Detecting Differential Transcription Factor Binding Based on Single-Cell DNA Accessibility

John Lin
Boston Latin School; Boston, USA
Email: johnzhuanglin@gmail.com
Mentor: Dr. Thouis (Ray) Jones

***Abstract* — Common genetic diseases—systemic diseases that are caused by thousands of mutations—affect millions of people around the world. Many of these mutations fall within regulatory regions. While the mutations associated with these diseases are widely known, the link between these mutations and their role in disease pathogenicity has largely gone undiscovered. This study harnesses single cell ATAC-seq data to differentiate bound and unbound sites in regulatory regions, serving as a first step to understanding these diseases. By computing observed and expected cuts for footprint regions, this study finds that regions with lower observed cuts than expected cuts conferred to protection from sequencing enzymes, indicating the presence of a bound transcription factor. In contrast, regions with higher observed cuts than expected indicate the absence of protection from sequencing enzymes, suggesting an absence of a bound transcription factor. In distinguishing between bound and unbound transcription factors, this study paves the way for using single cell ATAC-seq to understand common diseases by identifying the cell types and changes in transcription factor binding caused by mutations**

Key Words – transcription factor, single-cell sequencing, common genetic diseases, DNA accessibility

INTRODUCTION

Common genetic diseases are often caused by thousands of mutations and affect millions of people in the world each year. These diseases, which include heart disease, schizophrenia and Crohn's disease, are difficult to treat because their exact pathogenicity remains unclear. Despite the identification of mutations for these diseases, the link between these mutations and their role in expressing disease phenotypes has been largely undiscovered, particularly in understanding the cell types and biological pathways that are affected by the mutations. An analysis of these mutations, however, found that 80% of these mutations occur in distal regions, containing *cis* regulatory elements that drive gene expression [1]. The exact effects of these mutations on these regulatory regions remain unclear. Specifically, understanding the role of mutations in transcription factor binding is incredibly important as transcription factors play a large role in activating/suppressing gene expression. Despite this, there has been little information on detecting differences in transcription factor binding, serving as the premise for this study.

Sequencing technology has provided a plethora of information about gene regulation in cells. Specifically, Assay for Transposase-Accessible Chromatin sequencing (ATAC-seq) and DNase-seq identify accessible regions which are often regulatory regions. In the case of DNase-seq, this has led to the creation of DNase-I hypersensitive site (DHS) databases, creating a reference library for mapping regulatory regions [2] [3]. ATAC-seq harnesses a sequencing enzyme, Tn5 transposase, to attach adapters onto regions without nucleosomes which characterize accessible regions; these adapters are visualized by the sequencer in order to visualize the accessible regions. Furthermore, the rise of single cell ATAC-seq (scATAC-seq) has provided clearer resolution in identifying the accessible properties within a single cell as well as subpopulations in cells, particularly tissue [4]. This new technology has provided unprecedented ability to visualize accessibility regions in order to not only cluster different cell types but also explain the role of specific transcription factors in biological processes, such as cell differentiation [5].

These sequencing techniques have also been applied in DNA footprinting. ATAC-seq and DNase-seq use sequencing enzymes to cut accessible regions, which are not bound or "protected" by a protein. Footprinting involves observing drops in accessibility caused by the inability for a sequencing enzyme to cut a bound region. This technique has been harnessed to study within the regulatory region, especially in relation to transcription factor binding. Transcription factors typically recognize and bind to particular DNA motifs (short, distinct sequences) in accessible regions in the genome. However, not all accessible motifs have transcription factors to bind to them due to a variety of factors, including the lack of transcription factor expression or the absence of necessary cofactors. Footprinting allows us to visualize which motifs are actually bound by observing slight drops in accessibility caused by transcription factor:motif binding [6].

Despite the wealth of information that footprinting brings, there has been little knowledge about whether this information can be translated to single cell sequencing data. By determining whether footprinting can be detected in scATAC-seq data, this study identifies where these transcription factors may bind and also characterizes

whether they are bound or unbound. In relation to common genetic diseases, mutations in these regions often confer to changes in gene expression [6]. Thus, detecting differences in transcription factor binding, in the context of allele specific binding, for single cells has the potential to be an incredibly powerful tool in understanding these complex common genetic diseases [7].

