Using Machine Learning to Predict RNA Pol II Interactions of Metastatic Prostate Cancer

Ahmed Abbas[1], Chengcheng Liu[2], Michael Q. Zhang[2], Ram S. Mani[1,3,4]

[1]Department of Pathology, UT Southwestern Medical Center, Dallas, TX 75390, USA

[2]Department of Biological Sciences, Center for Systems Biology, The University of Texas at Dallas, Richardson, TX 75080, USA

[3]Department of Urology, UT Southwestern Medical Center, Dallas, TX 75390, USA

[4]Harold C. Simmons Comprehensive Cancer Center, UT Southwestern Medical Center, Dallas, TX 75390, USA

Background

The three-dimensional (3D) genome organization directly impacts diverse nuclear processes such as transcription, DNA repair, and replication. Therefore, it is crucial to understand how the distal regulatory elements (in the linear genomic distance) interact in 3D space. Several sequencing-based and imaging-based experimental methods have been developed in the last two decades to study the 3D chromatin organization [1]. High-throughput chromosome conformation capture (Hi-C) [2] and chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) [3] are some of the commonly used methods to study 3D genome organization. These experimental methods are costly and time-consuming. Thus, we sought to use machine learning to predict chromatin interactions.

Methods

We propose a gradient-boosting regression model [4] to predict the interaction strength between genes with their distal enhancers. Both RNA-seq and ATAC-seq experiments require a much smaller number of cells than those needed to perform a ChIA-PET experiment, which makes them suitable to use with human biopsies [5]. Thus, we used the normalized read count of ATAC-seq data for both anchor peaks and the FPKM expression values of the genes overlapping with them as inputs to our regression model. In addition, we used the genomic distance between the two anchor peaks as an additional input [6]. We trained the model on the high-quality data of the GM12878 ENCODE cell line [7].

Results

Our model can accurately predict interactions of different prostate cancer types

In [8], the authors provided the RNA-seq and ATAC-seq data of several prostate cancer (PCa) types. We focused on the samples representing the AR+ (ADPC) and neuroendocrine (NEPC) prostate cancer types. We used the RNA-seq and ATAC-seq data to predict RNA Pol II-associated interactions between the ATAC-seq peaks with genomic distance bigger than 50 Kb [9] and less than 1 Mb (a typical topologically associating domain (TAD) size [10]). To validate the accuracy of the predicted interactions, we first calculated the differentially expressed genes that are upregulated in each of the two PCa types using DESeq2 [11] (Fig. 1A). We then measured the strongest predicted interactions overlapping with the promoters of the upregulated genes in each of the two types. We found that for upregulated genes in ADPC, the predicted interactions overlapping with their promoters in the ADPC sample are significantly stronger than those in the NEPC sample (Fig. 1B). Likewise, the upregulated genes in the NEPC sample have stronger interactions associated with them in the NEPC sample than in the ADPC one (Fig. 1C).

Conclusion

We proposed a method for predicting RNA Pol II-associated interactions between genes and distal regions regulating them. It can be used to predict interactions of different cell types in a tissue using pseudobulk RNA-seq and ATAC-seq data obtained from the single-cell data of PCa biopsies.

Funding Acknowledgements
We thank the Prostate Cancer Foundation for granting Ahmed Abbas the Young Investigator Award in 2024. RSM acknowledges funding support from the NCI/NIH (grant R01CA245294), CPRIT Individual Investigator Research Award (RP230382), and US Department of Defense Breakthrough Award (W81XWH-21-1-0114).

