Genomics 112 (2020) 2886–2893

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

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

Saidi Wang[a], Haiyan Hu[a,*], Xiaoman Li[b,*]

[a] Department of Computer Science, University of Central Florida, Orlando, FL, 32816, USA
[b] Burnett School of Biomedical Science, School of Medicine, University of Central Florida, Orlando, FL, 32816, USA

ARTICLE INFO

Keywords:
Ribosomal protein genes
Chromatin interaction
Distal regulatory regions
Coordinated expression

ABSTRACT

To identify the potential distal regulatory regions of human ribosomal protein genes (RPGs) and to understand their characteristics, we studied the chromatin interactions in seven cell lines and four primary cell types. We identified 22,797 putative 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 groups of RPGs, which are the same groups across cell lines and cell types, share common regulatory regions. These shared regulatory regions by RPGs may contribute to their coordinated regulation. By studying the overrepresented motifs in the identified regulatory regions, we showed that there are about two dozen motifs in these regions shared across cell lines and cell types. Our study shed new light on the coordinated transcriptional regulation of human RPGs.

1. Introduction

It is important to study the transcriptional regulation of ribosomal protein genes (RPGs) [1,2]. 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 [3,4]. As a set of essential genes and one type of the most abundantly expressed genes [5,6], 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 [7]. 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 [1,8].

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 [9,10]. Later, high-throughput experiments showed that TFs such as RAP1 and FHL1 bind to their TFBSs in promoters of almost all RPGs in yeast [11,12]. 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 [7,13–15].

All above studies focused on RPG promoter regions. Rarely is there a study that explores the distal regulatory regions of RPGs. Here and in the following, promoters were defined as previously [16,17] 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 [16]. For the sake of simplicity, henceforth, we used “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.

Although our previous study shed new light on human RPG transcriptional regulation, it is limited in the following aspects [16]. 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 ($\geq$ 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 they may be further away from the target genes [18].

* Corresponding authors.
E-mail addresses: tjwangsaidi@knights.ucf.edu (S. Wang), haihu@cs.ucf.edu (H. Hu), xiaoman@mail.ucf.edu (X. Li).

https://doi.org/10.1016/j.ygeno.2020.03.028
Received 7 January 2020; Received in revised form 12 March 2020; Accepted 29 March 2020
Available online 30 March 2020
0888-7543/

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 [19,20] (Material and Methods). We identified 22,797 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 to be 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, we identified common TF binding motifs shared by samples. Our study shed new light on the distal regulation of the human RPGs.

2. Materials and methods

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 from the RPG database (http://ribosome.med.miyazaki-u.ac.jp/) and found that they were consistent.

We obtained chromatin interaction data from two studies (Supplementary Table S1). One was the Hi-C data in seven cell lines (GM12878, IMR90, HMEC, KBM7, HUVEC, NHEK, K562) from Rao et al. [20]. 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 [20]. 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 supported with at least 30 normalized Hi-C reads as the 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 [21,22]. Note that these pairs of interacting regions can be from different chromosomes, although the majority of them are intra-chromosomal interactions. We obtained the corresponding DHS data for each of the seven samples from the ENCODE project [23] (https://www.encodeproject.org/search/?type=Experiment).

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 [19]. These eight cell types were aCD4, nB, EP, tB, tCD8, FoeT, naCD4, and tCD4. The interactions between genomic regions were defined in the Supplementary Table S2 in the original study [19]. All pairs of interacting regions were from the same chromosomes. We were able to download the corresponding DHS data for the following four cell types: aCD4, nB, tCD8, and FoeT from https://www.encodeproject.org/search/?type=Experiment.

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 (Fig. 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 was calculated with the bedtools (https://bedtools.readthedocs.io/en/latest/content/tools/intersect.html) by 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 or 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.

