BIOINF 593, 2023, 1–6
doi: DOI HERE
4 December 2023
Paper

Predicting Gene Expression Levels From Chromatin Structure Using Graph Neural Networks

Sion Kim,[1] Aakash Patel,[2] Sawan Patel,[2] Shuyi Xie[1] and Yunxuan Xie[3]

[1]Bioinformatics, University of Michigan, 48109, MI, USA, [2]Computer Science and Engineering, University of Michigan, 48109, MI, USA and [3]Pharmacy, University of Michigan, 48109, MI, USA

Abstract

Advancements in single-cell multi-omic sequencing has enabled new innovative approaches for deciphering the intricate relationship between chromatin structure and gene expression. In this paper, we present a novel application of Graph Neural Networks (GNNs) to predict gene expression patterns based on RNA sequencing using Hi-C chromatin structure data. Leveraging the inherent graph representation of chromatin interactions, we employ Graph Convolutional Networks (GCNs) and Graph Attention Networks (GATs) to model the complex dependencies among genes. Our results demonstrate the ability of GNNs to capture the spatial organization and regulatory dynamics within the genome. In particular, we consider an inter-chromosomal mouse gene network responsible for ectoderm and mesoderm differentiation during embryonic development, and showcase our model's capability to predict coexpression of these genes with high accuracy. This work establishes a powerful framework for integrating chromatin structure data into predictive models, offering a deeper understanding of the regulatory mechanisms governing gene expression and paving the way for advancements in personalized genomics and therapeutic interventions.

Key words: single cell RNA-Seq, single cell Hi-C, HiRES, Graph Convolutional Network (GCN), Graph Attention Network (GAT)

Introduction

In multicellular organisms, distinct cell types possess genomes that are virtually the same, yet they exhibit significant differences in structure and function. Establishing varied cell identities throughout an organism's development entails regulating the gene expression, which includes the three-dimensional (3D) spatial interactions of chromatin in the nucleus [8]. For example, the locus control region (LCR) interacts with β-type globin genes in a stage-specific way during development, regulating transcription in erythroid cells [5]. Disruptions in the 3D organization of the genome have been linked to the emergence of numerous diseases, such as cancer, and mutations in IDH contribute to the development of gliomas by altering the chromosomal topology and enabling abnormal regulatory interactions that lead to the activation of oncogenes [7]. Despite this, the dynamic nature of the relationship between 3D genomic structure and gene expression is still a subject of debate. Substantial alterations in the 3D genome structure can be induced by the selective degradation of crucial regulatory proteins like CCCTC-binding factor (CTCF) or cohesin, yet these changes have only a slight effect on gene expression [13, 15]. In the embryos of *Drosophila*, while various cell types exhibit marked variations in gene expression, differences in chromatin structure are relatively minor [9].

RNA sequencing (RNA-Seq) is a sequencing technique to catalog and quantify gene expression on the transcriptome level across cells. Single-cell RNA sequencing (scRNA-Seq) extends these capabilities to the individual cell level. ScRNA-Seq generally has several steps: isolation of single cells, cell lysis, reverse transcription of RNA into cDNA, cDNA amplification by polymerase chain reaction (PCR), cDNA library generation, high-through put DNA sequencing and data analysis [18]. It is frequently used for cell type characterization and inferring gene regulatory networks.

Hi-C captures the 3D chromatin conformation, which helps to unravel how spatial organization of genomes and interactions between genomic regions affect gene expression and functions of cells. Hi-C is based on chromosome conformation capture (3C) assays, with recent extensions involving 4C (chromosome conformation capture-on-chip/circular chromosome conformation capture), and 5C (chromosome conformation capture carbon copy). Single-cell Hi-C (scHi-C) extends these capabilities to capturing chromatin interactions of individual cell rather than in bulk, thus enabling deciphering of cell-to-cell variation in chromosome structures. The process of scHi-C includes isolation of single cells, cross-linking of DNA-DNA interactions bridged by proteins using formaldehyde, cell lysis, chromatin digestion, ligation of proximal ends, reversal of crosslinking, DNA purification, library preparation, sequencing. Subsequently, contact map are generated by binning [2].

