Inferring Nucleosome Position Shifting with Single-cell Data from Prostate Cancer Neuroendocrine Transformation

James Bole Pan*
Department of Computer Science
Columbia University
New York, NY 10027
bole.pan@columbia.edu

Cassandra Burdziak
Computational and Systems Biology Program
Memorial Sloan Kettering Cancer Center
New York, NY 10065
burdziac@mskcc.org

Andrew Blumberg
Department of Mathematics
Columbia University
New York, NY 10027
andrew.blumberg@columbia.edu

Dana Pe'er
Computational and Systems Biology Program
Memorial Sloan Kettering Cancer Center
New York, NY 10065
peerd@mskcc.org

Abstract

Cancer progression is marked by the dysregulation of chromatin states, yet the precise mechanisms and dynamics of ATP-dependent chromatin remodelers during this process remain poorly understood. In this study, we investigate the role of chromatin remodelers in prostate cancer's neuroendocrine transformation using a single-cell multiomics dataset derived from mouse models. Employing a graph-based clustering approach, we grouped cells with similar phenotypes into distinct clusters and used a statistical testing-based algorithm to extract nucleosome positions across a substantial number of regulatory regions within all identified clusters. Dimensionality reduction analysis applied to this data revealed significant remodeling of the chromatin landscape during the transformation. Feature selection identified several key regulatory regions, showcasing two primary nucleosome dynamics: 1) initial eviction followed by eventual reassembly, and 2) shifting of positions. These findings enhance our understanding of the epigenetic alterations occurring during cancer progression. Applying this study's approach on expanded datasets could offer deeper insights and potentially guide the development of therapeutic strategies targeting the epigenome in prostate cancer and other aggressive cancer forms.

1 Introductions

The epigenetic dysregulation of chromatin states has been identified as a key hallmark of cancer (1). Chromatin states, which involve the packaging of DNA with both histones and non-histone proteins, are crucial in establishing and maintaining cell identities. Several mechanisms can alter chromatin states and lead to their dysregulation: DNA modification, histone modification, incorporation of histone variants, pathways mediated by noncoding RNA, and chromatin remodelers (2). This research focuses primarily on chromatin remodelers.

*Geometric Data Analysis Final Project, Spring 2024, Prof. Andrew Blumberg.
This project is adapted from research work supervised by Dr. Cassandra Burdziak and Dr. Dana Pe'er

Chromatin remodelers are ATP-dependent protein complexes involved in the exchange of histones and the repositioning and eviction of nucleosomes. Four families of chromatin remodelers have been well characterized: switch/sucrose nonfermentable (SWI/SNF), chromodomain-helicase DNA-binding protein (CHD), inositol-requiring mutant 80 (INO80), and imitation switch (ISWI) (3). They play a significant role in controlling nucleosome position and density in the chromatin, thereby directly influencing the accessibility and regulatory functions of genomic regions (4).

The link between chromatin remodelers and cancer has been extensively studied. Frequent mutations in the subunits of the SWI/SNF complex, the primary chromatin remodeler, are found in various human tumors (5; 6). Similar gene alterations in ISWI subunits have been identified in cancer and are correlated with prognosis (3). However, while these correlations are recognized, the specific mechanisms (e.g., the repositioning dynamics, the dominant mode of actions, etc.) behind the involvement of chromatin remodelers in cancer development remain largely unexplored (7).

Prostate cancer's neuroendocrine transformation serves as an excellent model for examining chromatin remodelers in greater detail. It is the most prevalent malignancy and the second leading cause of cancer-related deaths among men in the US. A subset of patients initially diagnosed with adenocarcinoma prostate cancer may, in the later stages of disease progression, develop small cell neuroendocrine carcinoma, a histology associated with increased aggressiveness and resistance to traditional androgen receptor (AR)-directed therapeutics (8). This transformation involves significant downregulation of known genes, characterized epigenetic changes, and activation of known oncogenic drivers (9), making it possible to pinpoint specific loci for more detailed study of chromatin remodeling mechanisms.

Recent genomics research has introduced methods for localizing nucleosome positions, the primary targets of chromatin remodelers, from sequencing data (10; 11). However, no studies have yet tracked changes in nucleosome positions across disease progression. Data on shifting nucleosome positions, combined with information on the expression levels of chromatin remodelers, could reveal detailed patterns of chromatin remodeler complexes' behavior in specific cancer development scenarios.

For this project, we aim to accomplish the first steps of the above vision. We hope to validate the hypothesis that nucleosome repositioning or position dysregulation could be a potential cause for prostate cancer's neuroendocrine transformation. Herein, we developed a computational pipeline to identify nucleosome positions in single-cell Assay for Transposase Accessible Chromatin using sequencing (scATAC-seq) data from the neuroendocrine transformation of prostate cancer in mice. Tracking these positions through the transformation, we characterize global patterns in this process.