Building upon this understanding, this study addresses this fundamental challenge in understanding common diseases by providing the link between mutation and disease phenotype through identifying the cell type, biological pathways and nearby genes that are involved in functional changes due to variants linked to disease phenotypes. In doing so, this study opens the doors for drug development that targets and stimulates or suppresses transcription factor binding sites in affected cells to address these common diseases.

METHODS

A. Extraction and Verification of Sequencing Data

Single cell ATAC-seq fragments data for CD8+ T cells was downloaded from Gene Expression Omnibus platform [8]. (Accession Number: GSM4293910). CD8+ T cell data was used specifically due to its relevance to a variety of autoimmune diseases, such as Crohn's disease as well as abundance of high-quality CD8+ ATAC data. DHS index and footprint data were downloaded from their respective papers to serve as a reference to map accessible fragments to footprints and index regions[2] [3]. The ATAC-seq fragments and the DHS index were overlapped in *bedtools* to map accessibility (in the form of these fragments) to footprints, and the number of overlapped cuts along with the location of the overlapped regions were mapped in order to confirm that the cuts came from lymphoid cells [9].

A. Enrichment Determination

In order to determine the most relevant or active transcription factors in the CD8+ T cell ATAC data, the enrichment of these transcription factors was calculated. To do so, a threshold was set at one cut per base (from arcsinh transformed data). If an overlapped region had more than one cut per base, then it was considered "more accessible;" If an overlapped region had less than one cut per base, then it was considered "less accessible." Each of these data sets (more accessible and less accessible) were overlapped with a DHS footprint library in order to match footprints and motifs with an accessible category. For each transcription factor, the frequency of the motif in the more accessible and less accessible categories were calculated separately. The equation for the enrichment or the fraction of the motif that was present in the more accessible category was developed and was calculated such that:

$$E = \frac{freq(more\ accessible)}{freq(more\ accessible) + freq(less\ accessible)}$$

B. Footprint Ranking and Candidate Identification

The enrichment ratio serves as an indicator for how active a certain motif was in the CD8+ T cell type. After calculating the enrichment ratio for each motif, we ranked all the transcription factors by the motif enrichment ratio. The ten most enriched motifs were the candidate footprints.

C. Calculating Sequence Bias and Expected Cuts

After determining the candidate footprints, the number of cuts expected within a region was calculated to serve as a control to detect differential binding. The Tn5 enzyme for ATAC-seq does not cut the fragments in a uniform manner, but instead, demonstrates a bias for cutting at certain sequences [10]. To account for sequence bias when determining the expected number of cuts, a table listing different 8-mers and their respective bias values was obtained [10]. For each candidate footprint, the DHS index element that makes up the general region of the footprint was divided into 8-mers. A query between the 8-mers in the sequence and the sequence bias table was executed, and the sum of the sequence bias values was determined in order to calculate the total bias in the sequence. The same procedure was repeated in the area of the footprint itself. The expected number of cuts within a given footprint region, *e*, was calculated based on the total cuts within the DHS element as well as bias within the footprint itself as well as in the entire region:

$$e = n_{cuts\ in\ total\ region} \cdot \frac{\Sigma(bias\ of\ footprint)}{\Sigma(bias\ of\ region)}$$

D. Determining Differential Binding

In order to investigate whether a transcription factor is bound or not bound, the number of observed cuts within a region was determined by overlapping the ATAC-seq fragments with the DHS footprints and calculating the number of intersections between ATAC-fragments and DHS footprints for each footprint. If the number of observed cuts is greater than or equal to that of expected cuts, the higher number of observed cuts indicates that the region was not protected from the fragmentation of the Tn5 enzyme during ATAC-seq; thus, footprint is unbounded. If the number of observed cuts was less than the number of expected cuts, the footprint is bound because the region faced protection from the Tn5 enzyme, resulting in fewer fragments and indicating the presence of a bound transcription factor.

E. Statistical Analysis

To calculate significance between the observed and the expected values, Anscombe transformation was first applied to normalize the distribution of the values. The transformation turns these values into a unit of standard deviation. Thus, the z-score is calculated by subtracting the

expected value (after transformation) from the observed value for each binding site. The z-score is then converted into a p value. Since many binding site locations (n=159) were tested under a single hypothesis (whether a footprint is bound or not), the p values runs the risk of Type I errors or false positives. To account this, Bonferoni correction was applied: each p value was multiplied by the total number of binding sites.

