 Discussion

SMS reconstructs bacterial strain genomes with a species reference genome and the raw sequencing reads. The reference is employed to map the cleaned reads. The chosen reference thus does not affect the predicted strain number and abundance, as they are inferred from the SNPs in strains that come from the mapped reads. SMS defines SNPs with an in-house procedure, which may affect the quality of individual SNPs. However, we do not think that the potential false SNPs will affect the predicted strain number and abundance, as they are determined by the coverage of the majority of SNPs in individual strains. Users may choose existing tools such as SAMtools [150] to define SNPs in samples. In addition, since reads are mapped to the reference genomes to predict bacterial strains, SMS can be applied to general metagenomic datasets instead of the shotgun samples for individual species illustrated here.

CHAPTER 5: CONCLUSION

5.1 Conclusion

In this dissertation, I have presented our work on distal gene regulation. We have studied motif pairs and their contribution to EPIs. We discovered 423 motif pairs that significantly co-occur in enhancers and promoters of interacting EP pairs. We demonstrated that these motif pairs are biologically meaningful and significantly enriched with motif pairs of known interacting TF pairs. We also showed that the identified motif pairs facilitated the discovery of the interacting EP pairs. Our study provides a comprehensive list of motif pairs that may contribute to physical EPIs, facilitating meaningful hypotheses for experimental validation.

We also study human RPGs and the role of their shared distal regulatory region in expression. We identified about 22,797 putative distal regulatory regions that directly or indirectly interact with human RPG promoters. A large proportion of these regions are only present in one cell line or one cell type, implying that RPGs may be differentially regulated across experimental conditions. We also noticed that subsets of RPGs share common regulatory regions across cell lines and cell types. The shared distal regulatory regions by RPGs may contribute to their coordinated regulation. By studying the overrepresented motifs in the identified regulatory regions, we showed that about two dozen motifs are common in these regions across cell lines and cell types. Our study shed new light on the coordinated transcriptional regulation of human RPGs.

We have explored the reconstruction of bacterial strain genomes from multiple shotgun metagenomic samples. The analysis of the bacterial strains is important for understanding drug resistance. Despite dozens of computational tools for bacterial strain studies, most are for known bacterial strains. Almost all remaining tools are designed to analyze individual samples or local strain regions. With multiple shotgun metagenomic samples routinely generated in a project, it is necessary to create methods to reconstruct novel bacterial strain genomes in multiple samples. We have developed a new tool called SMS that can de novo identify microbial strains from shotgun reads of multiple clonal or metagenomic samples without prior knowledge about the strains and their variations. Tested on 702 simulated and 195 experimental datasets, SMS reliably identified the strain number, abundance, and polymorphisms. Compared with the two existing approaches, SMS showed superior performance.

5.2 Future Work

5.2.1 EP motif pairs

Several directions may help to understand EP motif pairs better. First, although the identified motif pairs are likely to be useful in predicting EP interactions, they should be integrated with other features used previously [47, 74, 76] to fulfill their potential. Second, a more comprehensive collection of enhancers and their condition-specific activity may improve the quality of the predicted motif pairs. The number of enhancers we used is relatively small compared with the collected enhancers in other resources [151, 152]. Third, with more annotated known motifs, it may be better to discover motif pairs directly from known motifs. We look forward to exploring the EP motifs and their contribution to EPIs further.

5.2.2 RPG coordinated regulation

Although the progress we made, there is a long way to go to understand RPG coordinated regulation. First, chromatin interaction data with a much higher sequencing depth is greatly needed in other samples. With such data, we may understand how the number of the identified regulatory regions relates to the sequencing depth and how to more accurately define RPG regulatory regions. Moreover, we will be sure to know which regions are sample-specific and which are shared across samples and thus study how RPGs are able to orchestrate their coordinated expression with different regions under different conditions. Second, experimental validation of the functional consequences of certain RPG regulatory regions is a must. Such validation will not only generate new knowledge about RPG regulation but also provide guidelines to understand which of these regions may be truly functional. Third, integration of genomic and epigenomic data under the same conditions will greatly advance our understanding of RPG distal regulation. Finally, it is important also to study how RPGs are controlled at the translational level, which may contribute more to RPG coordinated regulation. We hope to work on these directions to under their regulation better.