Despite this, naively combining independent scRNA-seq and scHi-C data has a significant drawback in its inability to control for variation across cell stages and cell sub-types. Development in single-cell multi-omics sequencing techniques have facilitated the simultaneous profiling of the transcriptome alongside other biological features, such as protein levels, DNA methylation and chromatin accessibility. These advancements have significantly enhanced our comprehension of how cells determine their fate at the molecular level [1, 3]. Among these single-cell multi-omics techniques, a long-lasting gap has been the absence of a method to simultaneously obtain chromatin conformation data.

In 2023, Xing et al. [12] reported a new technique, HiRES (Hi-C and RNA-seq employed simultaneously), allowing the simultaneous profiling of transcriptome and 3D genome at the single-cell level (Figures 1, 2). The assay is detailed in [4]. After cell fixation and permeabilization, in-situ reverse transcription was performed, followed by a 3C procedure. Digestion and ligation were conducted, and cells were flow sorted and underwent quasi-linear amplification by several cycles of annealing and looping in a single-tube setup. After sequencing, the cDNA sequences were identified using the mRNA-specific tag added during the reverse transcription. A total of 432 single cells from adult mouse brain and 7716 single cells from F1 hybrid mouse embryos (C57BL/6J × CAST/EiJ) collected between embryonic day 7.0 (E7.0) and E11.5 were used to explore the relationship between genome organization and gene expression. The authors also suggested that rewiring of chromatin interactions prior to gene expression typically occurs in regions of active chromatin. For genes situated within repressed chromatin areas, condensed chromatin relaxation takes place before transcription activation. The pseudo-temporal relationship between chromatin conformation and transcriptome was also investigated, via pseudotime inference and residual analysis measuring the difference of gene-associated differential interactions (DI) and gene expression level.

In this work, we aim to extend the HiRES method to predict gene expression levels and co-expressed genes using Hi-C data by leveraging powerful deep learning tools including graph convolution and graph attention networks. We demonstrate the effectiveness of our method via a case study using several genes involved in ectoderm and mesoderm differentiation across multiple chromosomes, showing that our model can predict the coexpression of these genes with high accuracy.

Data Preprocessing

Hi-C

Hi-C data (.pairs format) was retrieved from Gene Expression Omnibus (accession no. GSE223917) and converted to .hic via *Juicer tools pre* with parameters `mm10 -r 100000,500000,1000000`. This corresponds to reference genome mm10, resolutions at 100Kb, 500Kb and 1Mb, and default normalization vectors (`VC`, `VC_SQRT`, `KR`, `SCALE`). Using the specified MLP model (Section 3.1), we performed a grid-search across different Hi-C parameters and optimized to observed counts with KR normalization, which was used for all subsequent analysis.

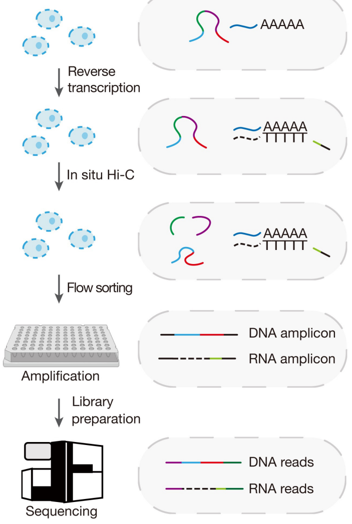

Fig. 1: HiRES enables the simultaneous profiling of transcriptomes and 3D genome structure. The workflow of HiRES is shown. Adapted from [12]

.

RNA-seq

RNA-seq data was retrieved from Gene Expression Omnibus (accession no. GSE223917), corresponding to two preprocessed files [12]:

1. GSE223917_HiRES_brain.rna.umicount.tsv.gz
2. GSE223917_HiRES_emb.rna.umicount.tsv.gz

Each dataset was subsequently merged with:

1. metadata (GSE223917_HiRES_brain_metadata.xlsx, GSE223917_HiRES_emb_metadata.xlsx, respectively) in order to generate datasets across different cell stages (e.g., E7) and
2. a genebank file generated from *BioMart* (mm39[1] with attributes for gene start, gene end, chromosome, gene name, gene symbol and *MGI* symbol).