Each direct or indirect region was about 5000 bps long (Supplementary Table S1), which depended on the Hi-C resolution and was on average much longer than known regulatory regions [24,25]. 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 [24,25]. We then obtained the DNA sequences for these processed regions.

2.3. Motif analyses in promoters and other regulatory regions

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 [26,27]. 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 [27,28]. 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 [24].

We compared the predicted motifs with the motifs in the JASPAR database [29]. 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 [30] similarity E-value smaller than 1e-5, a cutoff used in previous studies [31,32].

2.4. Other analyses

We downloaded the normalized gene expression data in 79 different tissues from the GNF Expression Atlas 2 [33], which is widely used to study gene transcriptional regulation [24,34]. 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 [35].

3. Results

3.1. About 22,797 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 [19,20] and the DHSs in the corresponding samples [23] (Material and Methods) (Fig. 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 interacts with this RPG promoter based on the corresponding chromatin interaction data and overlaps with at least a DHS region in this sample, and an indirect region of a RPG is a region that indirectly interacts with this RPG promoter and overlaps with at least a DHS region (Material and Methods). In total, we identified about 22,797 putative regulatory regions that interacted with RPG promoters in different samples (Supplementary Table S1). The majority of these regions were distal regions (Supplementary Table S2). The number of the putative regions varied

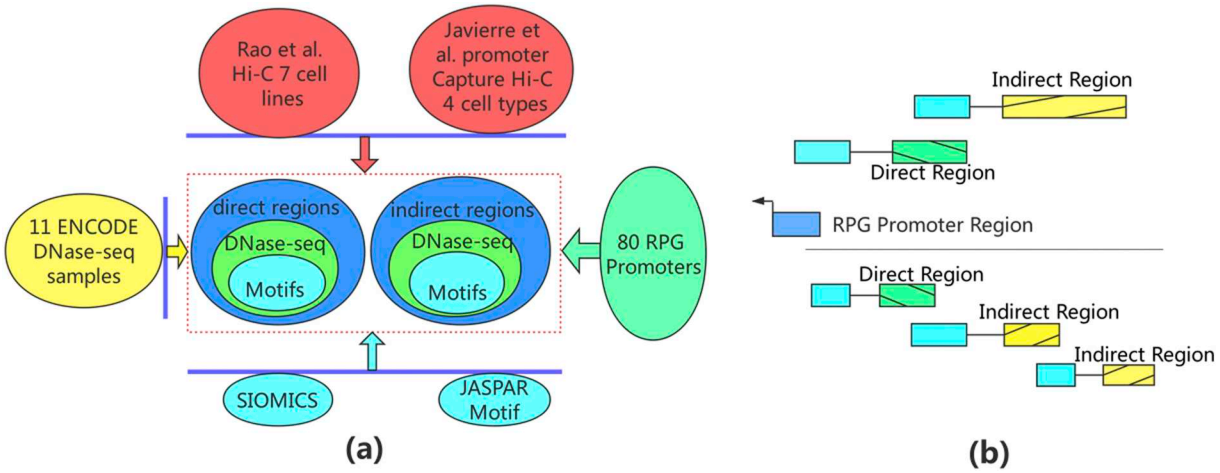

Fig. 1. The identification of the putative RPG regulatory regions: (a) Different sources of interaction data were used to infer RPG regulatory regions and the enriched TF binding motifs in these regions; (b) An example of direct and indirect regulatory regions of a RPG.

across samples. The details were in the following.

In seven samples from Rao et al., we identified 16,588 potential RPG regulatory regions (Supplementary Table S1). The number of regions in one sample varied from 338 to 16,541, depending on the sequencing depth and the nature of the samples (Fig. 2A, C, E). For instance, in GM12878, there were 2226 direct regions and 14,315 indirect regions identified, which was at least eight times of the direct and indirect regions identified in other samples (Supplementary Table S1). This was because GM12878 had a sequencing depth about nine to seventeen times of that in other samples [20]. In general, with a larger sequencing depth in a sample, there are more potential RPG regulatory regions identified in this sample (Supplementary Fig. S1). 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 of RPG regulatory regions instead of the effect of different sequencing depth (Fig. 2C). On average, we identified 470 direct and 2745 indirect regions in a sample excluding GM12878.

