3-30-2018

Integrative Machine Learning and Network Mining Models for the Inference of Regulatory Elements and Interactions in Human Cells

Asa Thibodeau

University of Connecticut - Storrs, asa.thibodeau@uconn.edu

Follow this and additional works at: https://opencommons.uconn.edu/dissertations

Recommended Citation

https://opencommons.uconn.edu/dissertations/1713
Integrative Machine Learning and Network Mining Models for the Inference of Regulatory Elements and Interactions in Human Cells

Asa Jake Thibodeau, PhD

University of Connecticut, 2018

With the increase in diverse genome profiling technologies and publicly available ontology databases ranging from open chromatin profiles to the 3D structure of the genome, it is imperative to build novel computational methods that take full advantage of these diverse datasets to uncover the regulatory mechanisms behind cellular functions. Integrating these datasets offers the opportunity to identify regulatory elements (i.e., promoter, enhancers, etc.) and interactions critical for cell-type-specific functions. Here, the goal’s two fold: 1) inference of regulatory interactions and networks from 3D chromatin interaction datasets and 2) inference of cell-specific and non-specific regulatory elements such as enhancers (regulatory elements that target gene promoters and regulate their expression).

To address the first goal, two software tools were developed: (1) a web-accessible application: Querying and visualizing chromatin Interaction Network (QuIN) and (2) a pathway analysis prioritization tool: Triangulation of Perturbation Origins and Identification of Non-Coding Targets (TriPOINT). QuIN enables users to easily mine chromatin interaction datasets and integrate them with other sources such as SNPs and epigenetic marks to ultimately build networks to query and visualize them in downstream analyses and to prioritize genomic loci (i.e., disease-causing variants). Similarly, TriPOINT uses pathways in conjunction with chromatin interaction networks to identify perturbed genes in treatment vs. control cases, implementing pathway topology based approaches for identifying inconsistencies in pathways and incorporating the capabilities of QuIN to integrate non-coding regulators targeting genes in these pathways through chromatin interaction data. The second goal was achieved using two approaches. First, features obtained from network mining were trained on support vector machines to assess the predictive power in identifying cell-type-specific promoters (broad domains) and enhancers (super enhancers) from chromatin interaction networks. Network signatures were mined in three cell lines...
(MCF-7, K562, and GM12878) using QuIN across multiple chromatin interaction assays (ChIA-PET, Hi-C, and HiChIP) and it was discovered that network related features could effectively discriminate typical promoters and enhancers from cell-type-specific ones. Second, features from Assay for Transposase Accessible Chromatin (ATAC-seq) were profiled to identify enhancers from accessible chromatin in neural network models. Models were highly predictive of enhancers; useful for individual specific and clinical sample settings.
Integrative Machine Learning and Network Mining Models for the Inference of Regulatory Elements and Interactions in Human Cells

Asa Jake Thibodeau

B.S., University of Connecticut, 2010

A Dissertation
Submitted in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy
at the
University of Connecticut

2018
Acknowledgements

I would first like to thank my major advisor Dr. Shin for all of his advice and guidance throughout the years. Dr. Shin not only brought me into the Ph.D. program as an undergraduate but also introduced me to many wonderful people at UCONN Health and The Jackson Laboratory for Genomic Medicine. I am very fortunate to have such an exciting future ahead all thanks to Dr. Shin taking notice of me during my undergraduate studies.

I owe much of my growth to the CSE professors at the University of Connecticut. My knowledge in computer science is a direct result of their teaching efforts since Fall 2006 when I began my journey as an undergraduate. My goals were different back then, but over time a new world slowly began to open up, forging new goals with higher impacts. I will always be thankful for their time and effort throughout the years. I would also like to thank Rebecca Randazzo and the late Debra Mielczarek in the CSE department and appreciate all the help they have provided me over the years with administrative matters.

I was very fortunate to meet with many wonderful people at the UCONN Health Center and The Jackson Laboratory for Genomic Medicine, all of which I am greatly appreciative for the time spent with me and their help with my research goals. Dr. Bayarsaihan from UCONN Health was pivotal in my research in chromatin interactions and epigenetics, and I thank him for his time to meet with me personally to discuss these topics, which would later become my main research focus. I also thank Paola Vera-Licona, Eladio Márquez, Michael Stitzel, Asli Uyar, Joshy George, and Francesca Menghi, for their insights and collaborations, and all the members of the Ucar, Stitzel, and Lee lab at JAX GM who provided constructive feedback from broad perspectives. I would especially like to thank Dr. Ucar for her invaluable guidance when undergoing research in computational biology. Being part of her lab in addition to Dr. Shin’s lab has been a wonderful experience and I cannot thank her enough for everything I have learned while conducting research with her.
I thank my advisory committee Dr. Shin, Dr. Bansal, Dr. Mandoiu, Dr. Ucar, and Dr. Wu for their patience and time in reviewing this research.

Finally, I would like to thank Caitlin Pach and my parents Daniel and Patricia Thibodeau for all their support throughout my studies.
# Table of Contents

**CHAPTER 1: INTRODUCTION** .......................................................................................................................... 1

**CHAPTER 2: QUERYING AND VISUALIZING CHROMATIN INTERACTION NETWORKS (QUIN)** ......................................................................................................................... 5

  2.1 Introduction .................................................................................................................................................. 5
  2.2 Software and Databases ............................................................................................................................. 7
  2.3 Constructing Networks from Chromatin Interaction Data ............................................................................. 9
  2.4 Annotation of Chromatin Interaction Networks Through Data Integration .............................................. 10
  2.5 Network Analysis Methods .......................................................................................................................... 10
  2.6 A Case Study Identifying Non-Coding Variants Associated with Breast Cancer ..................................... 12
  2.7 Discussion .................................................................................................................................................. 17

**CHAPTER 3: TRIANGULATION OF PERTURBATION ORIGINS AND IDENTIFICATION OF NONCODING TARGETS (TRIPOINT)** ........................................................................ 19

  3.1 Introduction ................................................................................................................................................ 19
  3.2 Software & Implementation .......................................................................................................................... 21
  3.3 Metrics for Identifying Pathway Perturbations .......................................................................................... 22
  3.4 Identification of Non-Coding Regulators .................................................................................................... 24
  3.5 Identification of Perturbed Gene RALGDS and its Non-Coding Regulators ........................................... 25
  3.6 Discussion .................................................................................................................................................. 28

**CHAPTER 4: INFERENCE OF CELL-TYPE-SPECIFIC REGULATORY ELEMENTS FROM CHROMATIN INTERACTION NETWORKS** ................................................................................. 29

  4.1 Introduction ................................................................................................................................................ 29
  4.2 Chromatin Interaction Networks from Diverse Cell Lines and Assays ..................................................... 31
  4.3 Chromatin Interaction Network Annotation ............................................................................................... 32
  4.4 Increased Interaction Frequency Among Broad Domains and Super Enhancers ..................................... 34
  4.5 Broad Domains and Super Enhancers are Hubs in Interaction Networks ................................................ 37
  4.6 Broad Domains and Super Enhancers Have Unique Connectivity Patterns ........................................... 39
  4.7 Network Connectivity Patterns are Predictive of Cell-Specific Activity .................................................. 43
  4.8 Discussion .................................................................................................................................................. 51

**CHAPTER 5: INFERENCE OF ENHANCERS FROM ACCESSIBLE CHROMATIN: PREDICT ENHANCERS FROM ATAC-SEQ (PEAS)** ................................................................................. 54

  5.1 Introduction ................................................................................................................................................ 54
  5.2 Extracting Features from ATAC-Seq ............................................................................................................ 56
  5.3 Assigning Ground Truth Class Labels with CHROMHMM ...................................................................... 57
  5.4 Neural Networks Discriminate Three Class Labels Across Five Cell Types ........................................... 61
  5.5 Comparison of Machine Learning Models & Model Tuning .................................................................... 63
  5.6 Individual Specific Enhancer Prediction ..................................................................................................... 67
  5.7 Cross Cell Type Enhancer Prediction ........................................................................................................ 68
  5.8 Inference of Enhancers via Cell-Agnostic Models ...................................................................................... 69
  5.9 Discussion .................................................................................................................................................. 71

**CHAPTER 6: CONCLUSIONS AND DISCUSSION** ............................................................................................ 73
Chapter 1

Introduction

Recent advancements have identified diverse cis-regulatory elements with diverse functions relating to transcription and gene regulation. Assays such as chromatin immunoprecipitation followed by sequencing (ChIP-Seq)\(^1\) identify regions marked by histone modifications whose combinatorial signals correspond to each of these different elements. Furthermore, efforts and algorithms to identify these regulatory elements have been developed\(^2,3\), allowing for their systematic study. One set of regulatory elements of particular interest are enhancers, which have been shown to regulate gene expression in cell or tissue-specific contexts\(^4\). Moreover, enhancers can regulate gene expression over long distances and do not necessarily target genes that are in close proximity with respect to linear distances of the genome. Genomic technologies have shown that regulatory elements distal in the linear genome can actually be within close proximity due to the 3D chromatin structure. Assays such as Chromosome Conformation Capture based methods (3C)\(^5\), 4C\(^6\), 5C\(^7\), Hi-C\(^8\), HiChIP\(^9\) and
Chromatin Interaction Analysis by Paired-End Tag Sequencing (ChIA-PET)\textsuperscript{10}, capture this 3D chromatin structure, resulting in chromatin interactions that provide evidence for two loci being in close proximity, even when they are mega-bases away from each other in the linear genome. Chromatin interactions, however, do not provide any information on the types of regulatory elements that are interacting, implying that data integration is crucial for determining different types of interactions such as enhancer-promoter interactions. Furthermore data integration improves prioritization of problematic loci, such as those harboring single nucleotide polymorphisms (SNPs) found in diseases, out of potentially hundreds of thousands of regions.

Enhancers and promoters, and the interactions among them, are of particular interests as recent studies have identified classes of promoters and enhancers that are cell-specific, namely, broad domains\textsuperscript{11} demarcated by broad H3kme3 domains, super enhancers\textsuperscript{12} demarcated by enrichment of H3k27ac signal, and stretch enhancers\textsuperscript{13} identified by long stretches of ChromHMM\textsuperscript{3} enhances states. However, little is known about how these cell-specific enhancers and promoters interact and whether or not there are differences in interactions, which could provide insights into their cell-specificity. Furthermore, as these loci are cell-specific, they have increased importance and are often used for prioritization of disease-associated targets. In addition to cell-specific annotations, enhancers may be active at the individual level; however, capturing such enhancers proves difficult as enhancer annotations are often defined from references of different individuals, and assays for capturing this information require high cell counts in addition to requiring multiple ChIP-seq experiments.

In addition to identifying enhancer targets, downstream analyses such as pathway analyses\textsuperscript{14-20}, could be further improved with chromatin interaction and enhancer information as non-coding targets are typically not included in their analyses, and typically restricted to gene
expression or proteomics data. The combination of chromatin interactions and enhancers in pathways is a less explored area, but would be able to connect non-coding regulators to their target genes to further explain how genes in pathways are regulated.