Genes with a match to either gene name, gene symbol, or *MGI* symbol were used as the final set, resulting in a 96% (48, 461 / 50, 463) match with the genes in the RNA-seq data.

Neural Network Models

In recent years, neural networks have emerged as powerful tools to unravel the complexities of biological data. Deep learning has been leveraged to extract meaningful patterns from genomic, proteomic, and other omic data [6]. Neural networks are capable of learning highly nonlinear patterns in the input data, making them particularly suited for complex tasks such as prediction of gene expression. This paper explores three categories of neural network

[1] Despite the disparity of reference genome between RNA-seq and Hi-C data (mm39 vs mm10), there is no effect on the model performance. This is due to the units for the training and testing data being discrete genes rather than fixed bin sizes.

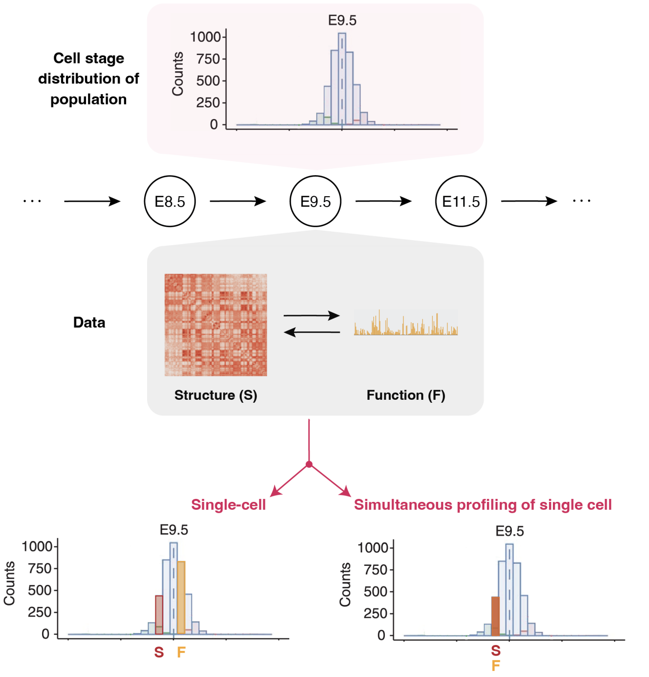

Fig. 2: Single-cell Hi-C (structure) and RNA-seq (function) studies have previously been used to investigate embryonic development [16, 14], but suffered from a lack of temporal precision between samples: single-cell Hi-C or RNA-seq samples may be drawn from two different time points (left), whereas simultaneous profiling of single-cell studies (right) controls this variability and ensures that data is identically and independently distributed.

models for this task: multilayer perceptrons, graph convolutional networks, and graph attention networks.

Multilayer Perceptrons

The multilayer perceptron (MLP) is one of the simplest of neural network architectures, and is inspired by the interconnected nature of biological neurons in the brain. A MLP consists of multiple layers, where each layer performs a linear transformation of the input followed by a nonlinear activation function. The MLP architecture used in this work is depicted in Figure 3 (left). We use five fully-connected layers, with ReLU activation and Dropout following each but the last layer. Since the features in an MLP must be one-dimensional, we flatten the adjacency matrix as input to the model. The model outputs a N_d size vector for each sample, where N_d is the number of genes we would like to predict.

Graph Convolution Networks

Graph Convolutional Networks (GCNs) are another type of neural network specifically designed to handle structured data in the form of graphs, where nodes represent Hi-C contact regions and edges represent the interaction between regions. GCNs excel in capturing spatial interactions between features by aggregating information from neighboring nodes for each node in the graph. Each graph convolutional layer computes a weighted sum of features from neighboring nodes, where the weights are determined by the interactions between nodes (edges in the graph). For each feature vector $h_i^{(l)}$ of node i at layer l, the graph convolutions compute

$$h_i^{(l+1)} = \sigma\left(\sum_{j \in \mathrm{nbhd}(i)} \frac{1}{c_{ij}} W^{(l)} h_j^{(l)} \right) \tag{1}$$

where nbhd(i) denotes the neighboring nodes of node i, $W^{(l)}$ is the weight matrix at layer l, σ is an activation function, and c_{ij} is a normalization constant [11].