To assess the statistical significance of the identified regulatory regions, we randomly chose 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 number of direct and indirect regions that interacted with the 80 random regions (Fig. 2A, C, E). 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 significantly more potential regulatory regions.

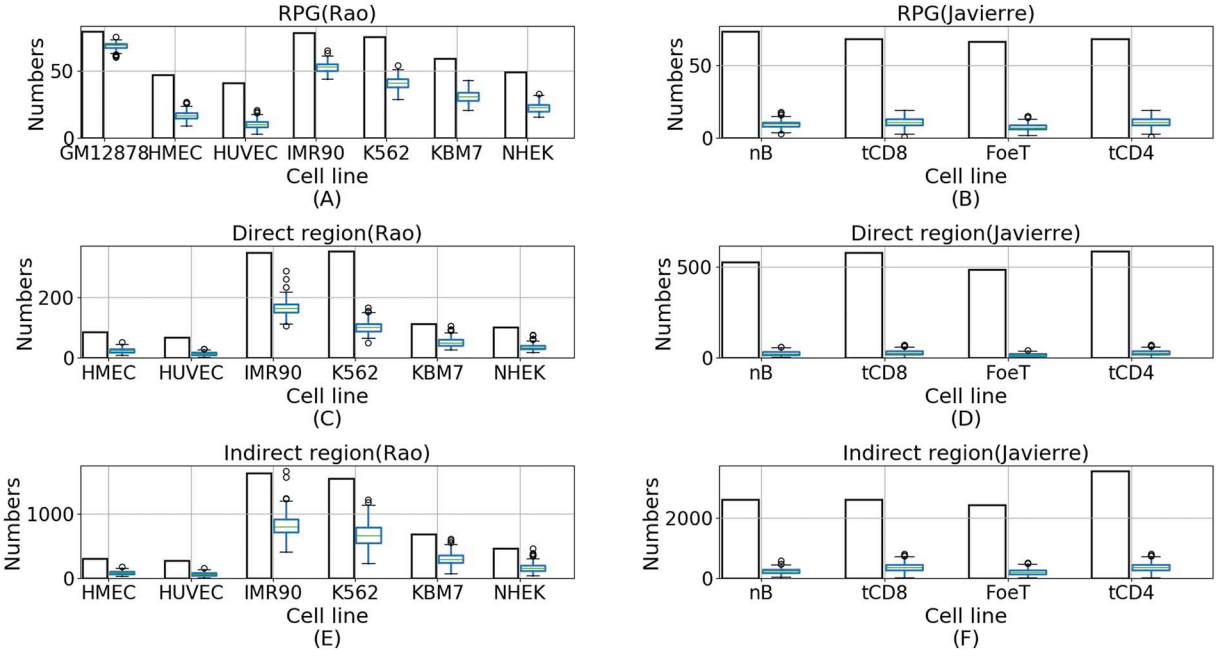

Fig. 2. The identified putative RPG distal regulatory regions: (A) & (B) The number of RPGs with identified regulatory regions in a sample; (C) & (D) The number of identified direct regions in a sample; (E) & (F) The number of identified indirect regions in a sample. In each section, the box plot is from 200 simulated sets of 80 random genomic regions. There are 2226 direct and 14,315 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.

Similarly, we identified in total 6209 regions that were likely to regulate RPGs in four samples from Javierre et al. [19]. 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 (Fig. 2B). The number of regulatory regions in a sample varied from 2902 to 4139, depending on the samples instead of the sequencing depth (Fig. 2D and F). For instance, the sample FoeT had the largest sequencing depth, while the number of regions identified in FoeT was the smallest (Supplementary Table S1). 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 (Fig. 2D, F, and Supplementary Table S1). 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 [19,20], we relaxed 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 Fig. 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 defined RPG direct and indirect regions 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 were 299,917 bps and 10,000 bps, respectively, in Rao et al.'s samples; and 93,913 bps and 3919 bps, respectively, in Javierre et al.'s samples.