These challenges have been addressed in multiple ways. First, to address the need for data integration with chromatin interaction datasets, a web accessible platform specifically for data integration with chromatin interaction data was developed: Querying and Visualizing Interaction Networks (QuIN)\textsuperscript{21}. QuIN combines chromatin interactions into networks to facilitate data integration needs while providing network-based analyses to easily interrogate these interactions in terms of their network topology. To further focus on gene prioritization and to interrogate pathway topologies, Triangulation of Perturbation Origins and Identification of Non-Coding Targets (TriPOINT) was developed. TriPOINT enriches pathways with non-coding regulator activity through chromatin interactions and identifies perturbed genes based on a number of network topology measures and the number of interacting non-coding regulators targeting the loci. Once perturbed genes are identified, users can prioritize their non-coding regulators, leading to a more focused set of loci for further experimentation. These two software tools, discussed in detail in chapters 2 and 3, aim to provide an easy way to interrogate data while providing novel analyses for prioritization. Cell-specific promoters and enhancers become the main focus in chapter 4, as inferences made from chromatin interaction networks are used in order to identify broad domains and super enhancers, and to understand their interactivity in the 3D genome. Network mining approaches were used to identify network related features to discriminate broad domains from promoters as well as super enhancers from enhancers. As a complementary interest, chapter 5 describes methods aimed to infer individual specific enhancers as well as enhancers in cell types with missing reference annotations, using machine learning
methods to discriminate enhancers from other regulatory elements through features extracted solely from Assay for Transposase Accessible Chromatin (ATAC-seq). The low cell count required for ATAC-seq makes it ideal for identifying individual specific enhancers in comparison to other methods which require multiple assays and an order of magnitude (or more) greater cell numbers. Overall, two data integration tools that facilitate the need to connect non-coding regions or regulators such as enhancers to their gene targets were developed alongside two studies which further identify and define enhancers, promoters and their interactivity.
Chapter 2

Querying and Visualizing Chromatin Interaction Networks (QuIN) \(^1\)

2.1 Introduction

Chromatin structure plays a major role in basic cellular functions. Advances in genomic technologies have revealed information regarding three-dimensional (3D) chromatin conformation and have shown that many regulatory elements that are distal on the linear genome map are actually in close physical proximity with each other as a result of the 3D chromatin structure. Current technologies for capturing this 3D structure include Chromosome Conformation Capture based methods (3C) \(^5\), 4C\(^6\), 5C \(^7\), Hi-C\(^8\), HiChIP\(^9\) and Chromatin Interaction Analysis by Paired-End Tag Sequencing (ChIA-PET)\(^10\). These technologies identify chromatin interactions between promoters, enhancers, and other regulatory elements. The data generated by these technologies are the starting point from which we can infer distal regulatory interactions and their system-level effects by modeling them in the form of interaction networks.

---

\(^1\) Sections of this chapter were previously published in PLOS Computation Biology\(^21\) under the Creative Commons Attribution License 4.0 [https://creativecommons.org/licenses/by/4.0/](https://creativecommons.org/licenses/by/4.0/) and modified for this dissertation.
Moreover, integrating interaction datasets with additional data types and public repositories facilitate the discovery of regulatory elements and interactions that are critical for cellular functions and for disease biology, such as gene targets of non-coding regulatory elements harboring disease-causing Single Nucleotide Polymorphisms (SNPs). It is therefore imperative to have an easy-to-use software platform that enables biologists to model and study their chromatin interaction datasets under the light of other data sources, such as SNP databases and epigenetic marks.

Chromatin interaction data are typically visualized using a genome browser in a linear fashion, providing one-dimensional representation of the data. A commonly used tool for this is the UCSC Genome Browser. However, two and three dimensional representations of chromatin interactions in the form of networks and three-dimensional models can provide a global view of the interactions and facilitate the use of established network analysis methods and measures on these datasets. For example, network representation of RNA Pol2 ChIA-PET data revealed that loci harboring disease-associated SNPs are differently connected in chromatin interaction networks. Most tools such as HOMER’s HiC analysis suite, SIMA, HiBrowse, GWAS3D, and GenomicInteractions for analyzing interaction data currently are unable to take advantage of network representations or are limited in data integration capabilities. The CytoHiC Cytoscape plugin for example is limited in annotation and query capabilities.

In order to overcome the limitations of current tools, a single platform for Querying and visualizing Chromatin Interaction Networks (QuIN) was developed. QuIN enables: 1) building and visualizing chromatin interaction networks from ChIA-PET or HiC interactions; 2) annotating these networks with functional information from epigenetic datasets, SNPs, gene definitions, Gene Ontology terms, etc.; 3) querying network components for specific genes, loci,
or disease-causing SNPs; and 4) utilizing network-based algorithms and measures to prioritize genomic sites for functional validation (Figure 2.1). QuIN mines chromatin interaction datasets such as those generated by the ENCODE consortium\(^2\) (or user supplied ones) and integrates these datasets with other functional information such as chromatin states that can be inferred from histone modification datasets or SNP databases. QuIN is designed to enable biologists to easily represent and annotate their chromatin-interaction datasets in the form of networks and to use these datasets for discovering important interactions or targets.

![Figure 2.1. A screenshot of QuIN’s web interface highlighting its features.](image)

1. menus for uploading data and building networks,
2. options for visualizing and annotating a network,
3. target discovery menu for visualizing and exporting direct and indirect targets from source annotations to target annotations
4. network visualization panel,
5. options for searching, querying, or exporting the network,
6. the menu for performing GO Enrichment Analysis on the current subnetwork,
7. tools for summarizing network construction statistics, centrality measures and enrichment of interactions between annotations,
8. dialog box showing additional information about a selected node, including centrality measures, SNPs, and associated diseases.

### 2.2 Software and Databases

QuIN was developed as a web application using Java for backend computations and Javascript for front-end development. For graphical user interface development, Cytoscape JS ([http://js.cytoscape.org/](http://js.cytoscape.org/)) was used for network visualization while JQuery/JQuery UI
(https://jquery.com/) packages were used for the remainder of user interface elements. For downstream analyses and graph image generation, R was employed. Server to R communication was facilitated using RServe\textsuperscript{32}.

For data storage and to more easily mine chromatin interaction networks built using QuIN, a MySQL database was employed. In addition to storing user data for a limited amount of time, the MySQL database includes local version of publicly available data, including ClinVar\textsuperscript{33}, dbSNP\textsuperscript{34}, GWAS Catalog\textsuperscript{35}, NCBI’s Gene2Refseq and Gene2GO databases\textsuperscript{36} and UCSC’s hg19 RefSeq database\textsuperscript{37}. Figure 2.1 summarizes software and databases used in QuIN as well as the data workflow used with them.
2.3 Constructing Networks from Chromatin Interaction Data

Two approaches were implemented for constructing networks from chromatin interactions obtained from preprocessing tools such as ChIA-PET tool for ChIA-PET and HOMER for HiC. The first approach infers networks from the chromatin interaction anchors (i.e., the two genomic loci interacting), without integrating other knowledge to filter out false positive interactions. The second approach allows the user to predefine node locations defined by, for example, accessible chromatin assays such as DNASE-seq, FAIRE-seq or ATAC-seq. Each approach utilizes
a three-step process of 1) extracting nodes, 2) network edge generation, and 3) connected component discovery. Each step is described in more detail in Appendix A.

2.4 Annotation of Chromatin Interaction Networks Through Data Integration

QuIN was designed to easily integrate various types of datasets including genomic regions of interests, gene lists, SNPs, and diseases/traits found in the GWAS catalog\textsuperscript{35}. Locations of Nodes defined in the chromatin interaction networks are annotated by intersecting genomic coordinates of annotation sets provided. For annotations such as genes and diseases/traits, genomic coordinates are determined through the database designed specifically to obtain this information quickly through the means of methods available from MySQL databases. Genomic locations of genes are obtained from local database storage of UCSC’s refseq\textsuperscript{37} definitions where as diseases/traits are obtained based on their associated SNP information and inferring genomic locations from dbSNP\textsuperscript{34}.

2.5 Network Analysis Methods

Four network metrics were implemented for analyzing chromatin interaction networks built using QuIN: 1) Connectivity degree, 2) Closeness, 3) Harmonic, 4) Betweenness. Connectivity degree defines the number of edges connected to a node. Closeness was computed as:

\[
\text{Closeness}(v) = \frac{1}{\sum_{u \in N_c} \text{nd}(u, v)}
\]  

(2.1)

where \(\text{nd}\) is the number of edges or network distance between nodes \(u\) and \(v\), and \(N_c\) represents the set of nodes within the connected component containing \(v\). Harmonic was computed as:

\[
\text{Harmonic}(v) = \frac{1}{\sum_{u \in N_c, u \neq v} \text{nd}(u, v)}
\]  

(2.2)

Finally, betweenness centrality was computed as:
Betweenness(v) = \frac{|\text{sp}(u,x,v)|}{|\text{sp}(u,x,x)|} \quad (2.3)

where \(sp(u,x,v)\) denotes all shortest paths between nodes \(u\) and \(x\) that include node \(v\). Normalized centrality measures of a node were calculated with respect to the size of the network component \(|N_c|\).

To decompose larger networks and understand local topologies of genomic regions of interests described by nodes in the network, Graphlet measures were made available through QuIN. Graphlets are defined as small, connected, and non-isomorphic sub-networks components that can be used to decompose and describe a large network\(^{33}\). In total 73 orbits (isomorphic nodes defined within a graphlet) are counted over 16 different graphlets of size 2-5 with respect to the number of nodes by incorporating the GraphletCounter\(^{44}\) Cytoscape plugin into QuIN.

Frequency of interactions between annotated nodes in networks is also measured with methods developed specifically in QuIN. Two expected frequency measures were made available to evaluate the observed number of interactions over the expected. The first expected frequency measure assumes all nodes in the network are equally likely to be connected by an edge and is theoretically calculated as follows:

\[
\text{Expected}(A, B) = 2 \frac{|E|}{|N|} \frac{|A|}{|N|} \frac{|B|}{|N|} \frac{1}{1} \quad (2.4)
\]

where \(A\) and \(B\) represent the nodes annotated with the respective annotations for \(A\) and \(B\), \(E\) represents the Edges in the network, and \(N\) represents all nodes in the network. A method permuting annotations was also implemented to calculate the expected frequency, however this approach requires more time and ends up converges to the theoretical expected frequency. The second expected frequency measure assumes edges cannot be connected if nodes map to different chromosomes and have distances farther than a user defined distance \(d\). This expected frequency is therefore calculated as
where $C_i$ represents the set of all nodes in chromosome $i$, $|E_{C_i}|$ represent the number of observed edges within the node subset $C_i$ and $gd(x,y)$ is the genomic distance from node $x$ to node $y$. In both expected frequency measures, values were adjusted accordingly to account for two nodes being selected from the same set of annotations.

As a final method to facilitate network analysis of chromatin interaction networks, a method for finding all shortest paths between source annotations and target annotations in the network was implemented with the term target discovery. Target discovery employs breadth-first search procedures to efficiently identify all connected target nodes from source annotations, providing information on the number of targets and their network distances.

### 2.6 A Case Study Identifying Non-Coding Variants Associated with Breast Cancer

To demonstrate the core functionalities of QuIN, a case study was conducted using RNA Pol2 ChIA-PET data in the MCF-7 breast cancer cell line. A chromatin interaction network was constructed using ChIA-PET (GSM970209) and DNASE-seq (GSM816627) data generated by the ENCODE consortium\(^2\). The resulting MCF-7 network comprised of 59,02 nodes, 65,308 edges, and 8,133 connected components. This network was used in order to discovery gene targets of non-coding variants associated with breast cancer.

The MCF-7 interaction network was first annotated with MCF-7 non-coding variants (NCV) obtained from the COSMIC database\(^4\) (cancer.sanger.ac.uk). In total 36 NCVs were identified as intersecting with genomic coordinates of nodes in the network. Using target discovery, gene targets of these NCVs were identified, using a 2kb distance upstream and downstream of their transcription start site to designate their promoter regions. Direct targets
(promoter nodes targeted with a single edge by a node harboring a NCV) and indirect targets (defined as having 2-4 edges between promoter nodes and NCV nodes) were captured using the target discovery analysis. This analysis revealed 90 genes through direct targets and 638 genes through indirect targeting.

To understand whether gene targets obtained from ChIA-PET interaction networks could be useful in discovering genes with potential implications in breast cancer, gene lists obtained using nearest TSS annotations were compared against the gene lists obtained from MCF-7 interaction networks (both direct and indirect targets). Using the nearest TSS approach, 30 genes were identified, which we compared against direct (90) and indirect targets (638) obtained from the ChIA-PET network target discovery analysis. Four gene lists were used in comparisons including known oncogenes,\textsuperscript{46-48} tumor suppressor genes,\textsuperscript{49} genes associated with poor prognosis in breast cancer patients,\textsuperscript{50} and a unified list including previous lists and other breast cancer associated gene lists.\textsuperscript{45,51,52} Enrichment of each gene list was calculated for the three NCV target strategies: 1) nearest TSS, 2) direct ChIA-PET targets and, 3) indirect ChIA-PET targets. These analyses revealed that ChIA-PET based targets were significantly more enriched for cancer-related genes than nearest gene targets. For genes associated with poor prognosis, p-values for direct and indirect targets were $3.64 \times 10^{-3}$ and $3.15 \times 10^{-8}$ respectively where as nearest TSS targets had a p-value of $6.38 \times 10^{-2}$ (calculations made using Fisher’s exact test) (Figure 2.3).

