Predicting cell-type specific combinatorial binding of neuronal transcription factor network by Deep Learning

and

Using Single-cell RNA seq to explore bone marrow immune landscape by IL-2 therapy.

Submitted by
Gunjan Dixit

August 2023

Supervisor
Assoc. Prof Jiayu Wen

A thesis submitted for the Degree of Doctor of Philosophy at
The Australian National University

Australian National University

Statement of originality

I declare that this dissertation is my own original work. It does not contain any material previously submitted for a degree to this or any other university or written by another person where due reference is not made in the text.

Gunjan Dixit

Table of Contents

List of Figures

List of Tables

Acknowledgements

Completing this chapter of my academic journey has been an incredible experience, and I am humbled by the support and kindness I have received from so many extraordinary people in my life. First and foremost, I want to express my sincere gratitude to my supervisor, **Assoc Prof. Jean Wen.** Her insightful guidance and unwavering encouragement have been the cornerstone of my PhD journey. Her belief in me and her constant motivation have been like a guiding light, helping me navigate through multiple projects and ultimately publishing papers. Without her remarkable supervision, this PhD journey wouldn't have been possible.

I'm also thankful to the Australian Government for awarding me the AGRTP scholarship, enabling my PhD journey in Australia. My sincere appreciation extends to the **Australian National University** and the **John Curtin School of Medical Research** for the opportunity to pursue my PhD in this prestigious institution. Their support, including the Covid-19 extension scholarship, has been invaluable during challenging times.

I'm truly grateful to my panel members, **Prof. Eduardo Eyras, Prof. Eric Stone,** and **Prof. Di Yu**, for their valuable guidance and insights that have shaped the trajectory of my PhD. My collaborators, **Prof. Qi Dai** (Stockholm University) and **Prof. Di Yu** (Queensland University) and **Dr. Brian Parker**, also deserve my heartfelt appreciation for providing valuable data and offering me the chance to work on significant projects. I'm thankful for the incredible people I've had the privilege of working with. Wen group—Ying Zheng, **Ke Ding (Stark), Lixinyu (Ema) Liu,** Jiajia Xu, Aobo Wang, Di Wang, and Xiangyu Chen—have been more than just colleagues; they've been friends and partners in this journey. A special shout-out goes to **Stark**, whose contribution has been nothing short of extraordinary. His exceptional guidance and support during the deep learning project have been invaluable.

My family's support has been my foundation. To my father, **Mr. K. K. Dixit**, who has been my role model, and my mother, **Mrs. Madhu Dixit**, whose teachings of determination and resilience have shaped my journey—I am immensely grateful for your unconditional love. To my siblings **Mrs. Jaya, Mrs. Pankhuri, Mr. Kushagra** and cherished family members, as well as my beloved aunt and guru, **Mrs. Anusuiya Dixit,**

your constant cheer and guidance have been a tremendous source of support.

I want to wholeheartedly express my gratitude to my fiancé, **Mr. Neel Thakkar**, for his steadfast support throughout my PhD journey. He possesses an exceptional ability to see the potential within me, even when I struggle to recognize it myself, and his continuous encouragement has been a driving force. His boundless love and caring nature have been an immense source of strength, and I feel incredibly fortunate to have him by my side and to have our paths crossed at ANU. His deep knowledge of Molecular Biology and science has sparked enlightening insights and discussions, which I wholeheartedly embrace and appreciate.

I am also incredibly grateful for my friends **Meetali, Pallavi, Alisha,** and **Sneha**, who have served as my cheerleaders, celebrating every milestone with sheer enthusiasm. Especially Meetali for being my constant and spreading joy from across the continent and Pallavi for sending me random desserts to celebrate my achievements or cheering me up when I was low. I also extend my gratitude to my fellow **"Toadies"** – Epi, Prima, Aishwarya, Ashwin, Ninad, Archica, and others – for their company and support that have enriched my journey in countless ways.

I'm also thankful to my previous mentors—**Prof. Andrew M. Lynn, Prof Bobby Paul**, and **Dr. Preeti Bajpai**—for shaping my passion for Bioinformatics. The Lynn Lab members, including **Dr. Amreesh Prakash, Dr. Madhulika Verma, Dr. Atanu Bannerjee, Dr. Poonam Vishwakarma, Dr Sunita Gupta, Mr. Mritunjay,** and **Mr. Nitin**, have also played an important role in my journey. Additionally, I am appreciative of my colleagues at JCSMR—**Anukriti, Yoshika, Akanksha, Jiaxin (Cindy), Zaka**, and **Emiliana**—for their motivation.