F. Identifying Potential Binding Sites

Upon calculating the p-value for each footprint binding site, the sites with the most potential for binding were identified for further examination. Sites with the highest binding potential were the footprint sites with significantly lower observed than expected ($p<0.05$) because areas with significantly lower observed than expected suggests the resistance of the footprint to the Tn5 cutting enzyme due to the presence of bound transcription factor.

G. Metaplot Analysis

After identifying the potential binding sites, the profiles for these binding sites were developed to visualize binding within these sites. For a given footprint region, the expected profile was compiled by applying the expected cuts equation for each base in the entire DHS index and mapping out these expected cut numbers in relation to their position on the DHS index element. In addition, the observed profile was developed by mapping where the observed ATAC-fragments fall in the DHS index. Upon calculating the profile, the mean of all the observed and expected profiles were mapped together in order to observe differences between the two categories for the binding sites.

RESULTS

A. Density Plots Identify Enriched Sites in ATAC data

Upon calculating the enrichment ratios, the footprint enrichment data was visualized using density plots to help determine the candidate footprints. As seen in Figure 1, HINFP1/3, the most enriched transcription factor had high frequencies at high log cuts per bases, demonstrating high enrichment. HINFP1/3 is a regulator for DNA methylation and transcription and is over-expressed in CD8+ T-cells [11]. In contrast, ZNF24, the least enriched transcription factor had high frequency at lower log cuts per bases, indicating low enrichment. As for all other transcription factors, the density plots compare the frequency of a footprint to the number of cuts per base. The higher the frequency was at a higher number of log cuts per base, the more enriched a particular footprint was. The plot features "with motif" which maps the relationship for a particular transcription factor and "without motif" which maps the relationship for all other transcription factors as a comparison.

TABLE 1: The ten most enriched transcription factors along with their enrichment ratios

Transcription Factor	Enrichment Ratio
HINFP1/2	0.468354
HINFP1/3	0.457364
ZBTB14	0.408665
KAISO	0.34715
HINFP1/1	0.34569
E2F/4	0.32182
MBD2	0.303204
CENBP	0.292683
AHR	0.275578
GMEB2/3	0.264368

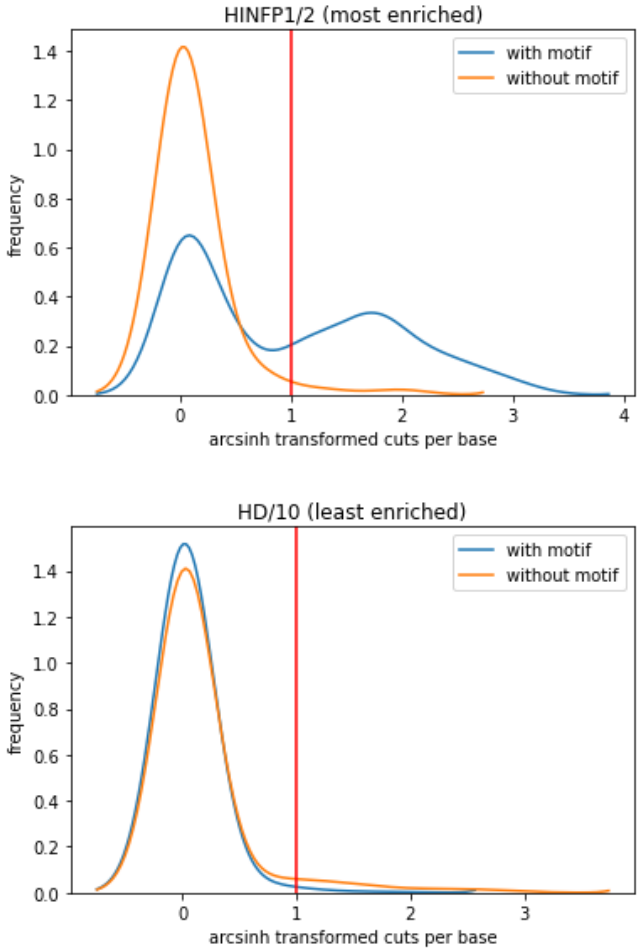

FIGURE 1: Representative density plots of the highest enriched (above) transcription factor (HINFP1/3) and the lowest enriched (below) transcription factor (ZNF24) in CD8+ T cells.