This analysis also revealed that with the help of chromatin interaction networks and QuIN, one can capture gene targets of NCVs that are more likely to be relevant for the disease than nearest TSS assignments, even if they are separated via multiple edges in the network.
To further assess the relevance of NCV gene targets as defined by chromatin interaction networks to breast cancer, the three lists of genes previously mentioned were studied once again by comparing their expression levels among TCGA samples\textsuperscript{53}. Differential expression of genes
between breast cancer samples and normal breast epithelium samples were calculated using RNA-seq data. Targets discovered by ChIA-PET were found to include genes that are more differentially expressed between cancer and normal tissues in comparison to nearest gene targets (Figure 2.4). Moreover, it was also observed that even indirect targets of the NCVs could be disease relevant, which highlights the system-level impact of disease-causing variants and the importance of studying these interactions at the network level. As an example, Figure 2.5 shows a network revealing the indirect relationships of a region harboring an NCV to genes associated with cancer, including the well-known tumor repressor TP53 and multiple breast cancer-associated genes, such as EIF4A, EIF5A, AURKB and CLDN7.
Figure 2.4. Direct and indirect interacting gene pairs exhibit elevated gene expression correlations. Gene expression correlations between gene pairs connected via direct and indirect interactions in MCF-7 ChIA-PET network compared to correlations between genes without interactions.

*Exactly one edge between promoters. Does not include promoter pairs within the same node (0 edges).
2.7 Discussion

QuIN is a web accessible, easy to use platform for for analyzing 3D chromatin interaction datasets hosted at https://quin.jax.org. Since its publication\textsuperscript{21}, QuIN has been extended, implementing intra-chromosome interaction frequency procedures and incorporating graphlet measures. Additionally, QuIN was extensively used to mine chromatin interaction signatures to identify cell-type-specific promoters and enhancers\textsuperscript{59}. A current limitation of this tool is its inability to capture loops entirely within other loops, which is not an easy task to incorporate for network representations. Additional information such as relative positions of CTCF motifs may also prove useful for classifying edges with respect to different loop types (such as convergent loops) previously described\textsuperscript{60} within these networks as well as provide further filters to drill-down putative loci of interests interacting with known disease-relevant genes. As more
information on chromatin interactions becomes available, tools available from QuIN will become increasingly important to infer regulatory changes related to loss or gain of interactions. QuIN is an open source project released under the GNU General Public License Version 3 and is available on GitHub (https://github.com/UcarLab/QuIN/).
Chapter 3

Triangulation of Perturbation Origins and Identification of Noncoding Targets (TriPOINT)

3.1 Introduction

Pathway analyses are often utilized to identify pathways that are enriched in differential genes between conditions, for example, cases vs. controls, to gain a better understanding of the biological processes that are affected by the phenotype of interest (e.g., a disease). Methods for pathway analysis over the years have fallen into three categories: i) over representation analyses which count the number of differentially expressed genes within a pathway; ii) functional class scoring which calculates enrichment scores of pathway gene sets, and iii) pathway topology analyses where pathways are translated into directed graphs or networks to incorporate directionality and interaction types such as activation or inhibition (see Figure 3.1 for an example activation/inhibition relationships). These analyses can lead to the identification of pathways whose functions are affected as a result of a disruption in the processes, through, for example a single nucleotide polymorphism (SNP), that might be
associated with a disease state. However, the majority of SNPs are located in non-coding regions, where determining their phenotypic outcome is a challenging task. Moreover, non-coding regions include enhancers, which are cis-regulatory elements that have been shown to precisely regulate a gene’s expression in cell-specific contexts, further reinforcing the importance of incorporating non-coding information with gene expression and pathway analyses. In recent years, several assays have been developed, including ChIA-PET, HiC, and HiChIP, to identify chromatin loops that bring non-coding regions in close proximity of their target genes’ promoters, which help uncover their phenotypic outcome. Furthermore, it was shown that the degree to which a gene interacts with non-coding targets has been associated with its importance in the studied cell type, which can be used to further prioritize non-coding regions and their targets for experimental validation. As more data and methods become available for linking non-coding regions to their target genes, it becomes more important to provide computational tools to incorporate non-coding regions into downstream analyses of differentially expressed genes and pathways.

Current approaches for pathway analyses are restricted to genes and fail to incorporate non-coding regulatory elements. To fill this gap, TriPOINT (Triangulation of Perturbation Origins and Identification of Non-Coding Targets), a software designed to identify genes perturbed in pathways and non-coding regulatory elements regulating them was developed. TriPOINT offers a novel method for pathway analysis by identifying the genes that are the most affected under a condition by using multiple scoring metrics to uncover the impact of a gene’s perturbation on the network and by providing the ability to integrate these genes with non-coding regions using chromatin interaction datasets. TriPOINT is an easy to use and flexible
tool for furthering existing methods for pathway analyses, which can lead to the identification of not only the most relevant genes for a phenotype but also their non-coding regulators.

### 3.2 Software & Implementation

TriPOINT was implemented in Java, incorporating pathway graphs from the GRAPHITE\textsuperscript{63} R package through RServe\textsuperscript{32}. Methods from QuIN\textsuperscript{21} were utilized to integrate chromatin interaction data to identify non-coding regulators. Figure 3.2 shows an overview of software and data used in TriPOINT. Finally, the Cytoscape\textsuperscript{31} Java application was used as a platform for visualization of TriPOINT JSON files which are easily imported and display pathways augmented with differential expression values and non-coding regulator information (see Figure 3.3 for an example).
3.3 Metrics for Identifying Pathway Perturbations

Four metrics were employed to identify perturbations in graph representations of pathways from gene expression: inconsistency, support, consistency, and impact. Let $U_s(g, p)$ define the set of all genes represented as nodes in the pathway graph that target a gene $g$ in pathway $p$ with an edge distance of 1. The inconsistency score defined in (6) measures the differential gene expression ($\text{Exp}(g)$) of immediate upstream genes ($U_s(g, p)$) of a target gene $g$ in a pathway $p$, and quantifies the number of times the expression of $g$ is going against the activation or inhibition relationships as shown in Figure 3.1.
The influence of differential gene expression is controlled by parameter $T_u$ while parameters $T_d$ and $T_u$ provide thresholds to determine whether or not genes are up-regulated or down-regulated respectively. Parameter $w$ controls the differences between strong and weak inconsistencies. Similar to inconsistency, the support score defined in (3.2) quantifies the number of times a target gene is supporting activation and inhibition relationships of upstream genes.

$$\text{Impact}(g_i, p_j) = \sum_{g_j \in \text{Us}(g_i, p_j)} i_s | \text{Exp}(g_j) | + w \sum_{g_j \in \text{Us}(g_i, p_j)} i_w | \text{Exp}(g_j) |$$

\[
i_s = \begin{cases} 
1 & \text{if } E_{\text{type}}(g_j, g_i, p_i) = \text{ACTIVATION} \\
1 & \text{if } E_{\text{type}}(g_j, g_i, p_i) = \text{INHIBITION} \\
0 & \text{otherwise}
\end{cases} \\
\text{Exp}(g_j) > T_u \text{ and } \text{Exp}(g_j) < T_d \\
\text{Exp}(g_j) > T_u \text{ and } \text{Exp}(g_j) < T_d
\]

\[
i_w = \begin{cases} 
1 & \text{if } E_{\text{type}}(g_j, g_i, p_i) = \text{ACTIVATION} \\
1 & \text{if } E_{\text{type}}(g_j, g_i, p_i) = \text{INHIBITION} \\
0 & \text{otherwise}
\end{cases} \\
\text{Exp}(g_j) < T_d \text{ and } \text{Exp}(g_j) > T_u \\
\text{Exp}(g_j) < T_d \text{ and } \text{Exp}(g_j) > T_u
\]

(3.1)

$$\text{Support}(g_i, p_j) = \sum_{g_k \in \text{Us}(g_i, p_j)} s_s | \text{Exp}(g_k) | + w \sum_{g_k \in \text{Us}(g_i, p_j)} s_w | \text{Exp}(g_k) |$$

\[
s_s = \begin{cases} 
1 & \text{if } E_{\text{type}}(g_k, g_i, p_i) = \text{ACTIVATION} \\
1 & \text{if } E_{\text{type}}(g_k, g_i, p_i) = \text{INHIBITION} \\
0 & \text{otherwise}
\end{cases} \\
\text{Exp}(g_k) > T_u \text{ and } \text{Exp}(g_k) > T_u \\
\text{Exp}(g_k) > T_u \text{ and } \text{Exp}(g_k) > T_u
\]

\[
s_w = \begin{cases} 
1 & \text{if } E_{\text{type}}(g_k, g_i, p_i) = \text{ACTIVATION} \\
1 & \text{if } E_{\text{type}}(g_k, g_i, p_i) = \text{INHIBITION} \\
0 & \text{otherwise}
\end{cases} \\
\text{Exp}(g_k) < T_d \text{ and } \text{Exp}(g_k) > T_u \\
\text{Exp}(g_k) < T_d \text{ and } \text{Exp}(g_k) > T_u
\]

(3.2)

Combined, the inconsistency and support scores define the consistency score (3.3) where negative values reflect perturbed genes and positive values reflect genes following the expected expression pattern in the pathway.

$$\text{Consistency}(g_i, p_j) = S(g_i, p_j) \cdot I(g_i, p_j)$$

(3.3)
Finally, the *impact* score defined in (3.4) quantifies the downstream effect of a gene’s expression. Expression values of downstream genes are normalized with respect to their edge distance $d(g_1, g_2, p)$, which defines the number of edges between two genes $g_1$ and $g_2$ in pathway $p$ using exponential decay with decay rate $r$.

$$\text{Impact}(g_i, p_i) = \sum_{g_m \in Ds(g_i, p_i)} \frac{|\text{Exp}(g_m)|}{e^{r d(g_i, g_m)}}$$ (3.4)

Downstream genes, represented by $Ds(g, p)$, are defined as the set of genes in the largest subgraph (identified using breadth first search) such that all genes are consistent (i.e., supporting the activation/inhibition relationships of their source genes) starting from the root gene $g$. The combination of *consistency* and *impact* scores identifies perturbed genes with the highest downstream impact. To evaluate the significance of each score, permuted $p$-values are calculated, in which the base line is created by randomly reassigning differential gene expression values.