Finally, I wish to convey my heartfelt gratitude to the guiding force (**Baba)** that has steered me through this journey, endowing me with strength and tranquility. As I step forward, I eagerly anticipate new adventures and experiences.

Thank you.

Gunjan

List of publications

1. Regadas, I., Dahlberg, O., Vaid, R., Ho, O., Belikov, S., **Dixit, G.,** Deindl, S., Wen, J.* and Mannervik, M.*, 2021. A unique histone 3 lysine 14 chromatin signature underlies tissue-specific gene regulation. ***Molecular Cell***, 81(8), pp.1766-1780.
2. **Dixit, G.,** Zheng, Y., Parker, B. and Wen, J., 2021. Identification of Alternative Splicing and Polyadenylation in RNA-seq Data. ***JoVE (Journal of Visualized Experiments)***, (172), p.e62636.
3. Ding, K., **Dixit, G.**, Parker, B.J.* and Wen, J.*, 2023. CRMnet: a deep learning model for predicting gene expression from large regulatory sequence datasets. ***Frontiers in Big Data***, 6.

   * co-corresponding

List of manuscripts in preparation:

1. scTFNet: Predicting cell-type specific combinatorial binding of neuronal transcription factor network by deep learning

2. Using Single-cell RNA seq to explore bone marrow immune landscape by IL-2 therapy

List of Presentations

1. Poster presentation and a flash talk at the Australian Mathematical Sciences Institute (AMSI) BioInfoSummer 2019, University of Sydney (2019).
2. Short presentation at the Australian Bioinformatics and Computational Biology society (ABACBS) conference, ANU, Canberra (2020).
3. Poster presentation and a 3-minute flash talk at the Australian Genomics and Technologies Association Conference, Sunshine Coast, Queensland (2022).

List of Abbreviations

AEL	After Egg Laying
ANN	Artificial Neural Network
AP	Anterior-Posterior
APA	Alternative Polyadenylation
AS	Alternative Splicing
ATAC	Assay for Transposase-Accessible Chromatin
AUC	Area Under the Curve
AUPRC	Area Under the Precision-Recall Curve
AUROC	Area Under the Receiver Operating Characteristic Curve
BDGP	Berkeley Drosophila Genome Project
BEND5	Beta-like Endonuclease-Domain 5
BM	Bone marrow
C2H2	Cysteine2Histidine2 Zinc Finger
CBF1	Core-Binding Factor Subunit 1
CDS	Coding Sequence
CFRT40A	Cystic Fibrosis Transmembrane Conductance Regulator
CNN	Convolutional Neural Network
CNS	Central Nervous System
CPM	Counts Per Million
CRE	Cis-Regulatory Element
CRM	Cis-Regulatory Module
DEG	Differentially Expressed Gene
DGE	Differential Gene Expression
DL	Deep Learning
DNN	Deep Neural Network
DV	Dorsal-Ventral
ENCODE	Encyclopedia of DNA Elements

F1	F-Score (Harmonic Mean of Precision and Recall)
FDR	False Discovery Rate
FER1L5	Fer-1-Like Protein 5
FIMO	Find Individual Motif Occurrences
FN	False Negative
FP	False Positive
GEO	Gene Expression Omnibus
GMC	Glial Microtubule-Associated Cell
GO	Gene Ontology
GPU	Graphics Processing Unit
GRO	Global Run-On Sequencing
GSEA	Gene Set Enrichment Analysis
H3K14R	Histone H3 Lysine 14 to Arginine Mutation
HOMER	Hypergeometric Optimization of Motif EnRichment
HPC	High-Performance Computing
IL2	Interleukin-2
ILC	Innate Lymphoid Cell
ILC2	Type 2 Innate Lymphoid Cell
K14R	Lysine 14 to Arginine Mutation
KO	Knockout
MACS2	Model-Based Analysis of ChIP-Seq 2
MCC	Matthews Correlation Coefficient
MEME	Multiple Em for Motif Elicitation
ML	Machine Learning
MLP	Multi-Layer Perceptron
MNN	Mutual Nearest Neighbors
MXE	Mutually Exclusive Exon
NGS	Next-Generation Sequencing
NK	Natural Killer
OVX	Ovariectomy
PNS	Peripheral Nervous System