B. *ATAC-seq Analysis Maps Potential Binding Sites*

For each binding site, the expected and observed cuts were plotted in relation to each other, as shown in Figure 2. The scatter plots serve to identify potential binding locations. These areas are defined by areas with lower observed cuts and higher expected cuts, indicating protection from the sequencing enzyme due to a bound transcription factor. Binding potential was measured using the regularized ratio (see Methods) and visualized with a colormap. The twenty five footprints with the lowest regularized ratios were used to further visualize the binding location and the binding behavior using metaplots. On an additional note, in comparison to HINFP1/2 (most enriched footprint), HD/10 has higher regularized ratios, as seen in Figure 2, possibly due to a lower prevalence of binding sites due to lower expression of the footprint.

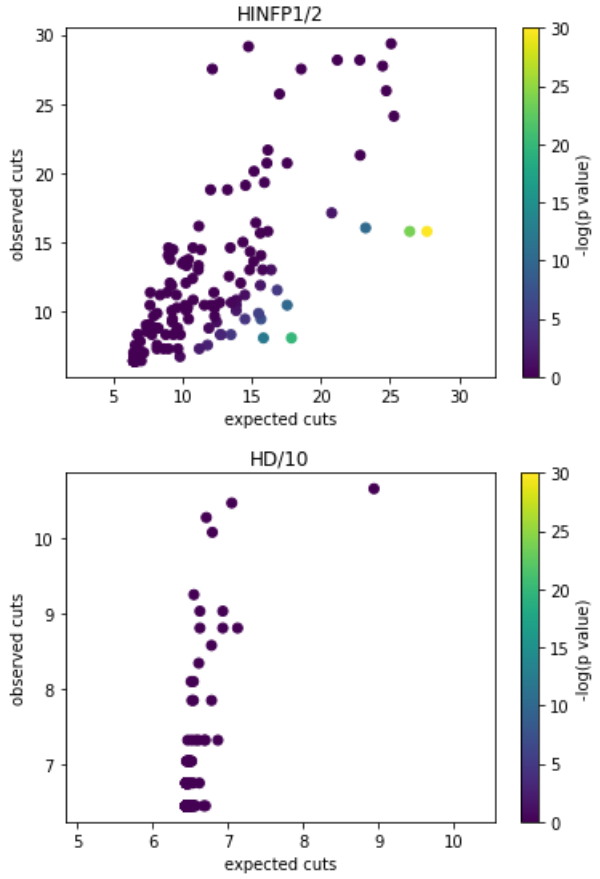

FIGURE 2: Representative scatter plots that map the relationship between observed cuts and expected cuts. Areas with higher binding potential lie in regions with high numbers of expected cuts but low numbers of observed cuts. A color map was developed based on the regularized ratio. More enriched transcription factors (HINFP1/2) have more binding sites with lower regularized ratios compared to less enriched transcription factors (HD/10).

C. *Metaplot Analysis Identifies Bound Locations*

After identifying potential binding locations, the expected and observed cut profiles for the DHS index were mapped. As seen in Figure 3A and Figure 3C, the lower number of cuts observed in the footprint region than expected due to protection from the cutting enzymes provides evidence for transcription factor binding on that site. In less enriched footprints, such as HD/10, the observed cuts are equal to or greater than the expected cuts, demonstrating that the region was not protected from the sequencing enzyme, providing evidence that the region is unbounded (Figure 3B, Figure 3D).

The expected and observed profiles could be further expanded to the other 24 regions with the highest binding potential in HINFP1/2 as seen in Figure 4. Consistent with the results from Figure 3, there is a large difference between the expected and the observed cuts in the enriched footprints in that the expected cuts are much higher than that of the observed, demonstrating that these regions are likely to be bound. For unenriched footprints, the opposite holds: the observed is equal or higher than the number of expected cuts.

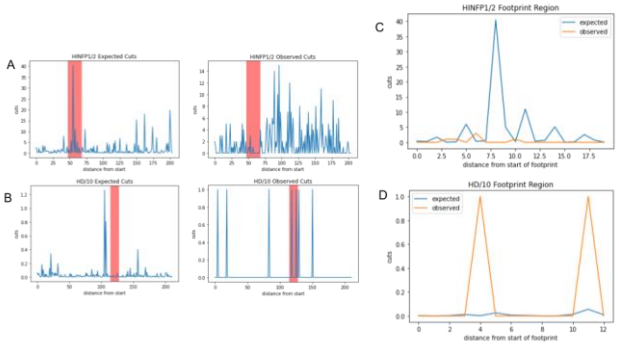

FIGURE 3: Expected and observed cut profiles of a representative HINFP1/2 (A) and HD/10 (B) with the footprint region shaded in red as well as the higher-resolution profile of the footprint itself. Within the enriched footprint (HINFP1/2), the observed cuts seem to be consistently lower than the expected cuts (C). In less enriched footprint (HD/10), the observed cuts are equal or higher than the expected cuts (D).