5.2.3 Update SMS with more functionalities

SMS is not designed for the strain analysis of novel species. With more and more sequenced bacterial genomes, this issue may not be of concern in the future. Moreover, SMS considers only the reference genomic regions to reconstruct bacterial strain genomes, and thus does not consider accessory genes that are not represented in the chosen reference. In this sense, what SMS reconstructs is similar to the strain core genomes. In the future, we may develop methods to further discover accessory genes in strains, with the inferred strain number and abundance in samples [153].

LIST OF REFERENCES

1. Behjati, S. and P.S. Tarpey, *What is next generation sequencing?* Arch Dis Child Educ Pract Ed, 2013. **98**(6): p. 236-8.
2. Slatko, B.E., A.F. Gardner, and F.M. Ausubel, *Overview of Next-Generation Sequencing Technologies.* Curr Protoc Mol Biol, 2018. **122**(1): p. e59.
3. Salk, J.J., M.W. Schmitt, and L.A. Loeb, *Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations.* Nat Rev Genet, 2018. **19**(5): p. 269-285.
4. Sarda, S. and S. Hannenhalli, *Next-generation sequencing and epigenomics research: a hammer in search of nails.* Genomics Inform, 2014. **12**(1): p. 2-11.
5. Mardis, E.R., *ChIP-seq: welcome to the new frontier.* Nature Methods, 2007. **4**(8): p. 613-614.
6. Ruppert, S.M., et al., *JunD/AP-1-mediated gene expression promotes lymphocyte growth dependent on interleukin-7 signal transduction.* PLoS One, 2012. **7**(2): p. e32262.
7. Kukurba, K.R. and S.B. Montgomery, *RNA Sequencing and Analysis.* Cold Spring Harb Protoc, 2015. **2015**(11): p. 951-69.
8. Belton, J.M., et al., *Hi-C: a comprehensive technique to capture the conformation of genomes.* Methods, 2012. **58**(3): p. 268-76.
9. Pan, F., et al., *Integrative Array Analyzer: a software package for analysis of cross-platform and cross-species microarray data.* Bioinformatics, 2006. **22**(13): p. 1665-7.
10. Rao, S.S., et al., *A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping.* Cell, 2014. **159**(7): p. 1665-1680.
11. Hafner, M., et al., *Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP.* Cell, 2010. **141**(1): p. 129-41.
12. Helwak, A. and D. Tollervey, *Mapping the miRNA interactome by cross-linking ligation and sequencing of hybrids (CLASH).* Nature Protocols, 2014. **9**(3): p. 711-728.
13. Takahashi, H., et al., *CAGE (cap analysis of gene expression): a protocol for the detection of promoter and transcriptional networks.* Methods Mol Biol, 2012. **786**: p. 181-200.
14. Ding, J., X. Li, and H. Hu, *CCmiR: a computational approach for competitive and cooperative microRNA binding prediction.* Bioinformatics, 2017. **34**(2): p. 198-206.
15. Wang, S., et al., *Computational annotation of miRNA transcription start sites.* Briefings in Bioinformatics, 2020. **22**(1): p. 380-392.
16. Talukder, A., X. Li, and H. Hu, *Position-wise binding preference is important for miRNA target site prediction.* Bioinformatics, 2020. **36**(12): p. 3680-3686.
17. Ding, J., X. Li, and H. Hu, *TarPmiR: a new approach for microRNA target site*

prediction. Bioinformatics, 2016. **32**(18): p. 2768-75.