We also studied the distances between the identified regions and their corresponding RPGs (Fig. 3, Supplementary Table S2). 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 28,007 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 551,403 bps. Since almost all human RPGs have neighboring protein-coding genes within 1 Mbps [16], this suggested that RPGs were not the closest genes to many of these regions (Supplementary Table S2).

3.2. The identified putative RPG regulatory regions varied dramatically across samples

We compared the identified RPG regulatory regions in different samples (Table 1). We found that the majority of them were not the same and not even overlapping across samples. This suggests that RPGs are likely to be regulated by different distal regions under different experimental conditions, which is consistent with our previous study [16].

More than 91% (15148) of the 16,588 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 to be 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 (Supplementary Table S1), 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 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%

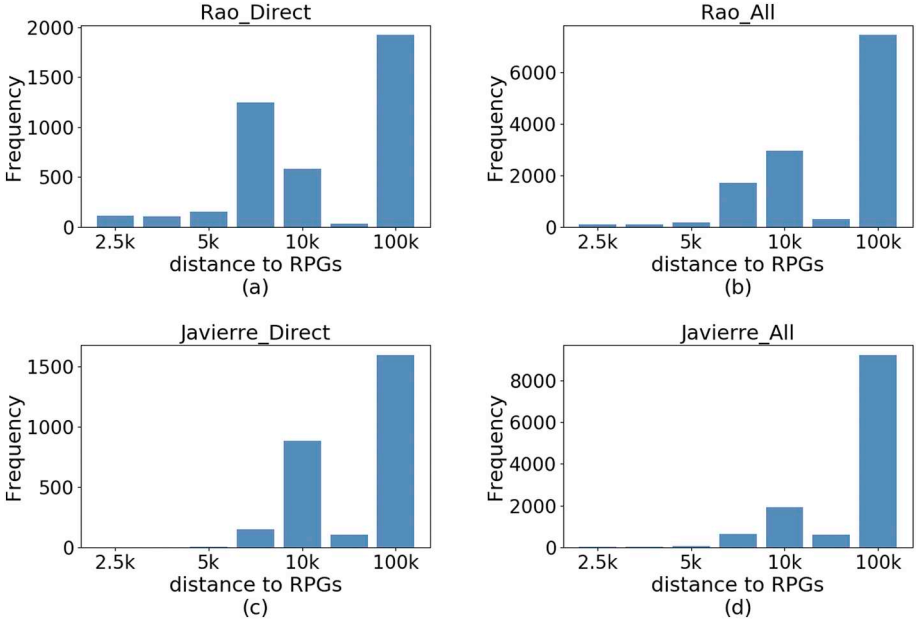

Fig. 3. The distance between a regulatory region and the corresponding RPG. We divided the distances into seven bins, such as the bin ≥2.5 k but < 5 k, where k means kilobase pairs. (a) and (b) Direct regions and all regulatory regions from Rao et al.'s data, respectively; (c) and (d) Direct regions and all regulatory regions from Javierre et al.'s data, respectively.

Table 1
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	16,588 (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.