### 3.4 Identification of Non-Coding Regulators

Two approaches for incorporating non-coding regulators into graph representations defining pathways were implemented. The first approach utilizes chromatin interaction loops from genome-wide assays such as ChIA-PET$^{10}$, HiC$^8$, or HiChIP$^9$ datasets. Methods available from QuIN$^{21}$ were employed to construct a chromatin interaction network to identify loci directly interacting with genes in a pathway. If chromatin interaction data is not yet available for the given cell type, non-coding regulators are identified based on proximity, assigning non-coding regions provided by the user to genes within a user-defined distance from their transcription start site. $P$-values relating to the significance of the number of non-coding regulators targeting a gene are calculated based on the Poisson distribution.
3.5 Identification of Perturbed Gene RALGDS and its Non-Coding Regulators

TriPOINT was demonstrated in a case study analyzing all stage-one breast cancer RNA-seq samples from females older than 50 (n=7) profiled by The Cancer Genome Atlas\textsuperscript{64}, which were obtained through the NCI Genomic Data Commons\textsuperscript{65} portal. Differentially expressed genes were identified for each individual by comparing the individual’s cancer sample with their matching normal sample using DESeq2\textsuperscript{66}. TriPOINT was employed on the differential expression data using KEGG\textsuperscript{67-69} pathway graphs available from GRAPHITE\textsuperscript{63}. Non-coding regulators were included in these graphs by integrating MCF-7 (an early stage breast cancer cell
line) DNASE-Seq (GSE32970) and RNA-Pol2 ChIA-PET (GSE39495) datasets from ENCODE\textsuperscript{2,70}. Genes/pathways were then selected with a p-value cutoff of 0.1 for consistency, impact, and number of non-coding regulators to identify the most impactful genes. In five out of seven stage-one breast cancer patients, \textit{RALGDS} was identified as a perturbed gene in the “choline metabolism in cancer” pathway, showing interactions with two non-coding regulators. Lack of \textit{RALGDS} has been previously shown to reduce tumor incidence in mice\textsuperscript{71}. Upon further inspection, \textit{RALGDS} was found to be down-regulated in six of the seven patients studied (Figure 3.3 and Figure 3.4). In the case where \textit{RALGDS} was up-regulated, \textit{PLD1}, a gene associated with cancer cell survival\textsuperscript{72} and activated by \textit{RALGDS}, was also up-regulated, suggesting \textit{RALGDS} may be in direct control of \textit{PLD1} activation but regulated by other unknown factors. To investigate the non-coding targets of \textit{RALGDS} further, each loci was integrated with stretch enhancers identified in MCF-7\textsuperscript{13}. Stretch enhancers have been shown to be more cell-specific and more likely to harbor disease causing variants\textsuperscript{13}. A 6kb stretch enhancer at chr9:136005000-136019600 was found to harbor the non-coding target chr9:136009033-136009882 of \textit{RALGDS}, revealing additional evidence the concerned loci is possibly in control of the gene’s expression and merits further experimental study.
Figure 3.4. *RALGDS* in the Choline Metabolism KEGG Pathway for Six Early Stage Breast Cancer Patients. Sub-graphs of the Choline Metabolism in Cancer KEGG pathway in the remaining six patients from TCGA. Gene node colors represent the log2 fold change of gene expression where green nodes represent down-regulated genes and red nodes represent up-regulated genes. Yellow nodes represent non-coding targets. Edges in the pathway represent activation and inhibition interactions between genes. Note the last patient is the only patient where *RALGDS* is being shown as activated.
3.6 Discussion

TriPOINT bridges the gap between non-coding regulators and pathways, opening the door for studying non-coding variants in pathway analyses. The capabilities of TriPOINT were demonstrated in a case study of early stage breast cancer samples, showing the usefulness of TriPOINT in connecting non-coding factors to pathway analyses and prioritizing genes in pathways, bringing closer a more complete picture of underlying mechanisms in the control of expression by uncovering potential therapeutic targets through data integration. TriPOINT is but a first step in Pathway analyses involving non-coding regulators. Further improvements can be made in two major ways. First, sequence motifs of these non-coding regulators can be used to identify transcription factors to further enrich pathways, updating them for cell-specific or individual specific contexts. Second, perturbation scores can be further improved to further incorporate non-coding regulator information of not only direct interactions, but also indirect ones. TriPOINT software and its source code released under the GNU General Public License Version 3 are available at: https://github.uconn.edu/ajt06004/TriPOINT.
Chapter 4

Inference of Cell-Type-Specific Regulatory Elements from Chromatin Interaction Networks

4.1 Introduction

Cell-type-specific functions of super enhancers and broad domains have been extensively studied and well established across diverse cell types and organisms\(^ {11,12,73,74}\), where their distinct epigenomic profiles were instrumental in their discovery. Super enhancers were demarcated by high levels of enhancer-associated histone modification mark H3 lysine 27 acetylation (H3K27ac) and were catalogued in 86 human cell and tissue types using this mark\(^ {73}\).

Pharmacological molecules have been used to effectively and specifically target super enhancer domains at oncogenes\(^ {75}\), further reinforcing their significance for disease biology. Similarly, cell type-specific promoters identified as broad domains are associated with expanded deposition of

\(^{2}\) Sections of this chapter were previously published in Scientific Reports\(^ {59}\) under the Creative Commons Attribution License 4.0 [https://creativecommons.org/licenses/by/4.0/](https://creativecommons.org/licenses/by/4.0/) and modified for this dissertation.
histone H3 lysine 4 tri-methylation (H3K4me3) mark - a signature conserved across diverse cell types (>99 in human cells) and organisms\textsuperscript{11}. Shortening of broad domains has been observed in cancer cells at tumor suppressor genes, enabling the discovery of novel tumor suppressors\textsuperscript{74}. Recently, super enhancers and broad domains overlapping super enhancers were shown to be more associated with chromatin interactions than their typical counterparts\textsuperscript{76} suggesting a unique organization of chromatin around cell-specific loci.

Chromatin structure plays a major role in governing cellular functions in a cell type- and condition-specific manner\textsuperscript{77}. Advances in genomewide chromatin interaction profiling have shown that many regulatory elements (i.e., enhancers and promoters) that are distal on the linear genome map are actually in close physical proximity with each other as a result of the 3D chromatin structure\textsuperscript{8,10,78}. Among these technologies, the Chromatin Interaction Analysis by Paired-End Tag Sequencing (ChIA-PET) combines chromatin immunoprecipitation with chromatin conformation capture to identify chromatin interactions that are mediated by a protein\textsuperscript{10}, such as RNA Polymerase II (Pol2) which mediates interactions between promoters and enhancers\textsuperscript{79}. More recently, an alternative method has been developed, HiChIP\textsuperscript{9}, to detect protein-centric chromatin interactions using 100-fold less input material, providing an opportunity to generate such maps in primary human cells and tissues. These datasets, particularly the ones capturing protein-mediated promoter and enhancer interactions enable genomewide study of chromatin interactions between broad domains and super enhancers.

This study utilizes advanced computational methods to uncover how broad domains and super enhancers interact in the 3D chromatin space. In particular, questions of whether these elements are associated with distinct connectivity patterns, whether these patterns are conserved across cell types and assays, and whether they are predictive of the cell-specific nature of
promoters and enhancers were each addressed. Chromatin interaction networks were built using diverse assays including ChIA-PET, Hi-C, and HiCHiP in three ENCODE cell lines: MCF-7 (breast adenocarcinoma), K562 (chronic myeloid leukemia), and GM12878 (lymphoblastoid cell line). These networks were annotated using ChromHMM states\(^3\),\(^\text{13}\) super enhancer\(^7\), and broad domain\(^1\)\(^\text{11}\) definitions in the corresponding cell types. Interaction frequencies, network centrality measures and graphlets\(^4\text{3}\) were studied to uncover distinct connectivity patterns associated with broad domains and super enhancers. Using machine learning models based on support vector machines (SVM)\(^\text{80},\text{81}\) it was shown that these chromatin connectivity patterns could effectively discriminate broad domains from regular promoters and super enhancers from regular enhancers. Overall these results suggested a unique and conserved chromatin organization around critical regulatory elements.

### 4.2 Chromatin Interaction Networks from Diverse Cell Lines and Assays

Chromatin interaction networks were built using Pol2 ChIA-PET data in three ENCODE cell lines: MCF-7, K562, and GM12878 and using Hi-C\(^\text{82}\) and HiChIP\(^9\) (targeting cohesion subunit Smc1a) data in GM12878. Pol2 ChIA-PET chromatin interactions\(^2,\text{60}\) were obtained for MCF-7 (GSE39495), K562 (GSE39495) and GM12878 (GSE72816). To minimize the number of false positive interactions, QuIN\(^2\)\(^\text{11}\) was employed to filter interaction calls, selecting only those where both interaction anchor overlapped DNase I hypersensitive sites (DHS) defined from DNASE-seq open chromatin peaks called using MACS2 software\(^\text{83}\) (version 2.1) after pooling replicates (GSE32970 and GSE29692) from ENCODE\(^2\). Pol2 ChIA-PET (and CTCF ChIA-PET for comparisons) Networks were constructed from intra-chromosome interactions, using 250bp extensions of accessible chromatin peaks, which also define network nodes, to facilitate capturing edges in the network by overlapping interaction anchors of interactions no greater than
1Mb in base pairs. Interaction pairs from HiChIP\(^9\) and contacts from HiC\(^8\) for GM12878 (GSE80820 and GSE63525 respectively) were obtained and processed independently based on their different protocols and resolution. Specifics on interaction calling can be found in Appendix B. Using open chromatin peaks as before for Pol2 ChIA-PET networks in GM12878, networks were constructed using QuIN\(^2\) using 0bp and 1,250bp extension parameters for HiChIP and Hi-C networks, respectively. These values were chosen to account for differences in read extension introduced in interaction calling steps. Resulting networks consisted of 20-50 thousand network nodes/edges and thousands of connected components (Table 4.1).

<table>
<thead>
<tr>
<th>Cell-line</th>
<th>Assay</th>
<th>Antibody</th>
<th># of CCs</th>
<th># of Nodes</th>
<th># of Edges</th>
<th>Avg PET per edge</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCF-7</td>
<td>ChIA-PET</td>
<td>Pol2</td>
<td>4,087</td>
<td>22,340</td>
<td>23,542</td>
<td>4.04</td>
</tr>
<tr>
<td>K562</td>
<td>ChIA-PET</td>
<td>Pol2</td>
<td>2,927</td>
<td>26,915</td>
<td>37,334</td>
<td>5.32</td>
</tr>
<tr>
<td>GM12878</td>
<td>ChIA-PET</td>
<td>Pol2</td>
<td>3,975</td>
<td>28,574</td>
<td>38,149</td>
<td>11.03</td>
</tr>
<tr>
<td>MCF-7</td>
<td>ChIA-PET</td>
<td>CTCF</td>
<td>5,123</td>
<td>36,032</td>
<td>40,261</td>
<td>5.43</td>
</tr>
<tr>
<td>K562</td>
<td>ChIA-PET</td>
<td>CTCF</td>
<td>2,995</td>
<td>16,578</td>
<td>16,385</td>
<td>4.18</td>
</tr>
<tr>
<td>GM12878</td>
<td>ChIA-PET</td>
<td>CTCF</td>
<td>2,383</td>
<td>30,851</td>
<td>44,501</td>
<td>20.69</td>
</tr>
<tr>
<td>GM12878</td>
<td>Hi-ChIP</td>
<td>Smc1a</td>
<td>6,077</td>
<td>38,548</td>
<td>48,578</td>
<td>6.81</td>
</tr>
<tr>
<td>GM12878</td>
<td>Hi-C</td>
<td></td>
<td>11,120</td>
<td>46,962</td>
<td>39,374</td>
<td>25.41</td>
</tr>
</tbody>
</table>

Table 4.1. Network statistics for chromatin interaction networks built from ChIA-PET, HiChIP, and Hi-C interactions in three ENCODE cell lines.

### 4.3 Chromatin Interaction Network Annotation

Nodes in these networks were annotated using ChromHMM states\(^3,13\) in conjunction with broad domain\(^11\) and super enhancer\(^73\) definitions in corresponding cell types. Network nodes were first annotated using ChromHMM states\(^3,13\). If a node overlapped with multiple ChromHMM annotations, the following priority schema was applied to dissolve ambiguities in annotations due to the genomic segmentation framework in ChromHMM: 1) enhancers/promoters 2) insulators 3) poised promoters 4) repressed elements 5) transcribed elements 6) low signal. If a node was annotated both as a promoter and an enhancer, known transcription start site (TSS) definitions were used to define promoter nodes found within 2kb of a known TSS. Broad
domains and super enhancers annotations were then assigned respectively to promoter and enhancer nodes using the previously defined broad domain\textsuperscript{11} and super enhancer\textsuperscript{73} regions in corresponding cells. If a broad domain was found to overlap multiple promoter nodes, the promoter node with the largest overlap with the broad domain in terms of base pairs (bps) was assigned as the broad domain node. In the case of ties, both promoters were labeled as broad domain nodes.