Our GCN architecture is shown in Figure 3 (left). We employ five graph convolutional layers to extract node embeddings from the input graph, followed by our previous MLP model to perform the regression.

Hyperparameter	MLP	GCN	GAT
Batch Size	64	8	8
Learning Rate	5×10^{-4}	1×10^{-4}	1×10^{-4}
Hidden Layers	4	5	4
Dropout Prob	0.5	0.5	0.5

Table 1. Hyperparameter values used for the different models.

Graph Attention Networks

Inspired by the concept of attention in natural language processing, Graph Attention Networks (GATs) utilize the powerful attention mechanism to selectively weigh the importance of different neighbors when aggregating information from the nodes of a graph [17]. The attention coefficients α_{ij} between nodes i and j are computed as

$$\alpha_{ij} = \frac{\exp(\sigma(a^\top [Wh_i || Wh_j]))}{\sum_{k \in \mathrm{nbhd}(i)} \exp(\sigma(a^\top [Wh_i \mid Wh_k]))} \tag{2}$$

where h_i is the feature representations of node i, W is the weight matrix, a is a learnable weight vector, $||$ denotes concatenation, and σ is an activation function. This mechanism allows nodes to selectively attend to relevant neighbors during the information aggregation process, allowing for more flexible and adaptive modeling of complex relationships within graph-structured data. This adaptability is particularly beneficial in scenarios like biological networks, where nodes may have varying degrees of importance and connectivity.

Our GAT architecture is identical to the GCN, only with graph attention layers instead of graph convolution. Because our Hi-C data has no features other than the graph structure, for the graph-based models we use the degree of each node as its feature. We perform a grid search optimization to guide model architecture and hyperparameter selection. The chosen hyperparameters are given in Table 1.

Experiments

Predicting Frequently Expressed Genes

To guide subsequent model development, we first experiment with training simple models on chromosomes 6 and 17 to predict the 10 most frequently expressed genes on those chromosomes. We train each model using the Hi-C data at 1Mb resolution for 50 epochs. A comparison of the performance for the different architectures is shown in Figure 4. The GCN and GAT models exhibit more stability in training and better performance than the MLP, but

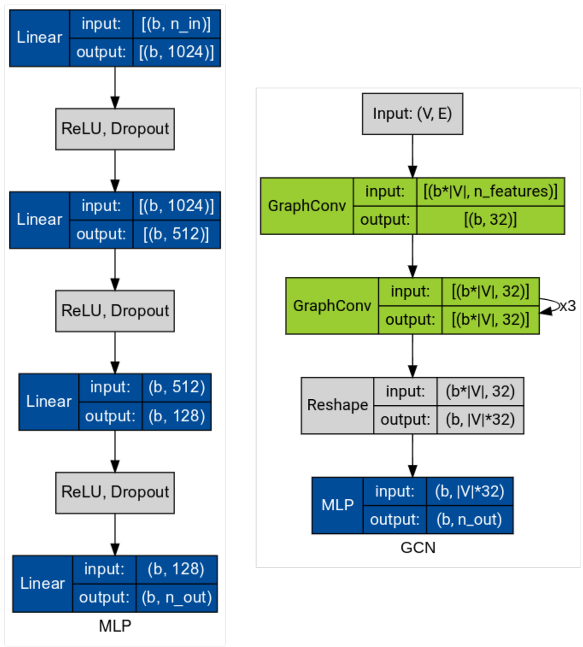

Fig. 3: Left: the MLP model architecture. We use five fully-connected layers with ReLU activation and dropout between each layer. Right: the GCN model architecture. We use five graph convolution layers to get node embeddings followed by the MLP model for regression. The GAT model is similar, with graph attention layers instead of convolution.

are roughly equal to each other. We suspect this similarity is due to the node structure in our graphs being relatively static, mitigating one of the biggest advantages GATs have over GCNs. We also experimented with normalizing the Hi-C data matrix and using different resolutions. However, these did not improve the performance, and so were not considered further.