2 Dataset and Methods

2.1 Dataset

The dataset used in this project consists of single-cell RNA sequencing (scRNA-seq) and sc-ATAC seq data collected from the progression of prostate cancer in genetically engineered mouse models, specifically designed to study the neuroendocrine transformation associated with advanced prostate cancer. The RPM genotype in this model involves the knockout of both Rb1 and Trp53, along with the activation mutation of cMyc (T58A), which are key alterations relevant to prostate cancer pathology.

The mice acquire prostate cancer starting with the transplantation of organoids. The cancer histology progresses from initial luminal histology to neuroendocrine histology over the course of 10 weeks, with samples collected for sequencing at the time points marked on the timeline. A series of images captures the morphological changes occurring at each stage, illustrating the dynamic process of cancer transformation and the development of aggressive cancer phenotypes resistant to conventional therapies. All subsequent analyses will be based on this dataset.

2.2 Phenograph Clustering

scATAC-seq data typically face the issue of sparsity due to inherent challenges in the technology (12). This necessitates the use of the pseudo-bulking technique, whose purpose is to aggregate data from multiple individual cells into virtual or "pseudo" bulk samples, thereby increasing the overall quantity of data for each cell group. Cells in the same pseudo bulk should be in similar enough conditions to be considered the same cell. To achieve this, we use Phenograph Clustering, an approach that dissects

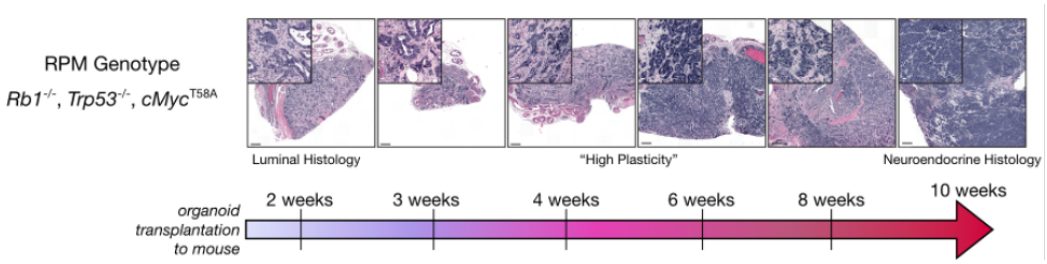

Figure 1: Histological progression of neuroendocrine transformation in RPM genotype ($Rb1^{-/-}$, $Trp53^{-/-}$, $cMyc^{T58A}$) prostate cancer in mice. The timeline shows organoid transplantation into mice and the subsequent progression from luminal histology through a state of high plasticity to fully developed neuroendocrine histology over a period of 10 weeks. Each panel represents a histological examination at two-week intervals, beginning at 2 weeks post-transplantation. Insets show magnified views of representative areas within each histological stage. The dataset, consisting of scRNA-seq scATAC-seq data from collected at the time points marked on the timeline, is obtained from Dr. Charles Sawyers' lab at the Memorial Sloan Kettering Cancer Center, New York.

high-dimensional single-cell phenotype data into cell subsets, each representing a distinct cellular population (13).

The Phenograph algorithm operates in three steps. First, given the input matrix of N single-cell RNA-seq measurements, it identifies the k nearest neighbors of each cell in Euclidean distance. This returns an $N \times k$ matrix, where each row represents the set of k nearest neighbors for each cell. Second, leveraging these sets, the Jaccard similarity coefficient is computed between every pair of cells, i.e., for every pair of cells a and b, where A and B represent their corresponding nearest neighbor sets, the Jaccard similarity coefficient between them is defined as follows:

$$J(a,b) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{1}$$

The algorithm then constructs a graph for N points, where each point represents a cell. The distance between every pair of cells is their Jaccard similarity coefficient. This effectively builds a weighted graph between cells that scales with the number of their common neighbors. In the third step, the Louvain community detection method is used to find partitions of the graph that maximize modularity, which measures the density of edges inside partitions compared to the density of edges between partitions. In the case of a weighted graph, the modularity is defined as follows:

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j) \tag{2}$$

where $A_{i,j}$ represents the edge weight between node i and node j; k_i and k_j represent the sum of all edge weights attached to node i and node j, respectively; m represents the sum of all edge weights in the graph; c_i and c_j represent the partition to which node i and node j belong; and the δ function $\delta(u, v)$ is 1 if $u = v$ and 0 otherwise. A heuristic approach is then used to optimize for local maximum values of modularity by adjusting the partitions (14).

2.3 Protein-Regulatory element Interactions at Nucleotide resolution using Transposition

To extract shifting nucleosome positions from scATAC-seq data, we utilize the Protein-Regulatory element Interactions at Nucleotide resolution using Transposition (PRINT) software (11). PRINT employs statistical testing to assess the likelihood of nucleosomes binding to DNA at each specific location on the genome.