18. Cha, M., et al., *A two-stream convolutional neural network for microRNA transcription start site feature integration and identification.* Scientific Reports, 2021. **11**(1): p. 5625.
19. Barham, C., et al. *Application of deep learning models to microrna transcription start site identification*. in *2019 IEEE 7th International Conference on Bioinformatics and Computational Biology (ICBCB)*. 2019. IEEE.
20. Buenrostro, J.D., et al., *ATAC-seq: a method for assaying chromatin accessibility genome-wide.* Current protocols in molecular biology, 2015. **109**(1): p. 21.29. 1-21.29. 9.
21. Song, L. and G.E. Crawford, *DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells.* Cold Spring Harb Protoc, 2010. **2010**(2): p. pdb.prot5384.
22. Gibcus, J.H. and J. Dekker, *The context of gene expression regulation.* F1000 Biol Rep, 2012. **4**: p. 8.
23. Latchman, D., *Gene regulation*. 2007: Taylor & Francis.
24. O'Brien, J., et al., *Overview of MicroRNA Biogenesis, Mechanisms of Actions, and Circulation.* Frontiers in Endocrinology, 2018. **9**.
25. Maston, G.A., S.K. Evans, and M.R. Green, *Transcriptional regulatory elements in the human genome.* Annu Rev Genomics Hum Genet, 2006. **7**: p. 29-59.
26. Heidari, N., et al., *Genome-wide map of regulatory interactions in the human genome.* Genome Res, 2014. **24**(12): p. 1905-17.
27. Jaari, S., M.-H. Li, and J. Merilä, *A first-generation microsatellite-based genetic linkage map of the Siberian jay (Perisoreus infaustus): insights into avian genome evolution.* BMC Genomics, 2009. **10**(1): p. 1.
28. Wingender, E., et al., *TRANSFAC: A Database on Transcription Factors and Their DNA Binding Sites.* Nucleic Acids Research, 1996. **24**(1): p. 238-241.
29. Castro-Mondragon, J.A., et al., *JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles.* Nucleic Acids Research, 2021. **50**(D1): p. D165-D173.
30. Wunderlich, Z. and L.A. Mirny, *Different gene regulation strategies revealed by analysis of binding motifs.* Trends Genet, 2009. **25**(10): p. 434-40.
31. Hu, H., *An efficient algorithm to identify coordinately activated transcription factors.* Genomics, 2010. **95**(3): p. 143-150.
32. Pennacchio, L.A., et al., *Enhancers: five essential questions.* Nature Reviews Genetics, 2013. **14**(4): p. 288-295.
33. Li, X., et al., *Integrative analyses shed new light on human ribosomal protein gene regulation.* Scientific reports, 2016. **6**: p. 28619.
34. Wang, S., H. Hu, and X. Li, *Shared distal regulatory regions may contribute to the coordinated expression of human ribosomal protein genes.* Genomics, 2020. **112**(4): p.

2886-2893.

35. Newburger, D.E. and M.L. Bulyk, *UniPROBE: an online database of protein binding microarray data on protein-DNA interactions.* Nucleic Acids Res, 2009. **37**(Database issue): p. D77-82.

36. Bailey, T.L., et al., *MEME SUITE: tools for motif discovery and searching.* Nucleic Acids Res, 2009. **37**(Web Server issue): p. W202-8.

37. Li, X., S. Zhong, and W.H. Wong, *Reliable prediction of transcription factor binding sites by phylogenetic verification.* Proceedings of the National Academy of Sciences, 2005. **102**(47): p. 16945-16950.

38. Hu, H. and X. Li, *Transcriptional regulation in eukaryotic ribosomal protein genes.* Genomics, 2007. **90**(4): p. 421-423.

39. Mager, W.H. and R.J. Planta, *Coordinate expression of ribosomal protein genes in yeast as a function of cellular growth rate*, in *Molecular Mechanisms of Cellular Growth*. 1991, Springer. p. 181-187.

40. Hariharan, N., D.E. Kelley, and R.P. Perry, *Delta, a transcription factor that binds to downstream elements in several polymerase II promoters, is a functionally versatile zinc finger protein.* Proceedings of the National Academy of Sciences, 1991. **88**(21): p. 9799-9803.