**Figure 4.1.** Regulatory annotations of Pol2 ChIA-PET network nodes.

In Pol2 ChIA-PET networks, a majority of nodes (68-80\%) overlapped promoters and enhancers, showing the utility of Pol2-mediated ChIA-PET interactions to capture interactions between regulatory elements\textsuperscript{77} (Figure 4.1). Majority of super enhancers and broad domains (>70\%) were represented in these networks (Figure 4.2), in agreement with recent reports on super enhancers being more involved in chromatin interactions\textsuperscript{76}. It was also observed that Hi-C networks included fewer (~25-39\% fewer) promoters, broad domains and super enhancers compared to Pol2-associated assays (Figure 4.3a-b), likely due to Hi-C capturing all DNA-DNA contacts.
Figure 4.2. Number of broad domains (inner chart) and super enhancers (outer chart) represented in ChIA-PET networks.

Figure 4.3. Representation of regulatory elements and interaction frequencies in Hi-ChIP and Hi-C networks. (a) Distribution of different annotations in Hi-ChIP (top) and Hi-C (bottom) networks in GM12878. (b) Distribution of broad domains (inner chart) and super enhancers (outer chart) in HiChIP (top) and Hi-C (bottom) networks in GM12878.

4.4 Increased Interaction Frequency Among Broad Domains and Super Enhancers

Frequency of interactions were calculated between all pairs of annotations including broad domains, typical promoters, super enhancers, typical enhancers, and other annotations, and
compared against theoretical expectations in QuIN\textsuperscript{21}. These analyses showed that in Pol2 ChIA-PET networks, broad domains were more connected to all other nodes than theoretically expected (2.9 times more than expected) in all three cell lines (Figure 4.4).

Furthermore, super enhancers interacted more frequently with broad domains (2.7-5.5 times more than expected) across the three cell types (Figure 4.4). Interestingly, super enhancer nodes also interacted more frequently among themselves (2.7-5 times more than expected), raising the possibility that distinct enhancer elements within a super enhancer region form highly interacting enhancer clusters in the 3D space. Further investigation of super enhancer-super enhancer interactions revealed that most of these (60-90\%) take place within the same super enhancer region (Figure 4.5a). Analyses were repeated to account for interactions within a single super enhancer region by representing the multiple nodes that belong to the same super enhancer domain as a single node. For this, separate interactions networks were created by merging super enhancer nodes that belong to the same super enhancer domain and representing the whole domain as a single node in the network. The number of edges that were connecting these merged nodes was counted and the interaction frequency analyses were repeated. After this adjustment,
enrichment of interactions among super enhancer nodes were mostly lost (Figure 4.5b). These analyses suggest that constituent enhancers within a super enhancer domain are in close proximity in the 3D chromatin space, however, these interactions do not typically span multiple distinct super enhancer domains. Finally, it was noted that an enrichment of interactions among promoter elements (both cell-specific and non-specific) was observed (1.9-5.0 fold over expected, Figure 4.4). HiChIP and Hi-C ChIA-PET assays revealed similar interaction frequency patterns: i) high interactions between broad domains and super enhancers, ii) high interactions among constituent enhancers of super enhancer regions (Figure 4.6). Robustness of these results across assays and across cell types suggests a strong link between 3D configuration of the genome and distinct characteristics of regulatory elements.

Figure 4.5. (a) (Top) Distribution of interactions within (turquoise) and across distinct (orange) super enhancer regions. (Bottom) Illustration of super enhancer nodes as defined by DNase-seq peaks within original super enhancer calls and the two different types of super enhancer interactions. (b) Interaction frequency of super enhancers after merging super enhancer nodes belonging to the same super enhancer region.
Figure 4.6. HiChIP (Smc1a) and Hi-C Interaction Frequencies. Enrichment of interactions between pairs of annotation classes in HiChIP (subunit Smc1a) and Hi-C networks. Colors represent log2 ratio of observed over expected number of edges, where red represents enrichment of interactions and blue represents depletion of interactions.

4.5 Broad Domains and Super Enhancers are Hubs in Interaction Networks

Network centrality measures suggest that cell-specific regulatory elements are more connected and exhibit hub-like connectivity in these networks in comparison to their typical counterparts (Figure 4.7). On average, promoters were connected to 2.63, 4.21, and 3.56 other nodes in MCF-7, K562, and GM12878 Pol2 ChIA-PET networks respectively, whereas the corresponding values for broad domains were 5.03, 5.83, and 4.49 (one-sided Wilcoxon test p-values < 4.4.e-32 for all three cell lines). The increased connectivity of broad domain promoters taken together with their frequent interactions with super enhancers might be essential in maintaining their robust and increased gene expression patterns. On the other hand, enhancer nodes were connected to an average of 1.81, 1.90, 2.25 other nodes respectively in MCF-7, K562 and GM12878, whereas super enhancer nodes averaged 4.61, 3.27, and 4.22 connections respectively (one-sided Wilcoxon test p-values < 5.1.e-114 for all comparisons). These results also revealed a higher betweenness score for broad domains relative to non-specific promoters (Figure 4.7) suggesting that broad domains act as connectors in the networks. For example, Figure 4.8 shows Pol2 ChIA-PET network involving the EMP2 oncogene that is upregulated in invasive breast
In the breast cancer cell line MCF-7, EMP2 maps to a broad domain node and is connected to multiple super enhancers (Figure 4.8, left panel), which are also connected to each other. In contrast, in K562 where EMP2 is an active yet non-specific promoter, this locus is connected differently and less densely (Figure 4.8, right panel). On the other hand, this locus was repressed in GM12878, and was not represented in the corresponding networks. This example illustrates that connectivity of a locus in the 3D chromatin reflects the functional importance of that region in the cognate cell type. Overall these analyses revealed that cell-specific regulatory elements are connected more frequently in the 3D genome in comparison to their non-specific counterparts.

**Figure 4.7. Centrality Score Distributions.** Distribution of centrality scores (connectivity degree, closeness, harmonic, normalized betweenness centrality) for different annotation classes. M, K, G represents MCF-7, K562, and GM12878 Pol2 ChIA-PET networks respectively. C and H represent HiChIP and Hi-C GM12878 networks respectively.
4.6 Broad Domains and Super Enhancers Have Unique Connectivity Patterns

Chromatin interaction networks offer the opportunity to explore higher-level chromatin connectivity patterns or network motifs beyond immediate interactions. However, enumerating all possible configurations in a large network is computationally intractable (subgraph isomorphism problem). To effectively and systematically uncover chromatin interaction patterns associated with cell-specific regulatory elements, graphlet orbit counts were measured using QuIN [21]. Graphlets are small, connected, and non-isomorphic subnetworks within a large network that enables systematically studying and quantifying the local network structure around a node of interest. The local structure of chromatin interaction networks was studied using all possible graphlets composed of two to five nodes, which encompassed 73 orbits (Figure 4.9a). For each node an orbit signature vector was compiled by counting the number of times each node possessed the local structure of the 73 orbits. Orbits were then clustered to account for their topological similarities revealing seven major orbit clusters that represent topologically distinct types of orbits (hierarchical clustering, Spearman coefficient cutoff = 0.3) (Figure 4.9b). For example, Cluster 1, C1, is composed of orbits occupying a central position across various graphlets (red nodes in Figure 4.9a). Therefore, a node that has a high C1 score occupies a central position in its chromatin interaction network. For each network node we calculated their
orbit cluster scores (n=7 scores) using 4.1 where $CS_{ki}$ denotes the score of node $i$ for cluster $k$, $O_{ij}$ denotes the orbit score for node $i$ and orbit $j$, $C_k$ denotes the set of orbits in cluster $k$, and $\mu_j$ and $\sigma_j$ represent the mean and standard deviation of orbit $j$'s score over all nodes.

$$CS_{ki} = \max_{j \in C_k} \left( \frac{O_{ij}}{j} \right)$$ (4.1)
example red nodes represent central nodes that are in cluster 1. (b) Hierarchical clustering of 73 orbits after pooling data from all Pol2 ChIA-PET networks. Seven distinct clusters are identified based on the topology of network graphlets.

Cluster scores allowed for systematically assessing recurrent chromatin interaction patterns associated with specific regulatory elements. Promoters (regular and broad domain promoters) in general had higher C1 scores than enhancers, indicating that they are more likely to be at the center of chromatin interaction networks (Figure 4.10). Broad domains held the most central positions in these networks, in agreement with their high connectivity degree (Figure 4.10). On the other hand, among the enhancer elements, super enhancers had a higher C1 score, implying they are more centrally located than regular enhancers (Figure 4.10). Furthermore, super enhancer nodes exhibited more clique-like structures (e.g. triangle, cycle, and mesh patterns) than typical enhancers as evident from their higher C2, C3, and C4 scores (Figure 4.10).

Strikingly, orbit cluster scores of different functional elements showed very consistent patterns across cell types and across different assays (Figure 4.11), suggesting that these connectivity patterns are not stochastic and have functional relevance. These results suggest that cell type-specific regulatory elements have unique connectivity patterns and tend to be central and form tightly connected sub-networks in the 3D chromatin space.
Figure 4.10. Seven Orbit Cluster Scores for Different Regulatory Elements. Trimmed mean cluster scores for ChIA-PET (MCF-7, K562, GM12878), Hi-ChIP (GM12878), and Hi-C (GM12878) networks for promoters (P), broad domains (BD), enhancers (E), super enhancers (SE) and other regulatory elements (O). Note that these measures are very similar across cell types and assays.

Figure 4.11. Cluster Scores are Consistent Across Cell Type. Pairwise correlations (Pearson’s coefficient) of 7 cluster scores computed over each annotation between all network pairs.
4.7 Network Connectivity Patterns are Predictive of Cell-Specific Activity

To determine whether chromatin connectivity patterns of regulatory elements can be predictive of their cell-specific activity, support vector machine\(^{80,81}\) (SVM) machine-learning models were employed on node features. Each network node was represented using two different types of features: (1) network related and (2) genomic-data related and SVM-based classification models were built to discriminate i) broad domains from regular promoters and ii) super enhancers from regular enhancers. The discriminative power of these models was quantified using receiver operating characteristic (ROC) curves and area under these curves (AUC).

![Enhancer vs. Super Enhancer](image)

Figure 4.12. SVM models to discriminate enhancers and super enhancers using Pol2 ChIA-PET data. Receiver Operating Characteristic (ROC) curves (top) and precision recall curves (bottom) for SVM models separating enhancers from super enhancers using ChIA-PET networks (baseline performance shown in gray). AUC: area under the curve. Colors represent different data features used in the models.
Figure 4.13. SVM models to discriminate promoters and broad domains using Pol2 ChIA-PET data. Receiver Operating Characteristic (ROC) curves (top) and precision recall curves (bottom) for SVM models separating promoters from broad domains for ChIA-PET networks (baseline performance shown in gray). AUC: area under the curve. Colors represent different data features used in the models.

SVM models efficiently discriminated super enhancers from regular enhancers with high accuracy (Accuracy = 0.91, 0.74, 0.84 at 0.2 probability threshold, AUC = 0.84, 0.72, 0.81 in MCF-7, K562, and GM12878 Pol2 ChIA-PET networks respectively) (Figure 4.12). Similarly, these models were also effective in discriminating broad domain promoters from regular promoters (AUC scores = 0.77, 0.70, 0.71 for MCF-7, K562, and GM12878 respectively) (Figure 4.13). Furthermore, both analyses revealed that integration of network related features, such as orbit cluster scores, with other genomic features improved the predictive ability of these models, suggesting that chromatin interaction networks and connectivity patterns harbor functional and non-redundant information. Prediction of super enhancers in HiChIP and Hi-C
was as effective (AUC= 0.81 and 0.76 respectively) (Figure 4.14). Similar results for broad domain classification models were achieved from Hi-C and HiChIP data (Figure 4.15), although the relatively smaller impact of network features in these analyses suggests that Pol2-mediated interactions may be better suited for capturing network patterns associated with promoters. Precision-recall curves (Figures 4.12-4.15) further emphasized the value of integrating network and genomic data features and the ability of these datasets to predict cell-specific regulatory elements.
Figure 4.14. SVM models to discriminate enhancers and super enhancers using HiChIP and Hi-C data. Receiver Operating Characteristic (ROC) curves (top) and precision recall curves (bottom) for SVM models separating enhancers from super enhancers for HiChIP (left) and Hi-C (right) GM12878 networks (baseline performance shown in gray). AUC: area under the curve.
Figure 4.15. SVM models to discriminate promoters and broad domains using HiChIP and Hi-C data. Receiver Operating Characteristic (ROC) curves (top) and precision recall curves (bottom) for SVM models separating promoters from broad domains for HiChIP (left) and Hi-C (right) GM12878 networks (baseline performance shown in gray). AUC: area under the curve.