Here is an overview of how PRINT operates: Initially, experiments were conducted on deproteinized DNA from bacterial artificial chromosomes (BACs) to determine the propensity of Tn5 transposase, the main enzyme used in scATAC-seq, to insert at each base pair. Subsequently, these data were used to train a machine learning model that predicts the likelihood of Tn5 transposase insertion in the

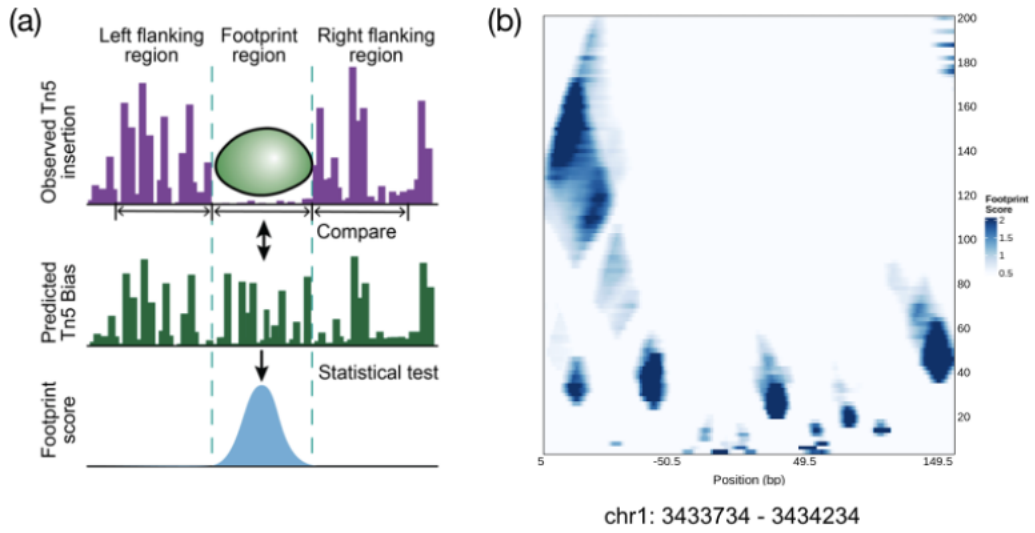

Figure 2: Analysis of nucleosome positions using PRINT software. Panel (a) illustrates the working mechanisms of PRINT. Observed Tn5 insertion profiles in the scATAC-seq data (purple bars) are compared to the predicted profiles (green bars) across a genomic region. Statistical testing is subsequently performed to determine protein-DNA binding likelihoods. An example binding protein is highlighted in light green oval, which resulted in fewer observed Tn5 insertions than predicted, and hence giving rise to a high footprint score. Panel (b) shows an example heat map of the footprint scores across a specific genomic region on chromosome 1 from the dataset described in part 2.1. The dark shades of blue indicate higher footprint score, and therefore likelihood of protein presence at each position. The horizontal axis indicates the genome position, while the vertical axis indicates the scale at which footprinting is performed. Larger proteins, like nucleosomes are captured using a larger scale, while smaller proteins, like transcription factors, are captured using a smaller scale.

absence of any binding proteins on the mouse genome (mm10). Lastly, these predicted likelihoods are compared with actual scATAC-seq data from mouse samples through statistical testing, producing a significance level (-log10 p-value) that is represented as a "footprint score" indicating the level of predicted protein binding.

During statistical testing, the size of the comparison windows determines the size of the DNA-binding proteins' footprints that can be detected. Although the PRINT software is capable of performing footprinting analysis at multiple scales, for the purposes of this project, we are specifically focusing on nucleosome positions. Consequently, we have set the scale to 100 for solely capturing nucleosome positions.

3 Results

3.1 Clustering of Dataset using Phenograph

To address the issue of data sparsity, we employed Phenograph for clustering the single-cell data derived from prostate cancer samples. Utilizing each cell's scRNA-seq data, the clustering algorithm groups cells into 25 distinct clusters based on their gene expression profiles. This method facilitated the identification of distinct cell populations, each representing different stages or types of cells present in the dataset.

The results of this clustering are depicted in the t-Distributed Stochastic Neighbor Embedding (t-SNE) plot shown in Figure 3. Notably, the plot highlights specific clusters that underwent neuroendocrine transformation, marked within a red oval, particularly Cluster 11 and Cluster 12. These clusters clearly differentiate from other cellular groups, illustrating the phenotypic shift from luminal to neuroendocrine histology in prostate cancer. Additionally, Cluster 1, which is closest to undergoing