41. Wagner, M. and R.P. Perry, *Characterization of the multigene family encoding the mouse S16 ribosomal protein: strategy for distinguishing an expressed gene from its processed pseudogene counterparts by an analysis of total genomic DNA.* Molecular and cellular biology, 1985. **5**(12): p. 3560-3576.

42. Lieb, J.D., et al., *Promoter-specific binding of Rap1 revealed by genome-wide maps of protein–DNA association.* Nature genetics, 2001. **28**(4): p. 327.

43. Martin, D.E., A. Soulard, and M.N. Hall, *TOR regulates ribosomal protein gene expression via PKA and the Forkhead transcription factor FHL1.* Cell, 2004. **119**(7): p. 969-979.

44. Li, X. and W.H. Wong, *Sampling motifs on phylogenetic trees.* Proceedings of the National Academy of Sciences, 2005. **102**(27): p. 9481-9486.

45. Ma, X., K. Zhang, and X. Li, *Evolution of Drosophila ribosomal protein gene core promoters.* Gene, 2009. **432**(1-2): p. 54-59.

46. Perry, R.P., *The architecture of mammalian ribosomal protein promoters.* BMC evolutionary biology, 2005. **5**(1): p. 15.

47. Talukder, A., et al., *EPIP: a novel approach for condition-specific enhancer–promoter interaction prediction.* Bioinformatics, 2019. **35**(20): p. 3877-3883.

48. Javierre, B.M., et al., *Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters.* Cell, 2016. **167**(5): p. 1369-1384. e19.

49. Zoetendal, E.G., A.D. Akkermans, and W.M. De Vos, *Temperature gradient gel electrophoresis analysis of 16S rRNA from human fecal samples reveals stable and host-specific communities of active bacteria.* Appl Environ Microbiol, 1998. **64**(10): p. 3854-9.

50. Anderson, S., *Shotgun DNA sequencing using cloned DNase I-generated fragments.* Nucleic Acids Res, 1981. **9**(13): p. 3015-27.

51. Qin, J., et al., *A human gut microbial gene catalogue established by metagenomic sequencing.* Nature, 2010. **464**(7285): p. 59-65.

52. Roumpeka, D.D., et al., *A Review of Bioinformatics Tools for Bio-Prospecting from Metagenomic Sequence Data.* Front Genet, 2017. **8**: p. 23.

53. Wang, Y., H. Hu, and X. Li, *MBMC: An Effective Markov Chain Approach for Binning Metagenomic Reads from Environmental Shotgun Sequencing Projects.* Omics, 2016. **20**(8): p. 470-9.

54. Wang, Y., H. Hu, and X. Li, *rRNAFilter: A Fast Approach for Ribosomal RNA Read Removal Without a Reference Database.* J Comput Biol, 2017. **24**(4): p. 368-375.

55. Li, X., et al., *When old metagenomic data meet newly sequenced genomes, a case study.* PLoS One, 2018. **13**(6): p. e0198773.

56. Hong, C.J., et al., *PathoScope 2.0: a complete computational framework for strain identification in environmental or clinical sequencing samples.* Microbiome, 2014. **2**.

57. Ahn, T.H., J.J. Chai, and C.L. Pan, *Sigma: Strain-level inference of genomes from metagenomic analysis for biosurveillance.* Bioinformatics, 2015. **31**(2): p. 170-177.

58. Roosaare, M., et al., *StrainSeeker: fast identification of bacterial strains from raw sequencing reads using user-provided guide trees.* PeerJ, 2017. **5**: p. e3353.

59. Ventolero, M.F., et al., *Computational analyses of bacterial strains from shotgun reads.* Briefings in Bioinformatics, 2022. **23**(2).

60. Li, X., H. Hu, and X. Li, *mixtureS: a novel tool for bacterial strain reconstruction from reads.* Bioinformatics, 2020.