The most predictive data features in these models were obtained using forward selection, which uncovered consistent ranking of discriminatory features. Forward selection was employed by adding features to the SVM model one at a time, choosing the feature with the best increase in performance using Matthews correlation coefficient (a performance measure that considers both class labels equally even in the presence of imbalances) in each round, until all features were
included. For broad domain predictions, the two most predictive features were node size and network features related to their centrality (Figure 4.16). Further investigation revealed that indeed broad domains are associated with expanded chromatin accessibility around their promoters (Figure 4.17), in agreement with the expanded H3K4me3 deposition observed at these loci. On the other hand, the most predictive features for super enhancers were related to their clique-like connectivity and high degree, reinforcing the importance of tight connectivity around super enhancers. In summary, results from these analyses showed that network connectivity patterns of a regulatory element are predictive of its importance for regulating critical cellular functions in that cell type.
Figure 4.16. Feature ranking for SVM models to discriminate promoters and broad domains. (a) Feature importance ranking for all networks derived from forward selection for broad domain prediction. Forward selection incrementally includes features with the highest Matthews correlation coefficient score in each step. (b) Performance of training models with individual features. Features are divided into network related features (blue labels) and genomic-data related features (green labels). Note that the most predictive features for broad domains are node size, and centrality related measures (degree, cluster 1 score).
Figure 4.17. Feature ranking for SVM models to discriminate enhancers and super enhancers. (a) Feature importance ranking for all networks derived from forward selection for super enhancer prediction. Forward selection incrementally includes features with the highest Matthews correlation coefficient score in each step. (b) Performance of training models with individual features. Features are divided into network related features (blue labels) and genomic-data related features (green labels). Note that the most important features are associated with clique-like patterns associated with super enhancers.
4.8 Discussion

Broad domains and super enhancers were shown to have distinct connectivity patterns that are conserved across cell types and captured using diverse assays. One key observation was that broad domains tended to be central nodes in these networks and were frequently targeted by constituent enhancers of a super enhancer. High connectivity among super enhancer elements and high connectivity between broad domains and constituent super enhancers were observed. This tightened connectivity among cell-specific regulatory elements ensures that the activity of super enhancer target genes, which are likely to be broad domains, is robust to the disruption of any single enhancer element within the enhancer cluster. Such a tightened connectivity pattern may be critical in establishing and maintaining robust expression patterns associated with broad domain genes\textsuperscript{11}, and offers explanation as to why machine learning models were able to identify super enhancers and broad domains.

Recent studies revealed that regulatory elements with frequent chromatin interactions, also known as hubs, are enriched in super enhancers and harbor more GWAS SNPs\textsuperscript{76,85,86}. Observations in this study were in alignment with these findings, observing increased connectivity for super enhancer nodes in comparison to regular enhancers using three different assays (ChIA-PET, Hi-C, and HiChIP) and in three different cell types (K562, MCF-7, GM12878). This study furthered existing knowledge on how super enhancers are connected in these networks and revealed that super enhancers form clique-like structures and typically connect to broad domains and other super enhancer elements within the same domain. Broad domains were associated with more chromatin interactions compared to typical H3K4me3 domains\textsuperscript{76}. In comparison to regular promoters, broad domains had more interactions overall. Moreover, these interactions were especially connected to super enhancers. It is likely that
increased targeting of broad domains by super enhancers ensures robust and increased expression of these important genes in cognate cells.

A challenge faced in this study was the difficulty of representing super enhancers that span long genomic regions (ranging from 10-20kb) in networks, which would harbor more interactions merely due to their genomic coverage. To overcome this challenge, network nodes were defined using open chromatin regions from DNASE-seq, which represent putative active regulatory elements. This methodology also enabled the study of individual enhancer elements within a super enhancer domain. It was observed that a super enhancer element typically is not a single and expanded regulatory element but a combination of constituent active enhancers that are in close proximity to each other in both linear and 3D space, and frequently interacted with one another and with their target gene to regulate that gene’s expression levels. It was also shown that although members of a super enhancer domain heavily interact among themselves, these enhancers typically do not interact with other super enhancer domains.

In conclusion, findings in this study revealed that chromatin connectivity patterns around super enhancers and broad domains are non-stochastic and conserved across cell types and can be captured via different assays. However, this study’s main caveat is that genome-wide chromatin interaction maps analyzed here are generated from millions of cells. Therefore it is not possible to dissect whether the connectivity patterns observed for super enhancers and broad domains take place in individual cells. Advances in single-cell chromatin interaction profiling techniques will be essential in studying these patterns at the single cell level. Hi-ChIP\(^9\) was a step towards this direction as it significantly reduced the input material required to profile protein-mediated chromatin interactions: a 100 fold decrease from 100 million cells to 1 million cells. Similarly, recent developments in single-cell Hi-C profiling techniques open the doors to
studying cell-to-cell-heterogeneity for DNA-DNA interactions\textsuperscript{87}. Advanced computational methods presented in this study will be critical in furthering understanding on how chromatin interactions might relate to establishing and maintaining critical cellular functions and how changes in these interactions might be associated with pathologies.
Chapter 5

Inference of Enhancers from Accessible Chromatin: Predict Enhancers from ATAC-seq (PEAS)³

5.1 Introduction

Enhancers are non-coding cis-regulatory elements that precisely regulate gene expression patterns in the control of cell-type-specific functions⁴. In eukaryotic cells, regulation of gene expression occurs at multiple levels and results from a complex organization of enhancers serving as binding sites for transcription factors (TFs), which together determine if a particular gene will be active or silent. Epigenomic maps have been effective in enumerating enhancer sequences in human cells. For example, mono-methylation of lysine 4 on histone H3 (H3K4me1) and acetylation of lysine 27 on histone H3 (H3K27ac) have been shown to mark active enhancer sequences ⁸⁸. Similarly, the transcriptional co-activator EP300 has been effective in identifying putative enhancers⁸⁹,⁹⁰. Efforts from ENCODE² and Roadmap Epigenomics⁹¹ projects, have systematically profiled reference epigenomes from diverse human cells and computationally

³ Sections of this chapter are the result of contributions made in a collaborative project with Asli Uyar with shared interests.
described regulatory states, including putative enhancers in these cell types\textsuperscript{3,91,92}. However, epigenomes of a majority of human tissues and cell types have not been profiled in health and disease states. These epigenomic profiles are of particular importance, as the majority of disease-associated sequence variants discovered via genome-wide association studies (GWAS) are found in non-coding enhancer sequences, likely altering enhancer activity\textsuperscript{13,93}.

The Hidden Markov Model (HMM)–based ChromHMM algorithm\textsuperscript{3} is often the preferred tool to obtain functional annotations by segmenting the genome into bins and annotating each bin with a chromatin state, where each state is associated with a distinct and combinatorial histone modification profile. ChromHMM has been effective in finding regulatory elements in human cell lines in the existence of multiple histone modification profiles\textsuperscript{91}. However, generating multiple ChIP-seq profiles, where each chromatin immunoprecipitation typically requires millions of cells, is not feasible for clinical samples. As an alternative, Assay for Transposase Accessible Chromatin (ATAC-seq) technology is able to interrogate chromatin accessibility from small cell numbers\textsuperscript{41,42} and has been used to profile chromatin accessibility of diverse clinical samples including immune cells\textsuperscript{94} and in pancreatic islets\textsuperscript{95}.

ATAC-seq is ideal for studying enhancers in clinically relevant human cells and across individuals, since active and poised enhancers are associated with accessible chromatin. However, ATAC-seq peaks include all active DNA sites including promoters, enhancers, insulators, protein binding sites, and potential false positive calls. To further classify these regions, a machine-learning framework based on neural networks (PEAS: Predicting Enhancers from ATAC-Seq data) was developed to infer genomic locations of active enhancers from ATAC-seq profiles. PEAS differs from previous enhancer prediction tools since it uses ATAC-seq data as the only genomic measurement. Models were demonstrated in GM12878, CD4+ T
cells, CD14+ monocytes, peripheral blood mononuclear cells (PBMCs) and pancreatic islets to evaluate the performance of its predictions. In addition, PEAS was evaluated in two relevant scenarios including identifying enhancers at the individual level by studying ATAC-seq profiles of multiple human islets (n=16) and predicting enhancers on a new cell line, EndoC-βH1 beta cell line, in the absence of reference annotations.

5.2 Extracting Features from ATAC-Seq

Peaks called using nucleosome free reads (insert size <= 150bp) in MACS2\textsuperscript{96} undergo a multi-step process in order to extract 1627 features for cell-specific models and 26 features for cell-agnostic models. Peak-related features (n=5) including peak score, peak length, fold change, summit pileup and summit position were obtained from MACS2. Summit positions values were adjusted to define relative summit positions instead of genomic coordinates to improve effectiveness in machine learning algorithms. All reads (including those with insert size > 150bp) from ATAC-seq reads were analyzed once more to extract insert and cut driven features (n=11). ATAC-seq reads spanning a peak were used to quantify the number of inserts, the number of cuts, the mean and median insert size, and the ratio of reads with inserts greater than or equal to 150bp over the number of reads less than 150bp. In addition reads within peaks were broken down by insert length into five categories: short inserts (0,50bp], nucleosome free inserts (50,150bp], mono-nucleosome inserts (150,300bp], di-nucleosome inserts (300,500bp], and multi-nucleosome inserts for reads with inserts sizes greater than 500bp (500bp,). The final cut related feature attempts to identify positions within peaks cut significant more than if cuts were distributed uniformly across the entire peak length. For each 5bp window within the peak, significance testing (p-value < 0.0005) was performed using the binomial test with respect to the total number of cuts found within the peak with the expected probability of $5/l$ ($l$ = peak length)
for a cut falling with in the 5bp window. Conservation features (n=2) were identified using vertebrate conservation scores (phastcons46way) to identify the mean and max conservation scores overlapping each peak. HOMER\textsuperscript{25} was employed to identify features related to known gene information (n=3), sequence related features (n=4) and motif driven features (n=1604 for cell-specific models, n=3 for cell-agnostic models). Known gene information includes ‘Annotation’ (i.e., promoter, exon, intron, etc.), ‘gene type’, and ‘distance to TSS’ while sequence related features include CPG and GC content percentages within the peak. Motif driven features incorporate 1604 transcription factor position weight matrices (PWMs), counting the number of times a motif theoretically binds to a position within a peak based on HOMER’s default parameters. Where cell-specific models use all 1604 PWMs, since it is expected that the transcription factors within the cell type do not change, cell-agnostic models used aggregate measures features which include ‘% of known motifs’, ‘% of denovo motifs’, and ‘# of CTCF motifs’, where CTCF (a universal protein that organizes chromatin architecture\textsuperscript{97}) motif hits were determined from a subset of the 1604 known motifs and denovo motifs were called using methods available from HOMER\textsuperscript{25}. Percentages for known and denovo motifs were calculated by counting the number of motifs with > 1 hits within the peak and dividing by the total number of motifs.