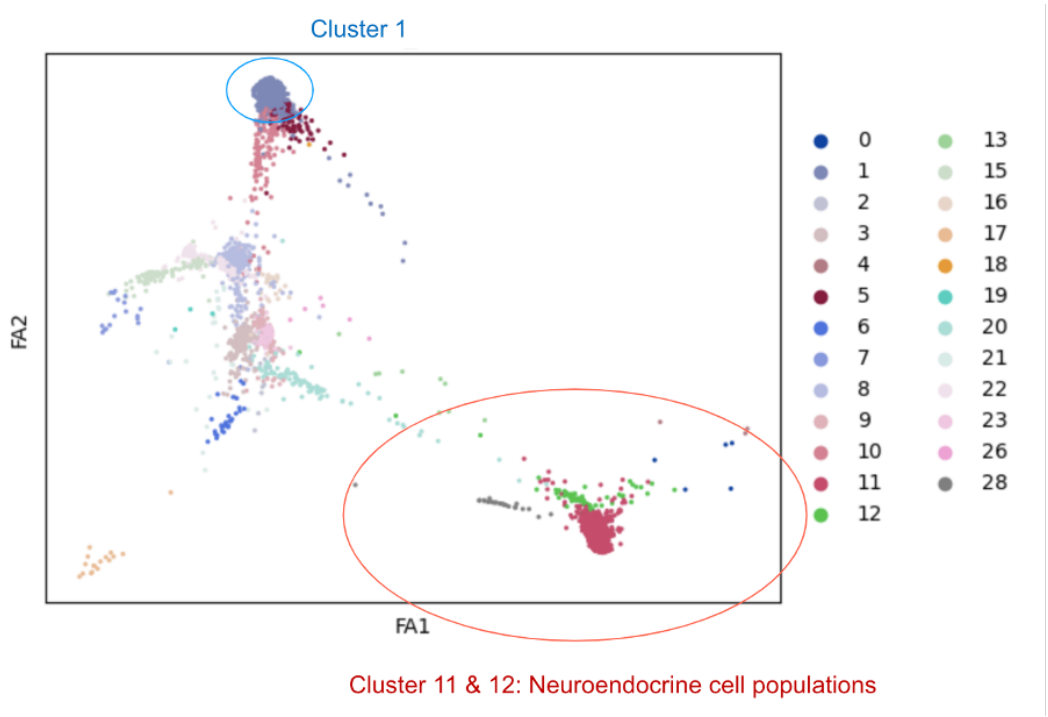

Figure 3: t-Distributed Stochastic Neighbor Embedding (t-SNE) plot showing the transformation of prostate cancer cells from luminal histology to neuroendocrine histology. Each point represents a single cell, colored according to different cell clusters as determined by Phenograph clustering. The neuroendocrine cell populations are highlighted within the red oval, demonstrating a distinct grouping from other cell types.

neuroendocrine transformation and still exhibits luminal histology, is identified as a likely progenitor cell population.

3.2 Extraction of Nucleosomal Positions across Regulatory Regions

Having assigned each cell a cluster label leveraging the scRNA-seq data, we shifted our focus to the scATAC-seq data. We utilized the PRINT software to extract the footprint score, an indication of the likelihood of a nucleosome being present at a specific location, from each of the 359,263 selected regulatory regions within each cell cluster. These regulatory regions cover all 20 chromosomes in mice, and each region is 501 base pairs wide.

This analysis aims to characterize the nucleosome landscape across diverse cellular states, providing insights into the chromatin remodeling dynamics associated with lineage transformation. The PRINT software was run at a scale of 100 base pairs to specifically target and extract nucleosome positions, excluding smaller scale elements such as transcription factors. This resulted in a three-dimensional matrix capturing the nucleosome footprinting across a substantial number of regulatory regions and cell clusters. To facilitate comparative analysis of chromatin remodeling dynamics across different cell clusters, we preserved the cell cluster dimension and concatenated all regions, ultimately forming a two-dimensional matrix suitable for downstream analysis.

3.3 Characterization of Global Patterns with Principal Component Analysis

Having extracted a data matrix containing nucleosome positions across diverse regulatory regions in all significant cell clusters along the prostate cancer progression timeline, we sought to identify potential global patterns. Specifically, we ask the following question: are there notable shifting patterns of nucleosome positions as prostate cancer undergoes neuroendocrine transformation? Using principal component analysis (PCA), we observed a clear variation along the first principal