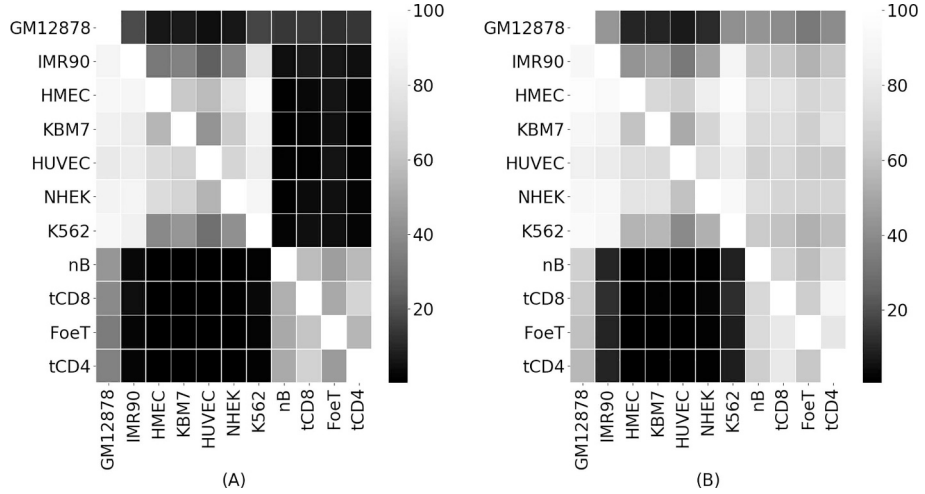

Fig. 4. The comparisons of regulatory regions across samples. The percentage of direct regions in a sample (row) overlapped with (A) the direct regions and (B) all regulatory regions in another sample (column) is represented by the heatmap.

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, 11.9% of the regions were shared by all four samples for RPGs, compared with the average 6.8% of the regions shared by four samples for the 80 random regions (Table 1). These observations are consistent with the fact that RPGs and their regulation are more conserved across samples than random regions.

To further understand the conservation of these regions across samples, we studied how the direct regions were shared across samples (Fig. 4). The direct regions were those that physically interacted with the RPG promoters and were detected by the Hi-C experiments (Fig. 1). We found that a large proportion of the direct regions in a sample did not overlap with the direct regions in any other sample 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, probably because of 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 that RPGs are likely to have different regulatory regions across samples. Otherwise, we should have seen 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 the 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 (Supplementary Table S3). Moreover, since GM12878 had a much higher sequencing depth than other samples, it should have included almost all direct regions in other samples, while in fact, around 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.

We also compared the direct regions in a sample with all regulatory regions in another sample (Fig. 4B and Supplementary Table S4). There were more direct regions in a sample identified in another sample, when we considered all regulatory regions instead of only the direct regulatory regions. However, the increment was small, only a handful of percentage, indicating that the majority of the direct 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 the direct regions not shared by samples, 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 the direct regions in HUVEC were not identified in GM12878 (supplementary Table S4).

We also studied how the indirect regions were shared across samples (Supplementary Tables S5 and S6). The indirect regions were not as conserved as the 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.

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

Table 2
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)

pair of RPGs that had their regulatory regions overlapped (Table 2). 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 but 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 2). 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 2 and Supplementary Table S7). The RPG pairs in GM12878 were often not identified in other samples, which may be due to the much smaller sequencing depth in other samples.

With the above pairs of RPGs in a sample, we grouped them into clusters of RPGs in two approaches (Table 2). 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 (Supplementary Table S7). 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 rather than independent.

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 even be up to a couple of thousand bps [24,25,36,37]. 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 [26,27], as the number of known motifs was still limited [29,38,39]. We found that about two dozen motifs were shared by different samples.

Table 3
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%

By de novo motif discovery (Material and Methods), we identified 68 to 1118 motifs in different samples (Table 3). 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. 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 number of motifs in the random sequences in every sample (Table 3), suggesting that the identified motifs in RPG regulatory regions were statistically significant and likely to be meaningful.

To assess the biological meaning of the predicted motifs, we further compared the predicted motifs with the known motifs in the JASPAR database [29]. In a sample, 38.28% to 54.36% of motifs were similar to known motifs (STAMP *E*-value < 1E-5 [30]). 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 (Table 2), suggesting that these motifs were likely to be 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 [16]. 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.