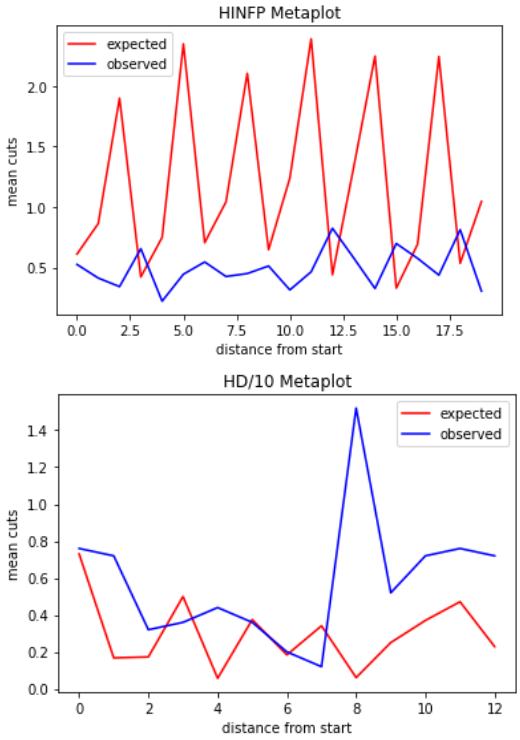

FIGURE 4: Metaplot analysis of the expected and observed cuts of the 25 regions with the most binding potential in HINFP1/2 (left) and HD/10 (right).

DISCUSSION

By comparing the observed cuts and expected cuts in a given footprint region, the results of this study harnesses scATAC-seq to characterize bound and unbound transcription factor sites. Specifically, the paper found that bound sites contained lower observed cuts than expected, resulting from protection by the bound transcription factor within the region. In contrast, unbound sites had an equal number or more observed cuts than expected, indicating that the region showed no protection (allowing the sequencing enzyme to freely fragment the area).

In distinguishing between bound and unbound sites, this study establishes a novel method for studying changes in expression through scATAC-seq. While past studies have identified DNA footprints and accessible sites [2] [3] using DNase-seq, these sequencing techniques lack efficiency to distinguish individual cell types without a large set of materials for input as well as differentiating between sub-populations within a cell type. This study lays the foundation for using scATAC-seq, which allows for identifying specific cell types and subtypes, to study the different transcription factor binding behavior within these specific cell types. In doing so, this study has the potential to influence the study of these common genetic diseases by examining the regulatory context of each cell type and linking mutations to a certain cell type, transcription factor or other regulatory element

Only through identifying these features within the regulatory regions can we fully identify the inner workings behind genetic diseases and develop specific drugs to correct for these mutation-induced differences in the regulatory environment. The potential for single-cell sequencing for understanding common genetic diseases and guiding targeted drug therapies is enormous, and this study serves as a first step for harnessing their potential to understanding and treating these diseases.

FUTURE WORK

This study hopes to expand the data set from CD8+ T cells to a wide variety of other cell types to characterize binding behavior in footprints for other cell types. In expanding the knowledge of transcription factor binding in more cell types, a reference library containing measurements of accessibility and binding for each transcription factor in a particular cell type can be developed. This can serve as a basis for future studies in comparing disease sequences and expression with this library.

It is important to note that DNA footprinting identifies potential sites of interest for binding, but does not confirm and verify binding. While we show that single-cell ATAC data does detect differences in transcription factor binding behavior, further validation is needed to confirm our results. This validation can take the form of ChIP-seq, which is a more direct way of confirming transcription factor binding to validate our results.

Another possibility for this study is to harness and categorize scATAC-seq data from Genome Wide Association Studies (GWAS) for different types of common genetic diseases. Upon categorizing the individual cell types and subtypes, this study will compare the transcription factor binding behavior and accessibility to that of a reference library. From there, differences between the GWAS data and the reference data can be detected, leading to a better understanding of the specific cell types, transcription factors and biological pathways that are relevant in causing diseases. This understanding would be extremely powerful in developing drugs in correcting these differences and alleviating the effects of these diseases.