61. Smillie, C.S., et al., *Strain Tracking Reveals the Determinants of Bacterial Engraftment in the Human Gut Following Fecal Microbiota Transplantation.* Cell Host & Microbe, 2018. **23**(2): p. 229-+.

62. Nayfach, S., et al., *An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography.* Genome Research, 2016. **26**(11): p. 1612-1625.

63. Jin, F., et al., *A high-resolution map of the three-dimensional chromatin interactome in human cells.* Nature, 2013. **503**(7475): p. 290-294.

64. Li, G., et al., *Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation.* Cell, 2012. **148**(1-2): p. 84-98.

65. Lieberman-Aiden, E., et al., *Comprehensive mapping of long-range interactions reveals folding principles of the human genome.* science, 2009. **326**(5950): p. 289-293.

66. Moore, J.E., et al., *A curated benchmark of enhancer-gene interactions for evaluating enhancer-target gene prediction methods.* Genome biology, 2020. **21**(1): p. 17.

67. Tang, Z., et al., *CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription.* Cell, 2015. **163**(7): p. 1611-27.

68. Mumbach, M.R., et al., *HiChIP: efficient and sensitive analysis of protein-directed genome architecture.* Nat Methods, 2016. **13**(11): p. 919-922.

69. Zhang, K., et al., *Systematic identification of protein combinations mediating chromatin looping.* Nature communications, 2016. **7**(1): p. 1-11.

70. Cao, Q., et al., *Reconstruction of enhancer–target networks in 935 samples of human primary cells, tissues and cell lines.* Nature genetics, 2017. **49**(10): p. 1428.

71. Corradin, O., et al., *Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits.* Genome research, 2014. **24**(1): p. 1-13.

72. He, B., et al., *Global view of enhancer–promoter interactome in human cells.* Proceedings of the National Academy of Sciences, 2014. **111**(21): p. E2191-E2199.

73. Okonechnikov, K., et al., *InTAD: chromosome conformation guided analysis of enhancer target genes.* BMC bioinformatics, 2019. **20**(1): p. 60.

74. Roy, S., et al., *A predictive modeling approach for cell line-specific long-range regulatory interactions.* Nucleic acids research, 2015. **43**(18): p. 8694-8712.

75. Singh, S., et al., *Predicting enhancer-promoter interaction from genomic sequence with deep neural networks.* Quantitative Biology, 2019. **7**(2): p. 122-137.

76. Whalen, S., R.M. Truty, and K.S. Pollard, *Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin.* Nature genetics, 2016. **48**(5): p. 488-496.

77. Zeng, W., M. Wu, and R. Jiang, *Prediction of enhancer-promoter interactions via natural language processing.* BMC genomics, 2018. **19**(2): p. 13-22.

78. Zhao, C., X. Li, and H. Hu, *PETModule: a motif module based approach for enhancer target gene prediction.* Scientific reports, 2016. **6**: p. 30043.

79. Zhuang, Z., X. Shen, and W. Pan, *A simple convolutional neural network for prediction of enhancer–promoter interactions with DNA sequence data.* Bioinformatics, 2019. **35**(17): p. 2899-2906.

80. Cao, F. and M.J. Fullwood, *Inflated performance measures in enhancer–promoter interaction-prediction methods.* Nature genetics, 2019. **51**(8): p. 1196-1198.

81. Xi, W. and M.A. Beer, *Local epigenomic state cannot discriminate interacting and non-interacting enhancer–promoter pairs with high accuracy.* PLoS computational biology, 2018. **14**(12): p. e1006625.

82. Talukder, A., H. Hu, and X. Li, *An intriguing characteristic of enhancer-promoter interactions.* BMC Genomics, 2021. **22**(1): p. 163.

83. Hnisz, D., D.S. Day, and R.A. Young, *Insulated neighborhoods: structural and functional units of mammalian gene control.* Cell, 2016. **167**(5): p. 1188-1200.