### 5.3 Assigning Ground Truth Class Labels with ChromHMM

ChromHMM\textsuperscript{3} annotations were used to assign ‘ground truth’ labels to ATAC-seq peaks. To obtain consistent labels for model training (specifically across cell types), ChromHMM was used to segment the genome into 15 states using H3K4me1, H3K4me3, H3K9me3, H3K27ac, H3K27me3, and CTCF (when available) ChIP-Seq datasets for the 5 cell types studied: CD14+, GM12878, islets, PBMCs, and CD4+ T cells (accession numbers for each ChIP-seq and cell type
are available in Appendix C). To obtain harmonized class labels across cell types, pairwise Pearson correlations of emission probabilities of states were calculated for GM12878, islet, and CD4+ T cells, which included all cells where CTCF ChIP-seq was available. Clustering resulted in nine clusters where eight unique labels were assigned by comparing their emission probabilities to ENCODE/Roadmap ChromHMM state annotations and studying their histone mark combinations (Figure 5.1). Insulator labels were assigned to the state with the highest CTCF emission probability. Harmonized annotations for CD14+ and PBMCs, which lacked CTCF ChIP-seq, were obtained by excluding the insulator associated state from the other three cell types and repeating the clustering of emission probabilities. States that do not have a clear emission probability distribution for histone marks were labeled as ‘ambiguous states’.
Figure 5.1. Top: Pairwise Pearson correlations between ChromHMM models based on emission probabilities for 15 state models in three studied cell types that have the CTCF data: CD4+ T(C), Islet(I), and GM12878(G) cells. ChromHMM state numbers (1 to 15) and the cell type were depicted as row/column names. 9 major clusters were identified using hierarchical clustering, where each cluster is marked with a different colored box and labeled for its functional annotation (bottom right). Bottom: Emission probabilities from 15 state ChromHMM models using 7 different ChIP-Seq signals in each cell type.
ATAC-seq samples in different cell types/individuals were annotated using the ChromHMM annotations in the matching cell type. Typically ~30-40% of ATAC-seq peaks map to ChromHMM-defined enhancers, 30-40% to promoters, and the rest to other functional states (insulators, transcription related loci, etc.) (Figure 5.2). Furthermore, 5-20% of ATAC-seq peaks were ambiguously annotated (labeled as ‘Ambiguous State’ in Figure 5.2). Comparing features after annotating peaks, it was observed that enhancers have a different characteristic than promoters and other functional states when their ATAC-seq or sequence features are compared (Figure 5.3). For example, peaks mapping to ChromHMM-defined enhancers harbor more nucleosome free reads (insert size <150 bps) than peaks mapping to promoters or other states (Figure 5.3, column ‘# of Inserts (50,150]’). On the other hand, sequences mapping to enhancers are lower in GC ratio than promoters or other states (Figure 5.3, column ‘GC Ratio’).

Furthermore, these patterns were conserved across cell types, suggesting that although the enhancers are distributed across the genome in a cell-specific manner, their data characteristics were conserved. These results suggest that different regulatory elements have different ATAC-seq data characteristics, which can be exploited for building predictive machine-learning models.
Figure 5.1 ChromHMM Annotations Reveal Similar Annotation Distributions. Distribution of ChromHMM annotations for ATAC-seq peaks called in CD4+ T, GM12878, CD14+, PBMC, and 16 islet samples. For each analysis, annotations in the same cell type are used. Note that ATAC-seq peaks include promoters (~30-40%), enhancers (~30-40%), and other states.

Figure 5.2 Features extracted are discriminative of promoters and enhancers. Average difference between enhancer and promoter peaks in terms of different data characteristics (left) and enhancer and other peaks (right). Note that enhancers have different data characteristics that are conserved across cell types and individuals.

5.4 Neural Networks Discriminate Three Class Labels Across Five Cell Types

As an initial assessment of how well machine learning algorithms are able to discriminate different class labels, a single hidden layer (nodes=100) multilayer perceptron neural network
Neural networks were ideal as they inherently support non-binary classification. Five-fold cross validation was performed on 7 to 9-way classification, incorporating all available labels for each cell type. Across each cell type, three clusters of annotations were consistently observed: promoters, enhancers, and other regulatory elements (Figure 5.4). Promoters included active promoters and other types of promoters, enhancers included active enhancers, genic enhancers and other types of enhancers, and other regulatory elements included insulators, transcribed regions, polycomb repressed regions, and other regions as annotated from chromHMM. In light of this observation, future models were trained using 3-way classification.

Figure 5.4. Confusion matrices for 9-way classification in CD4+ T, GM12878, and islets and 8- and 7-way classification in CD14+ and PBMC cells respectively, since the latter lacks CTCF signal hence the insulator state. For each cell type, classifiers correctly classified promoters and enhancers, however didn’t separate other states from each other. Furthermore, classifiers also failed to separate between different versions of promoters and enhancers. Based on these results, it was concluded that classifiers...
are the most effective in discriminating promoter states (active, poised, other), enhancer states, and all remaining annotations (‘others’). Following analyses are conducted using this 3-way classification models.

5.5 Comparison of Machine Learning Models & Model Tuning

Six machine learning algorithms available from scikit-learn were evaluated using 5-fold cross validation: neural network, support vector machine, random forest, k-nearest neighbor, quadratic discriminate analysis, and naïve Bayes. Models were trained for each cell type with the cell-agnostic feature set (n=26). One vs. rest procedures were employed in binary classification cases. Area under the curve (AUC) values for the micro-average receiver operator characteristic (ROC) revealed that neural networks outperformed other algorithms for 3-way (promoter, enhancer, and other) classification (Figure 5.5). Neural networks outperformed other algorithms consistently across all cell types. Support vector machines achieved similar performance, however, the time required to train these models was substantially greater than neural networks (minutes for neural networks compared to hours/days for support vector machines) further justifying the use of neural networks for future analyses.
Figure 5.5. Neural Networks Outperform Other Machine Learning Models. ROC AUC values and accuracies based on five-fold cross-validation for 3-way classifications are shown using six different algorithms: neural network, support vector machines (SVM), random forest, k-nearest neighbor, quadratic discriminate analysis, and naïve Bayes.

To improve performance of neural network models, parameter tuning was performed for cell-agnostic models, training on CD14+, GM12878, islets, PBMCs, and CD4+ T cells and testing on 6 other individuals for CD14+, PBMCs cells and EndoC-βH1 beta cells. Overall 8748 different parameter configurations were evaluated, calculating the average Matthews correlation coefficient for non-binary classification\textsuperscript{100} for each configuration. Although slight improvements were achieved in CD14+ and PBMC individuals, comparing default parameter configurations to tuned parameter configurations suggested that overfitting was occurring to achieve these results and were therefore not suitable for the aim of identifying enhancers in new cell lines. Default parameters were thus used in the remainder of analyses.

To evaluate the performance of models trained on different feature sets, neural network models were trained on five different feature sets from the cell-agnostic features (n=26): 1) all features 2) peak related features, 3) insert/cut driven features 4) sequence related & motif driven
features, and 5) known gene information related features. For each feature set evaluated, values for other features sets were set to 0 in order to maintain the same topology within the neural network. Sequence features showed the highest performance (micro-average ROC AUC values=0.86-0.91) overall among the feature subsets, however inclusion of all features achieved the best performance (micro-average ROC AUC values=0.92-0.94), implying integrating diverse features improved model performance overall (Figure 5.6).
Figure 5.6. Integrating Diverse Features Improves Model Performance. ROC curves for different feature sets using 5-fold cross validation in 5 different cell types: CD4+T, GM12878, CD14+, PBMCs, and islets. All features (blue) outperformed models that were obtained by different sets of features. Area stands for AUC values, acc stand for accuracy of models.
5.6 Individual Specific Enhancer Prediction

ChromHMM\textsuperscript{3} annotations are typically performed on a select few individuals as multiple ChIP-seq experiments are expensive and require high cell counts (\(10^5\)-\(10^7\) cells). ATAC-seq solves this issue, as it requires as few as 500-50,000 cells, making it feasible and cost effective to sequence multiple/all individuals. PEAS provides the opportunity for individual specific enhancers to be refined from reference samples. To evaluate the degree to which PEAS can identify enhancers at the individual level, leave one out validation was performed on 16 different islet samples. For each islet sample, a model was trained on the remaining 15 samples using cell-specific features (n=1627), which include specific motif information, yielding 16 models in total. To remove errors in training data, training data only included peaks that were annotated with class label other than ‘ambiguous state’ and those with an overlap higher than 50\% with the annotation. Islet specific models were highly effective at capturing enhancers defined by ChromHMM (micro-average ROC AUC values=0.93-0.96 across 16 individuals) (Figure 5.7). In total, 28682 individual specific enhancers of 164842 peaks among the 16 individuals were predicted using PEAS with only 11292 found in ChromHMM annotations. The high classification performance and high yield of individual specific enhancers is suggestive that PEAS is effective for identifying enhancers at the individual level, however further biological validation is required to accurately measure the performance of individual specific enhancer prediction.
5.7 Cross Cell Type Enhancer Prediction

As previously mentioned, enhancers identified by algorithms such as ChromHMM\(^3\) are costly both in terms of cell count and in terms of number of experiments required. Cross cell type models were evaluated using PEAS to explore the predictive power of enhancers in absence of reference enhancer annotations. Models were trained on CD14+, GM12878, islets, PBMCs, and CD4+ cells using cell-agnostic features (n=26) and evaluated on the remaining cell types. Overall these models were moderately effective in predicting enhancers (Figure 5.8). On average, the islet model was the best predictor of enhancers in other cell types (micro-average ROC AUC values=0.8-0.9), however it was also observed that the GM12878 model
underperformed when testing on other cell types (micro-average ROC AUC values=0.82-0.88). When testing GM12878 on other cell type models, predictions in GM12878 were similar underperforming relative to the other cell types (micro-average ROC AUC values=0.68-0.81), suggesting depth of sequencing, differences between cell lines and primary cells or a combination of the two may be a factor in classification performance.

![Cross cell-type predictions](image)

**Figure 5.8. Cross Cell-Type Predictions.** Barplots of ROC AUC values for cross cell-type predictions. Models were trained entirely on one cell type alone (depicted below the barplots) and tested on the remaining four cell types.

### 5.8 Inference of Enhancers via Cell-Agnostic Models

Continuing to evaluate models in the absence of reference annotations and to overcome the challenges presented by differences in cells and experimental procedures, a cell-agnostic model was trained, combining data from multiple cell types. Features matrices from CD14+, GM12878, islets (Islet16), PBMCs, and CD4+ were combined after performing standardization methods to remove variability between cells and experimental procedure further. Once again, to remove errors in training data, training data only included peaks that were annotated with a class label other than ‘ambiguous state’ and those with an overlap higher than 50% with the annotation. Enhancer predictions using the cell-agnostic model were evaluated in EndoC-βH1 beta cells, using ChromHMM annotations in EndoC-βH1 only for model evaluation. The cell-agnostic model predicted EndoC-βH1 enhancers with high efficacy (micro-average ROC AUC=0.87), achieving higher performance than the cell specific model built only from islet data.
(micro-average ROC AUC=0.84) (Figure 5.9). For comparison purposes, a model was trained using EndoC-βH1 features and ChromHMM states for training using 5-fold cross validation, which as expected performed better than the other two models (micro-average ROC AUC=0.92) (Figure 5.9). Combined, these analyses suggest that cell-agnostic PEAS models are effective in predicting enhancers in cell types missing reference annotations. An example genome browser shot is provided (Figure 5.10), revealing that EndoC-βH1 chromHMM enhancers not present in the training cell types were accurately predicted using the cell-agnostic model, further suggesting that the cell-agnostic model is capable of identifying new loci harboring enhancers by working on the premise that the enhancer features used are conserved across cell types.

![Cell-Agnostic Model Performance on ENDOC](image)

**Figure 5.9. Cell-Agnostic Model Performance on ENDOC.** ROC curves of EndoC-βH1 predictions using i) EndoC-βH1 data with 5-fold cross validation (purple), ii) islet-specific models (green), and iii) cell-agnostic model (blue).
5.9 Discussion

The machine learning framework, PEAS, demonstrated the ability to identify enhancer from a single measurement obtained from low cell counts in ATAC-seq, conditions ideal for studying enhancers from clinical samples, which is proving to be of great importance\textsuperscript{101}. Using neural network models, individual specific enhancers were further refined and it was shown that a cell-agnostic model could effectively identify enhancers in a new cell type (EndoC-βH1 beta cell line) without prior knowledge of its enhancer annotations.