Developmental Layers: Ectoderm and Mesoderm

Our initial experiments have shown that there are structural patterns in the Hi-C data that can be exploited to predict expression of arbitrary genes. We now turn our attention to predicting the expression of a network of genes that encode for a particular biological function. In particular, we aim to predict a subset of genes that are responsible for encoding ectoderm and mesoderm differentiation markers. We used the *Mouse Genome Informatics* repository of Gene Ontology Annotations to query for ectoderm and mesoderm specific genes.

Since our aim is to investigate diverging developmental layers across embryonic development, we combined the sets of genes together and generated an interaction confidence network on String-DB 5. Interestingly, sub-graphs with high edge confidence typically spanned chromosomes, such as *Oct4* (chr17), *Nanog* (chr6), and *Sall1* (chr8) [10], which motivated us to train our models using multiple chromosomes for greater flexibility in capturing meaningful inter-chromosomal interactions. Chromosome 17 was selected for *Pou5f1* (*Oct4*), a master regulator of embryonic stem cells, as well as *Six2*, involved in embryonic morphogenesis and organ development, chromosome 6 for the Hox gene clusters, implicated in spatial organization along the anteroposterior body axis and *Nanog*, and chromosome 11 for *Wnt3a*, which plays an important role in mesoderm formation and the development of the notochord.[2] A list of the set of genes used for model training is given in Table 2.

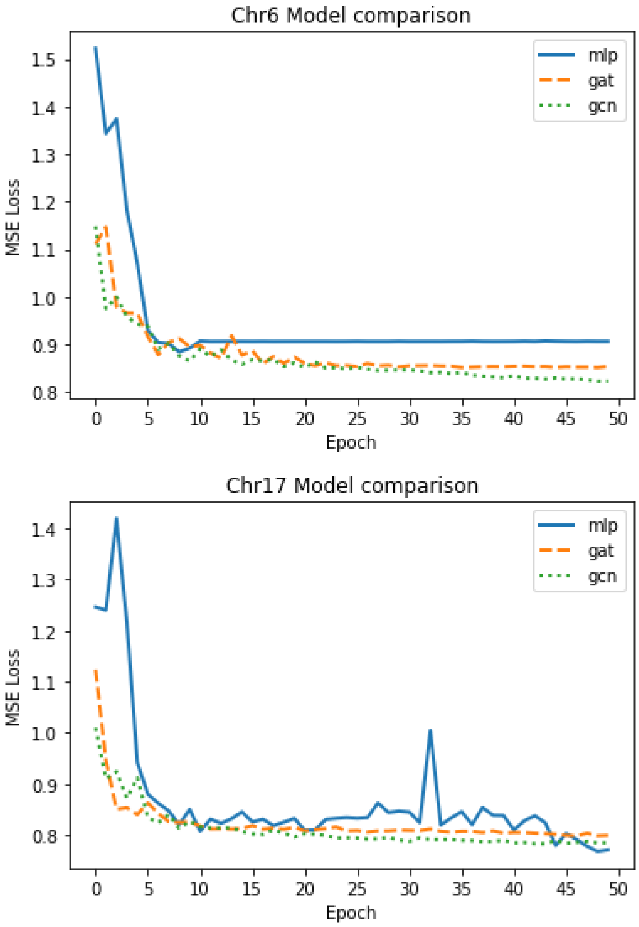

Fig. 4: Performance of different architectures on predicting top 10 genes on chromosomes 6 and 17.

For this experiment, we train a model on Hi-C data with block matrices for chromosome 2, 11, and 17 and predict the expression levels of 24 genes across the three chromosomes. We train both a GAT and a GCN on this combined data. The validation loss for both models is shown in Figure 6. The GCN model achieves a test loss of 0.16676 while the GAT is at 0.16904. Figure 7 highlights randomly selected cells from day 7, day 8.5, and 11.5. The combined Hi-C matrices for each cell are shown in the top row, and the true RNA-Seq values along with the predictions from the GCN model are shown below. There is overall good correspondence between labels and prediction. However, across time, the model predictions are similar, unable to capture the

[2] Including chromosome 3 was tempting due to the well-known interaction between *Sox2* and *Oct4*, but was left out as *Sox2* was not listed as markers of ectoderm or mesoderms on *MGI*.