84. Duren, Z., et al., *Modeling gene regulation from paired expression and chromatin accessibility data.* Proc Natl Acad Sci U S A, 2017. **114**(25): p. E4914-e4923.

85. Wong, K.C., Y. Li, and C. Peng, *Identification of coupling DNA motif pairs on long-range chromatin interactions in human K562 cells.* Bioinformatics, 2016. **32**(3): p. 321-4.

86. Whalen, S., R.M. Truty, and K.S. Pollard, *Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin.* Nat Genet, 2016. **48**(5): p. 488-96.

87. Jing, F., S.W. Zhang, and S. Zhang, *Prediction of enhancer-promoter interactions using the cross-cell type information and domain adversarial neural network.* BMC Bioinformatics, 2020. **21**(1): p. 507.

88. Ren, G., et al., *CTCF-mediated enhancer-promoter interaction is a critical regulator of cell-to-cell variation of gene expression.* Molecular cell, 2017. **67**(6): p. 1049-1058. e6.

89. Weintraub, A.S., et al., *YY1 is a structural regulator of enhancer-promoter loops.* Cell, 2017. **171**(7): p. 1573-1588. e28.

90. Khan, A., et al., *JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework.* Nucleic acids research, 2018. **46**(D1): p. D260-D266.

91. Strous, M., et al., *Deciphering the evolution and metabolism of an anammox bacterium from a community genome.* Nature, 2006. **440**(7085): p. 790-4.

92. Andersson, R., et al., *An atlas of active enhancers across human cell types and tissues.* Nature, 2014. **507**(7493): p. 455-461.

93. Consortium, E.P., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.

94. Weirauch, M.T., et al., *Determination and inference of eukaryotic transcription factor sequence specificity.* Cell, 2014. **158**(6): p. 1431-1443.

95. Mahony, S. and P.V. Benos, *STAMP: a web tool for exploring DNA-binding motif similarities.* Nucleic acids research, 2007. **35**(suppl_2): p. W253-W258.

96. Ding, J., H. Hu, and X. Li, *SIOMICS: a novel approach for systematic identification of motifs in ChIP-seq data.* Nucleic acids research, 2014. **42**(5): p. e35-e35.

97. Ding, J., X. Li, and H. Hu, *Systematic prediction of cis-regulatory elements in the Chlamydomonas reinhardtii genome using comparative genomics.* Plant physiology, 2012. **160**(2): p. 613-623.

98. Ding, J., et al., *Systematic discovery of cofactor motifs from ChIP-seq data by SIOMICS.* Methods, 2015. **79**: p. 47-51.

99. Vaquerizas, J.M., et al., *A census of human transcription factors: function, expression and evolution.* Nature Reviews Genetics, 2009. **10**(4): p. 252-263.

100. Hu, J., H. Hu, and X. Li, *MOPAT: a graph-based method to predict recurrent cis-regulatory modules from known motifs.* Nucleic acids research, 2008. **36**(13): p. 4488-4497.

101. Wang, Y., et al., *Prognostic cancer gene signatures share common regulatory motifs.* Scientific reports, 2017. **7**(1): p. 1-9.

102. Zheng, Y., X. Li, and H. Hu, *Comprehensive discovery of DNA motifs in 349 human cells and tissues reveals new features of motifs.* Nucleic acids research, 2015. **43**(1): p. 74-83.

103. Grant, C.E., T.L. Bailey, and W.S. Noble, *FIMO: scanning for occurrences of a given motif.* Bioinformatics, 2011. **27**(7): p. 1017-1018.

104. Ding, J., et al., *Chipmodule: systematic discovery of transcription factors and their cofactors from chip-seq data*, in *Biocomputing 2013*. 2013, World Scientific. p.

320-331.

105. Stark, C., et al., *BioGRID: a general repository for interaction datasets.* Nucleic Acids Res, 2006. **34**(Database issue): p. D535-9.

106. Breiman, L., et al., *Classification and Regression Trees. The Wadsworth statisticsprobability series. 1984.* Wadsworth International Group, Belmont, CA, 1984.