A major challenge addressed was the inconsistencies among different chromHMM annotations. As it is important to identify features capable of discriminating between different classes such as enhancers and other regulatory elements, it is equally important to have concise and consistent annotations. Correlating ChromHMM states and studying state calls made from consortiums such as ENCODE and Roadmap addressed this issue, however the determining the ‘ground truth’ is still a task requiring further improvement, especially when working with
reference annotations. Future studies using massively parallel reporter assays (MPRA) can be instrumental in revealing regions that behave differently in reference versus individual genomes.

Although only 3 classes were identified in this study, results in GM12878 revealed insulators to be strongly predicted. Two potential reasons for this strong prediction accuracy include depth of sequencing and better annotations for insulators. Further study into this regulatory element are ideal as insulators (often mediated by CTCF) are known to be in control of the 3D structure of chromatin, and may offer new insights for predicting the 3D structure in addition to identifying these regions.

The PEAS framework was designed to function in any cell type profiled with the ATAC-seq technology even in the absence of ChromHMM states. Source code of the PEAS framework was made available at https://github.com/UcarLab/PEAS.
Chapter 6

Conclusions and Discussion

In conclusion two data integration based tools were developed, and two studies furthering our knowledge of promoters, enhancers, their interactivity, and approaches for identifying them were conducted. QuIN\textsuperscript{21} was demonstrated as an easy to use web platform which facilitates the need for data integration with chromatin interaction data. Network analysis methods developed in QuIN were instrumental in inferring and interrogating the chromatin structure surrounding cell-type-specific promoters and enhancers. In addition, QuIN’s methods were useful for improving TriPOINT’s prioritization methods, allowing chromatin interactions to be used for further prioritization, utilizing information regarding the number of non-coding targets. Studying broad domains\textsuperscript{11} and super enhancers\textsuperscript{73}, led way to demonstrating the usefulness in using network analyses on chromatin interaction networks while leading to findings that suggest these cell-type-specific regulatory elements not only interact more frequently, but interact more frequently among each other. Finally, studying features from ATAC-seq revealed that these
features are capable of discriminating enhancers from other regulatory elements, moving one step forward in the direction of robustly identifying enhancers at the individual level.

Though these developments are diverse, they are very complementary to one another. For example, combining the results of chapter 4 and 5 suggests that given only chromatin interaction and ATAC-seq data, not only can promoters and enhancer be inferred, but cell-type-specific ones can be identified among them. A current limitation however is that chromatin interactions require high cell counts and are not currently feasible for individual or clinical samples. However, the results involving insulator predictions in GM12878 may prove useful in future studies for interrogating 3D chromatin structure of which platforms (QuIN) are now available to utilize machine learning approaches for inferring 3D chromatin structure from ATAC-seq alone.

Identification of individual specific enhancers also improves pathway-based analyses such as TriPOINT, where gene expression profiles of individuals can be fully interrogated based on their epigenome from putative individual specific enhancers with respect to known gene-to-gene interactions obtained from Pathways. Further information involving putative non-coding regulators can lead to better methods for scoring pathways and lead to therapeutic targets, potentially at the individual level, that regulate disease-causing factors that are specific to the tissue of which the disease is involved, thereby alleviating potential side effects in other unrelated tissues.
Appendix A

Details of Interaction Network Construction in QuIN

Chromatin interaction networks are constructed in three steps: 1) Node Extraction, 2) Edge Creation, and 3) Connected Component Discovery. Each method is described in detail below including both available methods for defining nodes within the network.

Node extraction procedures differ depending on the available data. Using only the interaction data, nodes are defined by merging interaction anchors. Anchors are first separated based on chromosome and sorted by start position. For each chromosome, the corresponding sorted anchor list is iterated once, maintaining a list of anchors to merge as well as the greatest end position seen. If the next anchor in the iteration has a start position less than that of the greatest end position seen then the anchor is added to the current list and the end position is updated accordingly if the new anchor’s end position is greater than the current greatest end position. If the next anchor’s start position is greater than the current greatest end position, then the current list of anchors defines a new node in the network and a new list is created to begin determining the next node in the network. Performing this procedure over all chromosomes defines all nodes represented in the network. An extend parameter can also be applied which will expand each anchor by the amount specified in both directions, offering flexibility for defining nodes. Regardless of the extend amount, nodes will be represented by the minimum start position and maximum end position of the anchors that define it. Algorithm 1 describes the above in more detail.
Algorithm 1 Node Creation with Interaction Anchors

```java
anchors[]; //array of interaction anchors
nodelist = new List();
chrgroups[] = separateByChromosome(anchors);

for i = 0 to chrgroups.length do
    sortByStartPosition(chrgroups[i]);
end for

for i = 0 to chrgroups.length do
    sortedanchors = chrgroups[i];
    anchorlist = new List();
    maxend = sortedanchors[0].endPosition;
    anchorlist.add(sortedanchors[0]);
    for j = 1 to sortedanchors.length do
        if sortedanchors[j].startPosition \leq maxend + extend * 2 then
            anchorlist.add(sortedanchors[j]);
        else
            nodelist.add(new Node(anchorlist));
            anchorlist = new List();
            anchorlist.add(sortedanchors[j]);
        end if
        maxend = max(maxend, sortedanchors[j].endPosition);
    end for
    nodelist.add(new Node(anchorlist));
end for
return nodelist;
```

With additional data provided for defining nodes, nodes are initially created using the node definitions (genomic regions) provided. If the regions within the data are found to be overlapping, then a step is performed to merge overlapping regions together into one region. The remainder of the algorithm focuses on determining the interaction anchors that overlap with the nodes in the network which is necessary for defining the edges of the network. For this, both nodes and anchors are separated based on chromosome where each list is sorted by start position. For each chromosome, the corresponding list of nodes and anchors are iterated concurrently as follows: 1) Select the first node in the sorted list and iterate over all anchors until the next
anchor's start position is greater than the node's end position. 2) Compare each anchor with the
current node as well as the next node in the list to determine whether or not the anchor overlaps
with either of these nodes after extending the nodes by user defined distance in both directions
for flexibility. 4) If an anchor overlaps with both nodes with extension, another comparison is
made without extension. If only one node overlaps with the anchor without extension, then the
anchor is assigned to the overlapping node. If both nodes are still overlapping with the anchor,
then the node that is overlapping with the anchor greater than half its length is assigned. 5) In the
case that neither node overlaps without extension, a final comparison is made with extension
again, checking whether the anchor overlaps with one node (with extension) greater than half the
length of the anchor while overlapping the other node less than half. If no assignment can be
made, then the interaction is categorized as ambiguous and will not be considered in the edge
creation step.

Edges are created by first initializing a tree-based map of node id keys, each referencing a
list of interactions. As interactions maintain a reference to their corresponding anchors and
anchors maintain a reference to the nodes they are assigned in the node creation step, this map is
created by iterating over each interaction once. For each iteration, the key for the map is
determined by concatenating the smallest integer node id with the largest node id (in that order)
using a delimiter and the interaction is added to the list referenced by the key in the map. If both
node ids are the same or one of anchors does not reference a node, then the interaction is not
included. Once the map is created, edges are created by iterating over the keys and values in the
map, using the key to determine the nodes to use for each edge. Finally, the edges are filtered
based on filtering parameters implemented in QuIN, removing them from the final network. The
procedure described above is detailed in Algorithm 2.
Algorithm 2 Edge Creation

\begin{verbatim}
interactions[];  //array of interactions in the network
edgemap;       //map of node id keys and interaction lists

for i = 0 to interactions.length do
    node1id = interactions[i].getAnchor1().getNode();
    node2id = interactions[i].getAnchor2().getNode();
    if node1id ≠ node2id then
        minid = min(node1id, node2id);
        maxid = max(node1id, node2id);
        key = minid + "," + maxid;
        if edgemap.containsKey(key) then
            edgemap.put(key, new List());
        end if
        edgemap.get(key).add(interaction[i]);
    end if
end for

for each key in edgemap do
    nodeids[] = split(K,",");
    createEdge(nodeids[0], nodeids[1], edgemap.get(K));
end for
\end{verbatim}

Connected components are determined in linear time by maintaining a Boolean array of visited nodes and performing Breadth-First Search on every node that has not yet been visited. The algorithm for this process simply iterates over the list of nodes where in each iteration, if the node has not been visited yet, a breadth-first search is performed putting all nodes and edges visited into the same connected component. Nodes visited when performing a breadth-first search are marked as visited such that breadth-first search is not repeated on the same component. After all nodes have been visited, all connected components in the network have been identified. Finally, single node components are removed from the network as they do not provide any interaction information and have proven to significantly increase the computational time for node annotation which is database query driven.
Appendix B

Methods for Interaction Calling in HiChIP and Hi-C

Interaction pairs from HiChIP\(^9\) (GSE80820) (targeting cohesion subunit \textit{Sme1a}) and Hi-C\(^82\) (GSE63525) data were called in the GM12878 cell line. Valid interaction pairs from HiChIP biological and technical replicates were pooled and filtered by extending 250bp in both directions, keeping only the pairs with both ends overlapping a DNASE-seq peak. Significant HiChIP interactions between peaks were called based on the hyper-geometric distribution as described in methods for ChIA-PET Tool\(^38\) and filtered using the Benjamini-Hochberg procedure (FDR < 0.05). Significant Hi-C intra-chromosomal interactions at 1kb resolution were identified implementing software based on the HiCCUPS method\(^82\) in order to specifically capture interactions at this higher resolution with less stringency. For each contact, expected values based on donut, vertical, horizontal and lower left filters were calculated using parameters P=20 and W=40 to calculate a P-value based on the Poisson distribution and filtered based on the Benjamini-Hochberg procedure (FDR < 0.025). A more stringent FDR cutoff is used for Hi-C data to make it more comparable with other assays, since Hi-C data were more deeply sequenced than the others. The intersection of these four filters was used to identify the final set of contacts.
Appendix C

ChIP-seq GSE Accession Numbers For ChromHMM

<table>
<thead>
<tr>
<th>ChIP/Cell Type</th>
<th>GM12878</th>
<th>CD4+ T</th>
<th>CD14+</th>
<th>PBMC</th>
<th>Islet</th>
</tr>
</thead>
<tbody>
<tr>
<td>H3k4me1</td>
<td>GSM733772</td>
<td>GSM772849,</td>
<td>GSM1102793</td>
<td>GSM1127143</td>
<td>PMID:24127591</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GSM997243</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H3k4me3</td>
<td>GSM733708</td>
<td>GSM772867,</td>
<td>GSM1102797</td>
<td>GSM1127126</td>
<td>GSM1261675</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GSM997240</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H3k27ac</td>
<td>GSM733771</td>
<td>GSM997239</td>
<td>GSM1102782</td>
<td>GSM1127145</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H3k36me3</td>
<td>GSM733679</td>
<td>GSM772866,</td>
<td>GSM1102788</td>
<td>GSM1127131</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>GSM997241</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H3k9me3</td>
<td>GSM733664</td>
<td>GSM772850,</td>
<td>GSM1102801</td>
<td>GSM1127133</td>
<td>PMID:24127591</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GSM997242</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H3k27me3</td>
<td>GSM733758</td>
<td>GSM772868,</td>
<td>GSM1102785</td>
<td>GSM1127130</td>
<td>PMID:24127591</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GSM997231</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CTCF</td>
<td>GSM733752</td>
<td>GSM325895</td>
<td>NA</td>
<td>NA</td>
<td>PMID:24127591</td>
</tr>
<tr>
<td>Control/Input</td>
<td>GSM733742</td>
<td>GSM772915,</td>
<td>GSM1102807</td>
<td>GSM1127151</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>GSM1112781</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Appendix C Table 1.** GSM Accession numbers and PubMed Ids of ChIP-Seq datasets used for generating ChromHMM states.
References