PRC	Precision Recall Curve
PRO	Precision-Recall Operating Characteristic
PSI	Percent Spliced-In
PSSM	Position-Specific Scoring Matrix
PWM	Position Weight Matrix
RBP	RNA-Binding Protein
RNN	Recurrent Neural Network
RPGC	Reads Per Genomic Content
RPKM	Reads Per Kilobase of Transcript per Million Mapped Reads
SC	Skip Connection
SE	Squeeze Excitation
SRA	Sequence Read Archive
STAR	Spliced Transcripts Alignment to a Reference
SVM	Support Vector Machine
TF	Transcription Factor
TFBS	Transcription Factor Binding Site
TFN	Transcription Factor Network
TMM	Trimmed Mean of M-values
TPR	True Positive Rate
TPU	Tensor Processing Unit
TSNE	t-Distributed Stochastic Neighbor Embedding
TSS	Transcription Start Site
UCSC	University of California, Santa Cruz
UMAP	Uniform Manifold Approximation and Projection
UMI	Unique Molecular Identifier
UTR	Untranslated Region
WT	Wild Type

Abstract

Project 1: scTFNet: Predicting cell-type specific combinatorial binding of neuronal transcription factor network by deep learning

Gene regulation is a multifaceted, with transcriptional regulation being pivotal. Transcription factors (TFs) are DNA binding proteins that recognize and bind to specific DNA nucleotide sequence or motifs, which recruit other proteins such as cofactors and RNA Polymerase II to initiate transcription. TFs play vital roles in regulating the expression of target genes and are involved in cellular processes such as development and differentiation. Notably, in neurons, cell-specific TFs play a crucial role in the differentiation process by guiding neuronal specification, migration and circuit assembly through transcriptional cascades. Mapping of transcription factor binding proteins at the genome and single-cell level allows for the identification of various cellular sub-types and their distinct roles in cell-type specific gene expression. Although recent deep learning methods have made progress in predicting TF binding sites, insights into TF combinatorial binding and cell-specific TF binding at single-nucleotide resolution remains elusive. To unravel the underlying mechanism that mediates combinatorial gene regulation and cell fate specification during development, we leveraged neuronal-specific TFs in *Drosophila* early embryo data to predict TF combinatorial binding profiles and associated gene expression patterns.

In this project, we developed scTFNet, a novel Transformer encoded U-Net model capable for predicting combinatorial binding sites from DNA sequences and chromatin accessibility. Our model accurately identified the binding of 14 *Drosophila* neuronal-specific TFs and their co-bindings, including Su(H), Scute (Sc), Charlatan (Chn), Asense (Ase), Senseless (Sens), Insensitive (Insv), Early boundary activity (Elba), Scratch (Scrt), Shaven (Sv/Dpax2), 48 Related 1 (Fer1), Hamlet (Ham), Tramtrak Ttk69), Hairless (H) and Sequoia (Seq), at single nucleotide resolution with an accuracy of 0.99. Our deep learning model employed two approaches, a classification model to predict TF binding sites and a regression model to predict gene expression. When applied to single-cell ATAC-seq data, our model adeptly predicts cell-specific binding of neuronal TFs at the

single cell level directly from DNA sequences. Using saliency maps, we interpreted our model's predictions by revealing the TF motifs at the predicted binding regions, providing valuable insights into the mechanisms underlying the predicted binding patterns. Additionally, we developed a regression model to infer gene expression from DNA sequences and validated these predictions with corresponding Precision nuclear run-on sequencing (PRO-Seq). In summary, this study presents a novel and robust method to determine cell-type-specific binding patterns and gene expression in *Drosophila* early embryo. We demonstrate that our model can predict the TF motif of various TFs in different cell types using DNA sequences, shedding light on the principles of combinatorial gene regulation and cell lineage by analysing TF binding and chromatin accessibility in conjunction.

Project 2: Using Single-cell RNA seq to explore bone marrow immune landscape by IL-2 therapy

Bone marrow (BM) contains multiple immune cell subsets with critical functions and is considered an immune regulatory organ. It contains osteoclasts and immune cells fundamentally involved in physiological and pathological bone remodelling. In autoimmune diseases, inflammation can impair the BM niche, disturb hematopoietic and immune development, and induce osteoporosis. Interleukin-2 (IL-2) is one of the crucial cytokines that exhibit pleiotropic effects on the immune system which is chiefly produced by CD4+ T cells, and to a lesser extent, by CD8+ T and Natural Killer cells. It is a 15 kDa protein containing 4-bundles of α-helices and was initially discovered as a growth factor for bone-marrow-derived T lymphocytes. IL-2 plays a significant role in the maintenance of CD4+ regulatory T cells, along with the differentiation of CD4+ T cells into different subsets. The strength of the IL-2 signal affects the differentiation process with strong signals leading to short-lived effector T cells and low-level signals leading to follicular helper (Tfh) or central memory T-cells. Discovery of IL-2 in the regulation of survival, differentiation and propagation of activated T cells paved the path for its direct clinical implications in immunotherapy. Since IL-2 is capable of activating both Treg and cytotoxic lymphocytes, using low-dose IL-2 administration increases the number of only Treg cells which is beneficial for conditions like autoimmunity and chronic inflammatory diseases,

whereas high-dose IL-2 treatment expands the population of cytotoxic lymphocytes, useful for metastatic cancer.