Axin1	Axin2	Grb2	Klrg1	Fgf6	Vps52
Nanog	Jup	Lhx1	Fgf18	Pou5f1	Tcf7
Vps53	Kremen1	Wnt3a	Hoxa11	Nog	Lrp6
Kdm6b	Vps54	Smo	Fgf23	Six2	Nf2

Table 2. Genes involved in differentiation of ectoderm and mesoderm. Genes were taken from *MGI* Gene Ontology Browser and filtered for chromosomes 6, 11, and 17 (see 5) for full list.

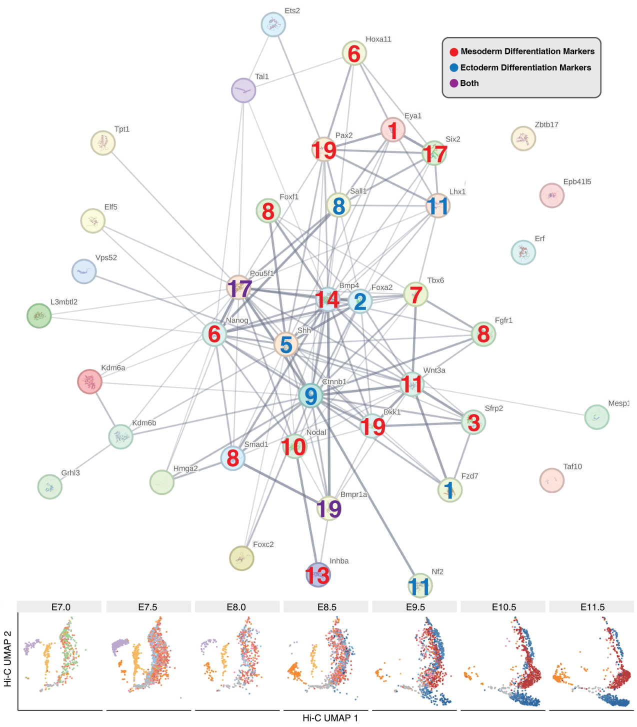

Fig. 5: (Top) Output of *String-DB* from marker genes of mesodermal and ectodermal differentiation processes. Marker genes were retrieved from *MGI* Gene Ontology Browser 2 with searches for **mesodermal cell differentiation** (GO:0048333) and **ectoderm development** (GO:0048333) yielding 36 and 19 markers, respectively. Nodes are overlaid with chromosome number and coloured according to functional groups identified from [12]: red and blue represent mesoderm and ectoderm differentiation markers, respectively. (Bottom) UMAP plots across time from Figure 2C [12].

temporal dependencies (e.g., consider *Smo* expressing in E7 but suppressed in E8.5). This may be due to the sparsity of the single-cell Hi-C contact maps despite a low resolution of 1Mb. This also highlights a limitation with our current model in that time is not explicitly taken into account. The training process pools together all samples and so the model effectively learns a conserved Hi-C to RNA-seq mapping rather than one which can effectively capture the progression of embryonic development. A possible approach may be to utilize a recurrent neural network.

Conclusion

Our work demonstrates the potential of using Graph Neural Networks for predicting gene expression from chromatin structure.

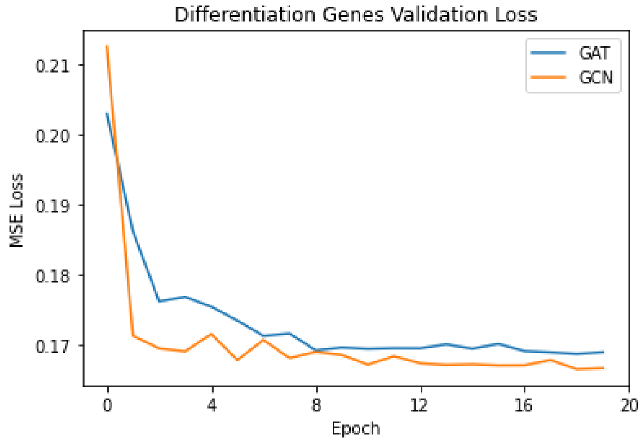

Fig. 6: Validation loss for GCN and GAT models for predicting differentiation genes across chromosomes 2, 11, and 17.