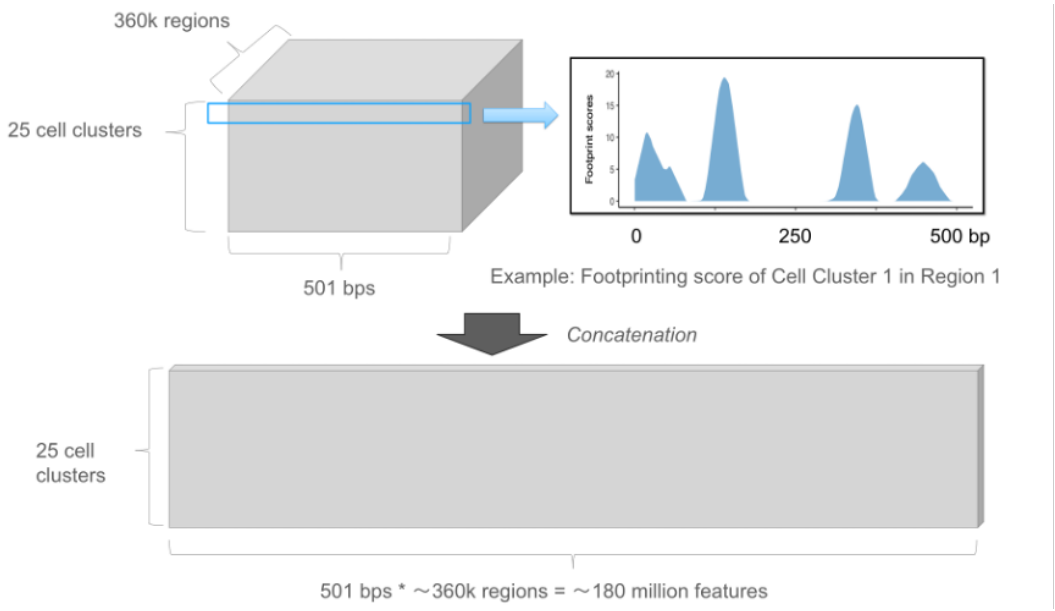

Figure 4: Schematic representation of the data processing workflow using PRINT software to extract nucleosome footprint scores. Initially, the PRINT software computes footprint scores from 359,263 selected regulatory regions, each 501 base pairs wide, for each of the 25 cell clusters, generating a three-dimensional matrix with dimensions 25 (cell clusters) by 501 (base pairs) by 359,263 (regions). The graph in the upper-right corner gives an example of footprint scoring for cell Cluster 1 in an example region. Subsequently, these scores are concatenated to form a final matrix with dimensions 25 by approximately 180 million features (501 multiplied by 359,263), which simplifies the data structure for further analysis.

component (PC1). The PCA plot, as shown below in Figure 5, highlights Cluster 11, one of the major neuroendocrine cell clusters, being significantly separated along PC1 from the other cell clusters. Cluster 1, which is identified as the closest cluster to neuroendocrine transformation and is likely the progenitor state, is also separated from rest of the clusters.

These provide evidence that PC1 strongly correlates with the neuroendocrine transformation of prostate cancer, with the separation along PC1 indicative of the substantial remodeling that the chromatin landscape in these clusters has undergone. However, adding nuances to the analysis, we observe that Cluster 12, another major neuroendocrine cell cluster, displays a similar PC1 value to clusters of luminal histology. This similarity could suggest that the nucleosome position shifts characterizing Cluster 12 might not be captured by PC1, but rather by other principal components. This indicates a potentially complex mechanism of different chromatin dynamics leading to different cell clusters during the cancer lineage transformation.

3.4 Analysis of Regulatory Regions with Most Correlation with PC1

Having observed a clear variation along Principal Component 1 (PC1), and noting its strong correlation with the neuroendocrine transformation of prostate cancer, it becomes the natural next step to identify which specific features contribute most significantly to PC1. To address this, we extracted PC1 and calculated the Pearson Correlation Coefficient between PC1 and all approximately 180 million features (359,263 regions multiplied by 501 positions). The Pearson Correlation Coefficient measures the linear correlation between two datasets and is defined as the following:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \tag{3}$$

where x_i and y_i represent elements from each of the two datasets, and $\bar{x}$ and $\bar{y}$ are the means of these datasets, respectively. This analysis enables us to pinpoint the nucleosome position dynamics that