107. Breiman, L., *Random forests.* Machine learning, 2001. **45**(1): p. 5-32.

108. Smola, A.J. and B. Schölkopf, *A tutorial on support vector regression.* Statistics and computing, 2004. **14**(3): p. 199-222.

109. Tibshirani, R., *Regression shrinkage and selection via the lasso.* Journal of the Royal Statistical Society: Series B (Methodological), 1996. **58**(1): p. 267-288.

110. Cai, X., et al., *Systematic identification of conserved motif modules in the human genome.* BMC genomics, 2010. **11**(1): p. 567.

111. Zhang, J., et al., *An integrative ENCODE resource for cancer genomics.* Nat Commun, 2020. **11**(1): p. 3696.

112. Vakoc, C.R., et al., *Proximity among distant regulatory elements at the beta-globin locus requires GATA-1 and FOG-1.* Mol Cell, 2005. **17**(3): p. 453-62.

113. Bond, H.M., et al., *ZNF423: A New Player in Estrogen Receptor-Positive Breast Cancer.* Front Endocrinol (Lausanne), 2018. **9**: p. 255.

114. Mesuraca, M., et al., *ZNF423 and ZNF521: EBF1 Antagonists of Potential Relevance in B-Lymphoid Malignancies.* Biomed Res Int, 2015. **2015**: p. 165238.

115. Matsubara, E., et al., *The role of zinc finger protein 521/early hematopoietic zinc finger protein in erythroid cell differentiation.* J Biol Chem, 2009. **284**(6): p. 3480-7.

116. Zhang, X., et al., *Analysis of high-resolution 3D intrachromosomal interactions aided by Bayesian network modeling.* Proc Natl Acad Sci U S A, 2017. **114**(48): p. E10359-e10368.

117. Bailey, S.D., et al., *ZNF143 provides sequence specificity to secure chromatin interactions at gene promoters.* Nat Commun, 2015. **2**: p. 6186.

118. Uechi, T., T. Tanaka, and N. Kenmochi, *A complete map of the human ribosomal protein genes: assignment of 80 genes to the cytogenetic map and implications for human disorders.* Genomics, 2001. **72**(3): p. 223-230.

119. Raiser, D.M., A. Narla, and B.L. Ebert, *The emerging importance of ribosomal dysfunction in the pathogenesis of hematologic disorders.* Leukemia & lymphoma, 2014. **55**(3): p. 491-500.

120. Vlachos, A., *Acquired ribosomopathies in leukemia and solid tumors.* Hematology 2014, the American Society of Hematology Education Program Book, 2017. **2017**(1): p. 716-719.

121. Angelastro, J.M., B. Töröcsik, and L.A. Greene, *Nerve growth factor selectively regulates expression of transcripts encoding ribosomal proteins.* BMC neuroscience, 2002. **3**(1): p. 3.

122. Blanchette, M., et al., *Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression.* Genome

research, 2006. **16**(5): p. 656-668.

123. Ding, J., H. Hu, and X. Li, *SIOMICS: a novel approach for systematic identification of motifs in ChIP-seq data.* Nucleic acids research, 2013. **42**(5): p. e35-e35.

124. Khan, A., et al., *JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework.* Nucleic acids research, 2017. **46**(D1): p. D260-D266.

125. Zheng, Y., X. Li, and H. Hu, *Comprehensive discovery of DNA motifs in 349 human cells and tissues reveals new features of motifs.* Nucleic acids research, 2014. **43**(1): p. 74-83.

126. Su, A.I., et al., *A gene atlas of the mouse and human protein-encoding transcriptomes.* Proceedings of the National Academy of Sciences, 2004. **101**(16): p. 6062-6067.

127. Xie, X., et al., *Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals.* Nature, 2005. **434**(7031): p. 338-345.

128. Wilcoxon, F., *Individual comparisons by ranking methods*, in *Breakthroughs in statistics*. 1992, Springer. p. 196-202.