CONCLUSIONS

This study developed an approach to characterizing the differences in bound and unbound transcription factors by examining ATAC-seq data. This was done in three steps. First, the most enriched footprints were identified in the DHS region by observing the distribution of cuts in relation to a particular footprint. From there, we identified HINFP1/2 as the most enriched footprint and HD/10 as the least enriched footprint. Next, we identified potentially bound sites by computing and comparing the observed and

expected cuts for the candidate footprints. Areas with low observed:expected ratios were selected and used for developing the metaplots. The metaplots compile the profile of the observed and expected to help us better understand the regulatory region specifically. From this compilation, this study identified bound sites as areas with less observed cuts than expected cuts, demonstrating protection from sequencing enzymes caused by transcription factor binding.

ACKNOWLEDGEMENTS

This paper and project would not be possible without the work and support of so many individuals and organizations. I would like to thank my principal investigator, Dr. Eric Lander, and the Lander Lab at the Broad Institute for providing funding and support in making this project possible. I would also like to thank Ms. Ana Lyons who thoroughly reviewed my paper and provided helpful edits. I am thankful for Haydn Bradstreet for providing so much feedback and offering support in reviewing my paper. I am deeply appreciative of the Research Science Institute, the Center for Excellence in Education and the Massachusetts Institute of Technology for providing the opportunity to take part in this incredible project. Lastly, I would like to thank my parents and family for helping and supporting me in nurturing my curiosity as well as my love for science.

REFERENCES

[1] C. P. Fulco, J. Nasser, T. R. Jones, G. Munson, D. T. Bergman, V. Subramanian, S. R. Grossman, R. Anyoha, B. R. Doughty, T. A. Patwardhan, et al. Activity-by-contact model of enhancer–promoter regulation from thousands of crispr perturbations. *Nature genetics*, 51(12):1664–1669, 2019.

[2] J. Vierstra, J. Lazar, R. Sandstrom, J. Halow, K. Lee, D. Bates, M. Diegel, D. Dunn, F. Neri, E. Haugen, et al. Global reference mapping and dynamics of human transcription factor footprints. *bioRxiv*, 2020.

[3] W. Meuleman, A. Muratov, E. Rynes, J. Halow, K. Lee, D. Bates, M. Diegel, D. Dunn,J. Neri, A. Teodosiadis, et al. Index and biological spectrum of accessible dna elementsin the human genome. *BioRxiv*, page 822510, 2019.

[4] C. A. Lareau, F. M. Duarte, J. G. Chew, V. K. Kartha, Z. D. Burkett, A. S. Kohlway, D. Pokholok, M. J. Aryee, F. J. Steemers, R. Lebofsky, et al. Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. *Nature Biotechnology*,37(8):916–924, 2019.

[5] L. S. Ludwig, C. A. Lareau, E. L. Bao, S. K. Nandakumar, C. Muus, J. C. Ulirsch, K. Chowdhary, J. D. Buenrostro, N. Mohandas, X. An, et al. Transcriptional states and chromatin accessibility underlying human erythropoiesis.*Cell reports*, 27(11):3228–3240, 2019.

[6] F. Aguet, A. N. Barbeira, R. Bonazzola, A. Brown, S. E. Castel, B. Jo, S. Kasela, S. Kim-Hellmuth, Y. Liang, M. Oliva, et al. The gtex consortium atlas of genetic regulatory effects across human tissues. *BioRxiv*, page 787903, 2019.

[7] E.-W. Yang, J. H. Bahn, E. Y.-H. Hsiao, B. X. Tan, Y. Sun, T. Fu, B. Zhou, E. L. Van Nostrand, G. A. Pratt, P. Freese, et al. Allele-specific binding of rna-binding proteins reveals functional genetic variants in the rna. *Nature communications*, 10(1):1–15, 2019.

[8] T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, et al. Ncbi geo: archive for functional genomics data sets—update. *Nucleic acids research*, 41(D1):D991–D995, 2012.

[9] A. R. Quinlan and I. M. Hall. Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010.

[10] Z. Li, M. H. Schulz, T. Look, M. Begemann, M. Zenke, and I. G. Costa. Identificationof transcription factor binding sites using atac-seq.Genome biology, 20(1):45, 2019.[11] A. D. Rouillard, G. W. Gundersen, N. F. Fernandez, Z. Wang, C. D. Monteiro, M. G. McDermott, and A. Ma'ayan. The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database*, 2016, 2016.