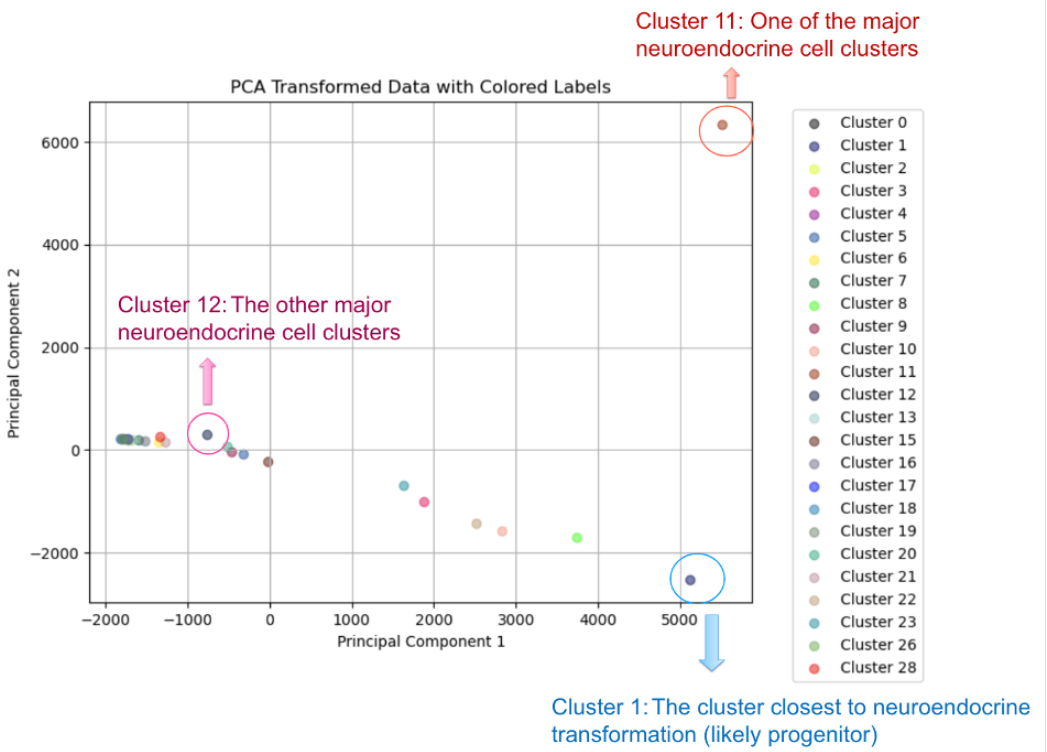

Figure 5: PCA plot illustrating the separation of cell clusters based on nucleosome position data across regulatory regions. Cluster 11, identified as one of the major neuroendocrine cell clusters, and Cluster 1, a potential progenitor state before the transformation, are highlighted and distinctly separated along Principal Component 1 from all other clusters. This is suggestive of the substantial remodeling processes that were undergone in Cluster 11 and Cluster 1. However, Cluster 12, another significant neuroendocrine cluster, is positioned closer to the clusters representing cells of luminal histology.

most closely correlate with PC1, providing insights into the specific chromatin remodeling changes associated with the transition to a neuroendocrine phenotype.

The distribution graph indicates that over half of the features exhibit moderate to strong correlation with Principal Component 1 (PC1), with correlation coefficients (r) greater than 0.5. To further understand the specific chromatin remodeling dynamics captured in PC1, we focused on visualizing regions containing features with the highest correlation to PC1 ($r > 0.98$). Analysis of these regions revealed two distinct modes of chromatin remodeling dynamics: 1) the initial eviction and subsequent reassembly of nucleosomes, and 2) shifts in nucleosome positioning along the DNA. These findings align partially with existing literature on the mechanisms of chromatin remodelers, but they also suggest an additional potential activity where remodelers recruit new nucleosomes to specific positions.

3.5 Future Directions

The findings presented from our current analysis provide initial insights into chromatin remodeling processes during prostate cancer's neuroendocrine transformation. To build on this foundation, several next steps are proposed for future research:

1. Wasserstein Barycenter Calculation: As a continuation of our work in Section 3.4, computing the Wasserstein Barycenter of all heatmaps for features with a Pearson Correlation Coefficient greater than 0.5 is a priority. This method would allow the synthesis of the distributional into a single representative heatmap, offering a more unified and intuitive visualization of the dominant patterns affecting PC1.

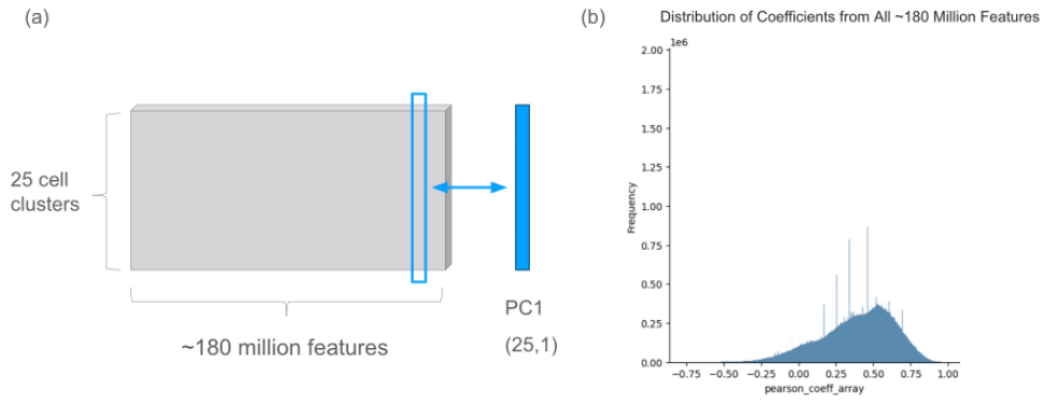

Figure 6: Schematics of the correlation analysis and the distribution of the coefficients. Panel (a) illustrates the process where we take each of the ∼180 million features from the data matrix with 25 cell clusters and calculate the Pearson Correlation Coefficient between the feature and the first principal component (PC1). Panel (b) illustrates the distribution of Pearson Correlation Coefficients calculated between PC1 and each of the ∼180 million features, showing the range of correlations that could be helpful in determining the features most strongly associated with the neuroendocrine transformation of prostate cancer.