By exploiting the graph structure inherent in Hi-C data, powerful GNNs like Graph Convolutional Networks and Graph Attention Networks exhibit a remarkable capacity to capture complex relationships and dependencies among genes. Through their ability to consider the connectivity patterns within gene regulatory networks, GNNs offer a nuanced approach to modeling the multifaceted dynamics of gene expression. The adaptability of attention mechanisms in GATs and the hierarchical learning capabilities of GCNs enable these models to discern structural patterns that might be elusive to traditional approaches. Despite the promise of graph-based models in capturing structural interactions, fully leveraging time-series data requires a better suited architecture with explicit modelling of time in order to capture the dynamics of chromatin structure and function. Nevertheless, the integration of GNNs into the realm of predicting gene expression marks a significant stride toward unraveling the complexities of the genomic landscape and holds immense promise for advancing our understanding of biological systems and informing personalized therapeutic strategies.

Contribution

S.K., S.X., and Y.X. developed the idea and provided biological framework. S.K. and A.P. processed data. A.P. and S.P. developed, debugged, and benchmarked the NN models. S.P. performed the exploratory analysis. All members contributed to writing the report and preparing the presentation.

References

1. C. Angermueller, S. J. Clark, H. J. Lee, I. C. Macaulay, M. J. Teng, T. X. Hu, F. Krueger, S. A. Smallwood, C. P. Ponting, T. Voet, et al. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nature methods*, 13(3):229–232, 2016.
2. H. Belaghzal, J. Dekker, and J. H. Gibcus. Hi-c 2.0: An optimized hi-c procedure for high-resolution genome-wide mapping of chromosome conformation. *Methods*, 123:56–65, 2017.

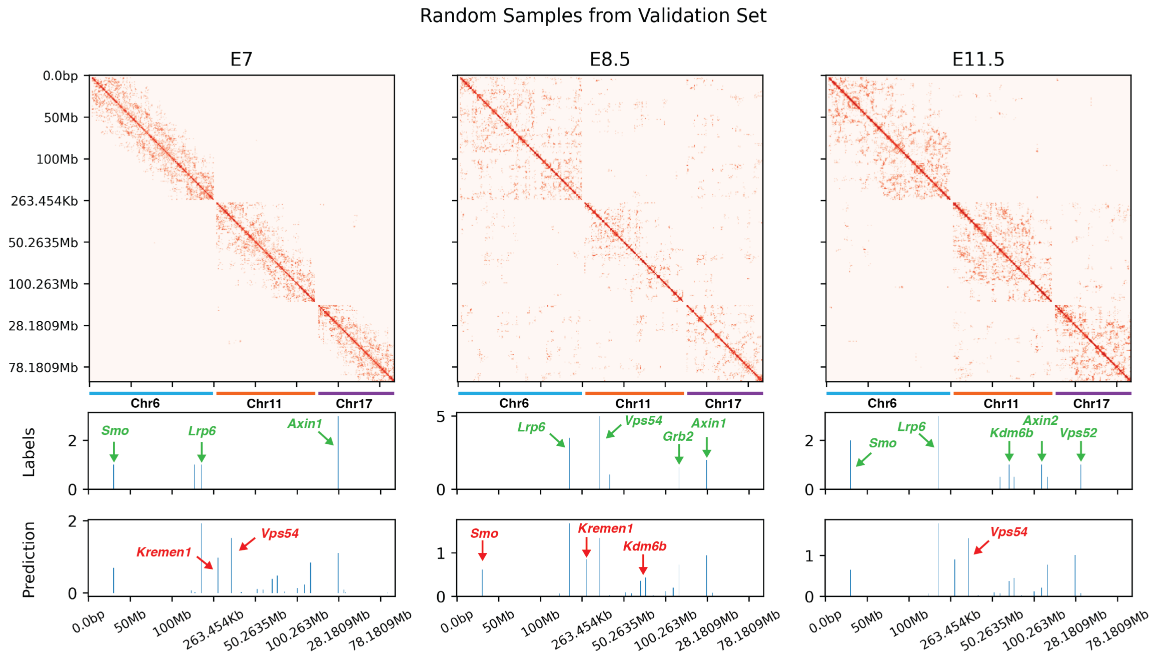

Fig. 7: Random samples (Hi-C, Top and RNA-seq, Middle) were taken from the validation set from E7, E8.5, and E11.5. Predicted outputs from the trained GCN model are shown below. Hi-C map is log normalized for visualization purposes.