In spite 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 (Supplementary Table S8). 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 (Supplementary Table S8).

4. Discussion

We studied the putative regulatory regions of human RPGs in eleven samples. We identified about 22,797 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 (Supplementary Table S3). Moreover, different RPGs may share common regulatory regions and form a co-regulated gene groups. Such co-regulated gene groups were conserved across samples. In addition, in different samples, common regulatory motifs were 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 16,588 regulatory regions that were likely to 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 [19]. 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 [16]. 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 the regions shared by the majority samples. It also implied that RPGs are likely to be 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 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 be RPG-specific and they may contribute to the RPG coordinated regulation as well. It is worth investigating whether these shared motifs (Supplementary Table S8), 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 that this is because the promoter capture Hi-C is biased, there is something different among the samples in the two studies, or something else.

Although the progresses 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 be able to understand how the number of the identified regulatory regions relate to the sequencing depth and how to more accurately define RPG regulatory regions. Moreover, we will be for 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 a 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 to also study how RPGs are controlled at the translational level, which may contribute more to RPG coordinated regulation. We hope to work on these directions in the future to understand their regulation better.

Author contributions

H.H. and X.L. designed the study. S.W., H.H. and X.L. analyzed data and wrote the manuscript.

Declaration of Competing Interest

The authors declare no competing interests.

Acknowledgement

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

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2020.03.028.