2. Expanding the Dataset: To refine and validate our findings, it would be beneficial to rerun the analyses on a larger and more detailed-annotated dataset of prostate cancer neuroendocrine transformation. Including a greater number of meta-cells could potentially reveal subtler aspects of chromatin remodeling not captured in this initial study with only 25 cell clusters.
3. Incorporating Spearman Correlation Coefficients: Experimenting with Spearman Correlation Coefficients to assess the non-parametric rank correlation between PC1 and the features could uncover non-linear relationships that Pearson's method may miss.
4. Gene Association Studies: Identifying genes proximal to the shifting nucleosomes and examining their biological functions would be the natural next step to elucidate the potential functional impacts of nucleosome repositioning. Such an analysis would help understand how these chromatin remodeling events influence the phenotype observed in neuroendocrine transformation.

These potential next steps, added with the initial results presented in the above findings, hold the promise of deepening our mechanistic understanding of chromatin remodeling in cancer development.

4 Conclusion

In summary, we investigated the mechanisms and dynamics of chromatin remodelers during the neuroendocrine transformation of prostate cancer using scRNA-seq and scATAC-seq datasets derived from mouse models. Using the Phenograph clustering approach, we partitioned cells into representative clusters and utilized the PRINT software to map nucleosome positions across 359,263 regulatory regions within all 25 identified clusters. Principal Component Analysis (PCA) applied to this dataset revealed that the first principal component (PC1) captures significant variations corresponding to the direction of the neuroendocrine transformation, emphasizing the extensive chromatin landscape remodeling involved in this process. By calculating the Pearson correlation coefficient between each feature and PC1, we examined the regions containing features with the highest coefficients. We identified two primary nucleosome dynamics: 1) initial eviction followed by eventual reassembly, and 2) positional shifts. These findings enhance our understanding of the epigenetic alterations occurring during cancer progression. This study sets the stage for future research using expanded datasets and additional statistical analysis techniques to further elucidate chromatin remodeling mechanisms, potentially offering therapeutic insights into epigenome-targeted approaches for prostate cancer and other aggressive forms of cancer.