3. J. Cao, D. A. Cusanovich, V. Ramani, D. Aghamirzaie, H. A. Pliner, A. J. Hill, R. M. Daza, J. L. McFaline-Figueroa, J. S. Packer, L. Christiansen, et al. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science*, 361(6409):1380–1385, 2018.
4. Y. Chen, H. Xu, Z. Liu, and D. Xing. Simultaneous profiling of chromosome conformation and gene expression in single cells. *Bio-protocol*, 13(22), 2023.
5. W. Deng, J. W. Rupon, I. Krivega, L. Breda, I. Motta, K. S. Jahn, A. Reik, P. D. Gregory, S. Rivella, A. Dean, et al. Reactivation of developmentally silenced globin genes by forced chromatin looping. *Cell*, 158(4):849–860, 2014.
6. G. Eraslan, Ž. Avsec, J. Gagneur, and F. J. Theis. Deep learning: new computational modelling techniques for genomics. *Nature Reviews Genetics*, 20(7):389–403, 2019.
7. W. A. Flavahan, Y. Drier, B. B. Liau, S. M. Gillespie, A. S. Venteicher, A. O. Stemmer-Rachamimov, M. L. Suvà, and B. E. Bernstein. Insulator dysfunction and oncogene activation in idh mutant gliomas. *Nature*, 529(7584):110–114, 2016.
8. E. E. Furlong and M. Levine. Developmental enhancers and chromosome topology. *Science*, 361(6409):1341–1345, 2018.
9. E. Ing-Simmons, R. Vaid, X. Y. Bing, M. Levine, M. Mannervik, and J. M. Vaquerizas. Independence of chromatin conformation and gene regulation during drosophila dorsoventral patterning. *Nature genetics*, 53(4):487–499, 2021.
10. E. Karantzali, V. Lekakis, M. Ioannou, C. Hadjimichael, J. Papamatheakis, and A. Kretsovali. Sall1 regulates embryonic stem cell differentiation in association with nanog. *Journal of Biological Chemistry*, 286(2):1037–1045, 2011.
11. T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.
12. Z. Liu, Y. Chen, Q. Xia, M. Liu, H. Xu, Y. Chi, Y. Deng, and D. Xing. Linking genome structures to functions by simultaneous single-cell hi-c and rna-seq. *Science*, 380(6649):1070–1076, 2023.
13. E. P. Nora, A. Goloborodko, A.-L. Valton, J. H. Gibcus, A. Uebersohn, N. Abdennur, J. Dekker, L. A. Mirny, and B. G. Bruneau. Targeted degradation of ctcf decouples local insulation of chromosome domains from genomic compartmentalization. *Cell*, 169(5):930–944, 2017.
14. N. Ranisavljevic, M. Borensztein, and K. Ancelin. Understanding chromosome structure during early mouse development by a single-cell hi-c analysis. *Epigenetic Reprogramming During Mouse Embryogenesis: Methods and Protocols*, pages 283–293, 2021.
15. S. S. Rao, S.-C. Huang, B. G. St Hilaire, J. M. Engreitz, E. M. Perez, K.-R. Kieffer-Kwon, A. L. Sanborn, S. E. Johnstone, G. D. Bascom, I. D. Bochkov, et al. Cohesin loss eliminates all loop domains. *Cell*, 171(2):305–320, 2017.
16. J. Shi, Q. Chen, X. Li, X. Zheng, Y. Zhang, J. Qiao, F. Tang, Y. Tao, Q. Zhou, and E. Duan. Dynamic transcriptional symmetry-breaking in pre-implantation mammalian embryo development revealed by single-cell rna-seq. *Development*, 142(20):3468–3477, 2015.
17. P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.
18. M. You, R. Rong, Z. Zeng, H. Li, X. Xia, and D. Ji. Single-cell rna sequencing: A new opportunity for retinal research. *Wiley Interdisciplinary Reviews: RNA*, 12(5):e1652, 2021.