129. Cai, X., H. Hu, and X. Li, *A new measurement of sequence conservation.* BMC genomics, 2009. **10**(1): p. 623.

130. Gallo, S.M., et al., *REDfly v3. 0: toward a comprehensive database of transcriptional regulatory elements in Drosophila.* Nucleic acids research, 2010. **39**(suppl_1): p. D118-D123.

131. Methe, B.A., et al., *A framework for human microbiome research.* Nature, 2012. **486**(7402): p. 215-221.

132. Human Microbiome Project, C., *Structure, function and diversity of the healthy human microbiome.* Nature, 2012. **486**(7402): p. 207-14.

133. Proctor, L.M., et al., *The Integrative Human Microbiome Project.* Nature, 2019. **569**(7758): p. 641-648.

134. Van Rossum, T., et al., *Diversity within species: interpreting strains in microbiomes.* Nat Rev Microbiol, 2020. **18**(9): p. 491-506.

135. Eyre, D.W., et al., *Detection of mixed infection from bacterial whole genome sequence data allows assessment of its role in Clostridium difficile transmission.* PLoS Comput Biol, 2013. **9**(5): p. e1003059.

136. Sobkowiak, B., et al., *Identifying mixed Mycobacterium tuberculosis infections from whole genome sequence data.* Bmc Genomics, 2018. **19**(1): p. 613.

137. Albanese, D. and C. Donati, *Strain profiling and epidemiology of bacterial species from metagenomic sequencing.* Nature Communications, 2017. **8**.

138. Anyansi, C., et al., *Computational Methods for Strain-Level Microbial Detection in Colony and Metagenome Sequencing Data.* Front Microbiol, 2020. **11**: p. 1925.

139. Li, X., et al., *BHap: a novel approach for bacterial haplotype reconstruction.* Bioinformatics, 2019. **35**(22): p. 4624-4631.

140. Luo, C., et al., *ConStrains identifies microbial strains in metagenomic datasets.* Nat Biotechnol, 2015. **33**(10): p. 1045-52.

141. Pulido-Tamayo, S., et al., *Frequency-based haplotype reconstruction from deep sequencing data of bacterial populations.* Nucleic Acids Res, 2015. **43**(16): p. e105.

142. Costea, P.I., et al., *metaSNV: A tool for metagenomic strain level analysis.* PLoS One, 2017. **12**(7): p. e0182392.

143. Quince, C., et al., *DESMAN: a new tool for de novo extraction of strains from metagenomes.* Genome Biol, 2017. **18**(1): p. 181.

144. Sankar, A., et al., *Bayesian identification of bacterial strains from sequencing data.* Microb Genom, 2016. **2**(8): p. e000075.

145. Scholz, M., et al., *Strain-level microbial epidemiology and population genomics from shotgun metagenomics.* Nat Methods, 2016. **13**(5): p. 435-8.

146. Tamburini, F.B., et al., *Precision identification of diverse bloodstream pathogens in the gut microbiome.* Nat Med, 2018. **24**(12): p. 1809-1814.

147. Truong, D.T., et al., *Microbial strain-level population structure and genetic diversity from metagenomes.* Genome Research, 2017. **27**(4): p. 626-638.

148. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nat Methods, 2012. **9**(4): p. 357-9.

149. Sczyrba, A., et al., *Critical Assessment of Metagenome Interpretation-a benchmark of metagenomics software.* Nat Methods, 2017. **14**(11): p. 1063-1071.

150. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-9.

151. Gao, T. and J. Qian, *EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species.* Nucleic acids research, 2020. **48**(D1): p. D58-D64.

152. Wang, J., et al., *HACER: an atlas of human active enhancers to interpret regulatory variants.* Nucleic acids research, 2019. **47**(D1): p. D106-D112.

153. Talukder, A., et al., *Interpretation of deep learning in genomics and epigenomics.* Briefings in Bioinformatics, 2020. **22**(3).