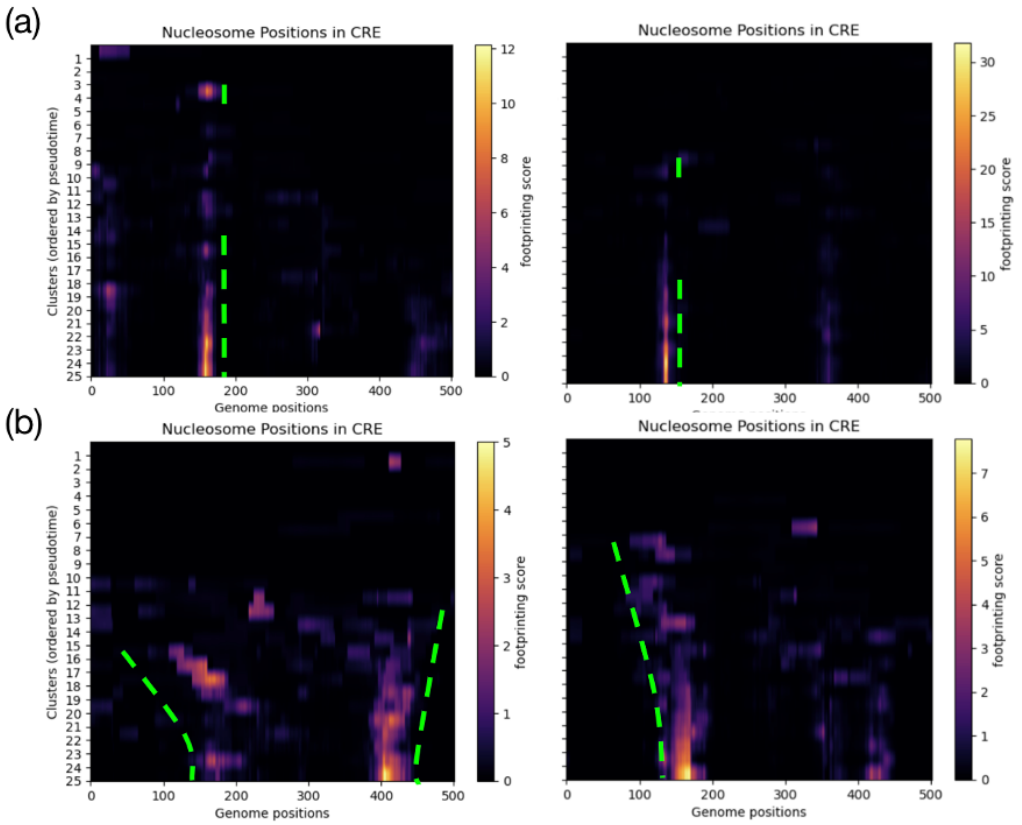

Figure 7: Visualization of nucleosome positioning dynamics in sample regulatory regions containing features with the highest correlation with PC1. Panel (a) and Panel (b) display heatmaps of nucleosome positions across different cell clusters, illustrating two main chromatin remodeling dynamics: (a) shows the initial eviction and reassembly of nucleosomes at later stages, indicated by the green dashed lines; (b) highlights shifting nucleosome positions during the transformation, with changes tracked also by the green dashed lines. The clusters listed on the vertical axis are reordered according to the magnitude of their PC1 value. These visualizations underscore some potential major remodeling strategies undertaken by chromatin remodelers during cancer progression.

Code Availability

The code for this project is available on GitHub at the following repository: `https://github.com/james-bole-pan/chromatin_remodeling`.

Acknowledgments

The author sincerely thank the guidance of Dr. Cassandra Burdziak, Dr. Dana Pe'er, and Dr. Andrew Blumberg in the course of completing this project.

References

[1] K. Kukkonen, S. Taavitsainen, L. Huhtala, J. Uusi-Makela, K. J. Granberg, M. Nykter, and A. Urbanucci, "Chromatin and epigenetic dysregulation of prostate cancer development, progression, and therapeutic response," *Cancers*, vol. 13, no. 13, p. 3325, 2021.

[2] T. Chen and S. Y. Dent, "Chromatin modifiers and remodellers: regulators of cellular differentiation," *Nature Reviews Genetics*, vol. 15, no. 2, pp. 93–106, 2014.

[3] Y. Li, H. Gong, P. Wang, Y. Zhu, H. Peng, Y. Cui, H. Li, J. Liu, and Z. Wang, "The emerging role of iswi chromatin remodeling complexes in cancer," *Journal of Experimental & Clinical Cancer Research*, vol. 40, pp. 1–27, 2021.

[4] C. P. Rodrigues, M. Shvedunova, and A. Akhtar, "Epigenetic regulators as the gatekeepers of hematopoiesis," *Trends in Genetics*, vol. 37, no. 2, pp. 125–142, 2021.

[5] J. A. Biegel, T. M. Busse, and B. E. Weissman, "Swi/snf chromatin remodeling complexes and cancer," in *American Journal of Medical Genetics Part C: Seminars in Medical Genetics*, vol. 166, pp. 350–366, Wiley Online Library, 2014.

[6] K. Mardinian, J. J. Adashek, G. P. Botta, S. Kato, and R. Kurzrock, "Smarca4: implications of an altered chromatin-remodeling gene for cancer development and therapy," *Molecular cancer therapeutics*, vol. 20, no. 12, pp. 2341–2351, 2021.

[7] B. Monterde and I. Varela, "Role of swi/snf chromatin remodeling genes in lung cancer development," *Biochemical Society Transactions*, vol. 50, no. 3, pp. 1143–1150, 2022.

[8] L. Puca, P. J. Vlachostergios, and H. Beltran, "Neuroendocrine differentiation in prostate cancer: emerging biology, models, and therapies," *Cold Spring Harbor perspectives in medicine*, vol. 9, no. 2, p. a030593, 2019.

[9] Y. Yamada and H. Beltran, "Clinical and biological features of neuroendocrine prostate cancer," *Current oncology reports*, vol. 23, pp. 1–10, 2021.

[10] G. E. Zentner and S. Henikoff, "High-resolution digital profiling of the epigenome," *Nature Reviews Genetics*, vol. 15, no. 12, pp. 814–827, 2014.

[11] Y. Hu, S. Ma, V. K. Kartha, F. M. Duarte, M. Horlbeck, R. Zhang, R. Shrestha, A. Labade, H. Kletzien, A. Meliki, *et al.*, "Single-cell multi-scale footprinting reveals the modular organization of dna regulatory elements," *bioRxiv*, 2023.

[12] Z. Li, C. Kuppe, S. Ziegler, M. Cheng, N. Kabgani, S. Menzel, M. Zenke, R. Kramann, and I. G. Costa, "Chromatin-accessibility estimation from single-cell atac-seq data with scopen," *Nature communications*, vol. 12, no. 1, p. 6386, 2021.

[13] J. H. Levine, E. F. Simonds, S. C. Bendall, K. L. Davis, D. A. El-ad, M. D. Tadmor, O. Litvin, H. G. Fienberg, A. Jager, E. R. Zunder, *et al.*, "Data-driven phenotypic dissection of aml reveals progenitor-like cells that correlate with prognosis," *Cell*, vol. 162, no. 1, pp. 184–197, 2015.

[14] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of statistical mechanics: theory and experiment*, vol. 2008, no. 10, p. P10008, 2008.