References

[1] H. Hu, X. Li, Transcriptional regulation in eukaryotic ribosomal protein genes, Genomics 90 (4) (2007) 421–423.
[2] T. Uechi, T. Tanaka, 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 72 (3) (2001) 223–230.
[3] D.M. Raiser, A. Narla, B.L. Ebert, The emerging importance of ribosomal dysfunction in the pathogenesis of hematologic disorders, Leuk. Lymphoma 55 (3) (2014) 491–500.
[4] A. Vlachos, Acquired ribosomopathies in leukemia and solid tumors, Hematology 2014, the American Society of Hematology Education Program Book, 2017(1) 2017, pp. 716–719.
[5] J.M. Angelastro, B. Töröcsik, L.A. Greene, Nerve growth factor selectively regulates expression of transcripts encoding ribosomal proteins, BMC Neurosci. 3 (1) (2002) 3.
[6] B. Li, C.R. Nierras, J.R. Warner, Transcriptional elements involved in the repression of ribosomal protein synthesis, Mol. Cell. Biol. 19 (8) (1999) 5393–5404.
[7] X. Li, S. Zhong, W.H. Wong, Reliable prediction of transcription factor binding sites by phylogenetic verification, Proc. Natl. Acad. Sci. 102 (47) (2005) 16945–16950.
[8] W.H. Mager, R.J. Planta, Coordinate expression of ribosomal protein genes in yeast as a function of cellular growth rate, Molecular Mechanisms of Cellular Growth, Springer, 1991, pp. 181–187.
[9] N. Hariharan, D.E. Kelley, R.P. Perry, Delta, a transcription factor that binds to downstream elements in several polymerase II promoters, is a functionally versatile zinc finger protein, Proc. Natl. Acad. Sci. 88 (21) (1991) 9799–9803.
[10] M. Wagner, 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, Mol. Cell. Biol. 5 (12) (1985) 3560–3576.
[11] J.D. Lieb, et al., Promoter-specific binding of Rap1 revealed by genome-wide maps of protein–DNA association, Nat. Genet. 28 (4) (2001) 327.
[12] D.E. Martin, A. Soulard, M.N. Hall, TOR regulates ribosomal protein gene expression via PKA and the Forkhead transcription factor FHL1, Cell 119 (7) (2004) 969–979.
[13] X. Li, W.H. Wong, Sampling motifs on phylogenetic trees, Proc. Natl. Acad. Sci. 102 (27) (2005) 9481–9486.
[14] X. Ma, K. Zhang, X. Li, Evolution of Drosophila ribosomal protein gene core promoters, Gene 432 (1–2) (2009) 54–59.
[15] R.P. Perry, The architecture of mammalian ribosomal protein promoters, BMC Evol. Biol. 5 (1) (2005) 15.
[16] X. Li, et al., Integrative analyses shed new light on human ribosomal protein gene regulation, Sci. Rep. 6 (2016) 28619.
[17] C. Zhao, X. Li, H. Hu, PETModule: a motif module based approach for enhancer target gene prediction, Sci. Rep. 6 (2016) 30043.
[18] A. Talukder, et al., EPIP: a novel approach for condition-specific enhancer–promoter interaction prediction, Bioinformatics 35 (20) (2019) 3877–3883.
[19] B.M. Javierre, et al., Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters, Cell 167 (5) (2016) 1369–1384 e19.
[20] S.S. Rao, et al., A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping, Cell 159 (7) (2014) 1665–1680.
[21] F. Jin, et al., A high-resolution map of the three-dimensional chromatin interactome in human cells, Nature 503 (7475) (2013) 290.
[22] G. Li, et al., Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation, Cell 148 (1–2) (2012) 84–98.
[23] E.P. Consortium, An integrated encyclopedia of DNA elements in the human genome, Nature 489 (7414) (2012) 57.
[24] M. Blanchette, et al., Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression, Genome Res. 16 (5) (2006) 656–668.
[25] X. Cai, et al., Systematic identification of conserved motif modules in the human genome, BMC Genomics 11 (1) (2010) 567.
[26] J. Ding, et al., Systematic discovery of cofactor motifs from ChIP-seq data by SIOMICS, Methods 79 (2015) 47–51.
[27] J. Ding, H. Hu, X. Li, SIOMICS: a novel approach for systematic identification of motifs in ChIP-seq data, Nucleic Acids Res. 42 (5) (2013) e35.
[28] C.E. Grant, T.L. Bailey, W.S. Noble, FIMO: scanning for occurrences of a given motif, Bioinformatics 27 (7) (2011) 1017–1018.
[29] A. Khan, et al., JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework, Nucleic Acids Res. 46 (D1) (2017) D260–D266.
[30] S. Mahony, P.V. Benos, STAMP: a web tool for exploring DNA-binding motif similarities, Nucleic Acids Res. 35 (suppl_2) (2007) W253-W258.
[31] Y. Zheng, X. Li, H. Hu, Comprehensive discovery of DNA motifs in 349 human cells and tissues reveals new features of motifs, Nucleic Acids Res. 43 (1) (2014) 74–83.
[32] J. Ding, X. Li, H. Hu, Systematic prediction of cis-regulatory elements in the Chlamydomonas reinhardtii genome using comparative genomics, Plant Physiol. 160 (2) (2012) 613–623.
[33] A.I. Su, et al., A gene atlas of the mouse and human protein-encoding transcriptomes, Proc. Natl. Acad. Sci. 101 (16) (2004) 6062–6067.
[34] X. Xie, et al., Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals, Nature 434 (7031) (2005) 338–345.
[35] F. Wilcoxon, Individual comparisons by ranking methods, Breakthroughs in Statistics, Springer, 1992, pp. 196–202.
[36] X. Cai, H. Hu, X. Li, A new measurement of sequence conservation, BMC Genomics 10 (1) (2009) 623.
[37] S.M. Gallo, et al., REDfly v3. 0: toward a comprehensive database of transcriptional regulatory elements in Drosophila, Nucleic Acids Res. 39 (suppl_1) (2010) D118-D123.
[38] J. Ding, et al., Chipmodule: systematic discovery of transcription factors and their cofactors from chip-seq data, Biocomputing 2013, World Scientific, 2013, pp. 320–331.
[39] M.T. Weirauch, et al., Determination and inference of eukaryotic transcription factor sequence specificity, Cell 158 (6) (2014) 1431–1443.