Conflicts of Interest

No

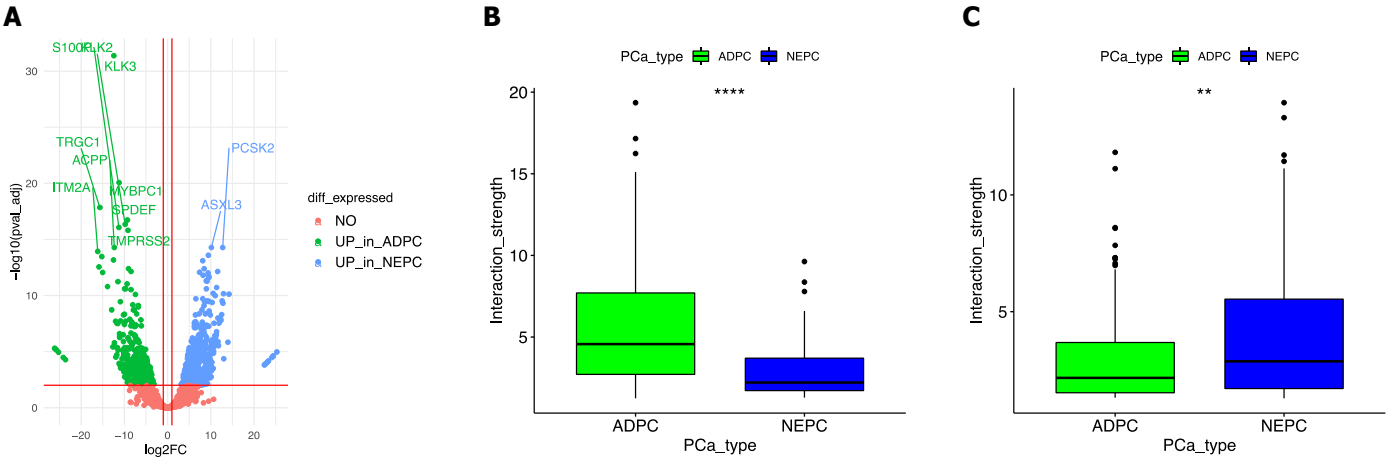

Figure 1: Predicted interactions associated with upregulated genes have higher interaction strength. (A) Volcano plot showing the genes upregulated in ADPC and NEPC, respectively. (B) Predicted interactions associated with the promoters of genes upregulated in the ADPC samples have significantly higher interactions' strength in the ADPC sample vs. the NEPC one. (C) Predicted interactions associated with the promoters of genes upregulated in the NEPC samples have significantly higher interactions' strength in the NEPC sample vs. the ADPC one. ****: p-value < 0.0001, **: p-value < 0.01, Wilcoxon rank sum test.

References

1. Cavalli, G., *Understanding 3D genome organization by multidisciplinary methods.* Nature Reviews Molecular Cell Biology, 2021. **22**(8): p. 511-528.
2. Lieberman-Aiden, E., et al., *Comprehensive mapping of long-range interactions reveals folding principles of the human genome.* science, 2009. **326**(5950): p. 289-293.
3. Fullwood, M.J., et al., *An oestrogen-receptor-α-bound human chromatin interactome.* Nature, 2009. **462**(7269): p. 58-64.
4. Wade, C. and K. Glynn, *Hands-On Gradient Boosting with XGBoost and scikit-learn: Perform accessible machine learning and extreme gradient boosting with Python*. 2020: Packt Publishing Ltd.
5. Xu, D., et al., *Recapitulation of patient-specific 3D chromatin conformation using machine learning.* Cell reports methods, 2023. **3**(9).
6. Abbas, A., et al., *ChIPr: accurate prediction of cohesin-mediated 3D genome organization from 2D chromatin features.* Genome Biology, 2024. **25**(1): p. 1-27.
7. Consortium, E.P., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57.
8. Cejas, P., et al., *Subtype heterogeneity and epigenetic convergence in neuroendocrine prostate cancer.* Nature communications, 2021. **12**(1): p. 5775.
9. Belokopytova, P.S., et al., *Quantitative prediction of enhancer–promoter interactions.* Genome research, 2020. **30**(1): p. 72-84.
10. Dixon, J.R., et al., *Topological domains in mammalian genomes identified by analysis of chromatin interactions.* Nature, 2012. **485**(7398): p. 376-380.
11. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome biology, 2014. **15**: p. 1-21.