This project was designed to address the fundamental question of how low-dose IL-2 therapy modulates BM immune landscape by comprehensively mapping the therapy-induced changes using single-cell technologies and then dissecting the underlying mechanisms by mouse models. I analysed the scRNA-seq data obtained from BM of CD45+ cells of mice to identify different immune cell types and compared their expression across four experimental conditions- a control (sham), control treated with IL-2 (Sham + Treatment), ovariectomy-induced osteoporosis (OVX) and OVX treated with IL-2 (OVX + Treatment). This study uncovers significant cellular heterogeneity within mouse BM and identifies the rare ILC2 cell type. The observed increase in ILC2 cells in IL-2 treated samples suggests a functional role in osteoclast suppression, which is crucial for preventing bone resorption in osteoporosis. Gene expression profiling reveals a substantial decrease in differentially expressed genes in IL-2 treated samples, indicating a regulatory effect of the treatment. Pathway analysis demonstrates that the osteoclast differentiation pathway is downregulated in both treated and control groups, while upregulated in the diseased group. Trajectory analysis confirms the cluster identification and outlines the developmental direction of B and T cell populations and subtypes. Collectively, these findings underscore the therapeutic potential of IL-2 in modulating bone-related diseases and pave the way for novel treatment strategies.

PROJECT 1

scTFNet: Predicting cell-type specific combinatorial binding of neuronal transcription factor network by deep learning

Chapter 1: INTRODUCTION

1.1. Biology of Gene Regulation Driven by Transcription Factors

1.1.1 Eukaryotic gene expression and its regulation

Gene expression refers to the process by which the information encoded in DNA is transcribed into RNA and further translated into a functional protein. In multicellular organisms, cells exhibit specialization and perform distinct functions based on their appearance and structure in different tissues due to the expression of set of genes unique to each cell type. Nearly all the cells of an organism contain identical genome, however, the different set of genes expressed at a given time (differential gene expression) gives rise to various cell-types.

Gene regulation is a crucial biological process that controls how genes are expressed within an organism. Precise regulation of genes is an essential step for normal development and functioning of cells and the entire organism. Gene expression can be regulated at both transcriptional and post-transcriptional levels, the first level is transcriptional regulation by sequence-specific transcription factors and their interaction with *cis*-regulatory modules (CRMs), the second level includes modulation by RNA binding proteins (RBPs) and microRNAs, and third level involves chromatin modifications as a result of transcription factor-DNA interactions (Peter & Davidson, 2015). The process of controlling gene expression is intricate, and dysfunctions in this process can be detrimental to the cell, leading to the development of various diseases, including cancer, diabetes and more.

1.1.2 Transcription Factors and their role in gene expression

Transcriptional control is often mentioned in regard to “cis” and “trans” acting components. The term “cis” pertains to the regulatory elements present on the same DNA as the target gene including enhancers and promoters, while “trans” refers to regulatory

proteins that are transported from elsewhere to interact with the DNA, such as transcription factors.

Transcription Factors (TFs) are DNA binding proteins that recognize and bind to specific DNA nucleotide sequence (motif) to recruit cofactors and RNA Polymerase II for transcription initiation. TFs bind to the DNA in a sequence specific pattern by virtue of DNA binding domains, which are known to be conserved across different families of animal kingdom. TFs regulate the expression of specific target genes and play a vital role in development and differentiation (Lambert et al., 2018).

1.1.3 Transcription Factor Binding sites and Motifs

Transcription factor binding sites (TFBS) refers to short stretches of DNA ranging in length from 6 to 20 nucleotides, located in the promoter regions or enhancer elements of genes. They serve as docking sites for transcription factors and are critical for the precise regulation of gene expression (Jolma et al., 2013).

TF binding to a TFBS can lead to either activation or repression of gene expression, depending on the specific TF and the context of binding. Typically, an activator TF binding to a TFBS results in transcriptional activation by facilitating the recruitment of RNA polymerase and other transcriptional machinery to the promoter region (**Figure 1.1**). This process enhances the transcription of the target gene. Conversely, when a repressor TF binds to a TFBS, it impedes the binding of RNA polymerase to the promoter region and causes transcriptional repression, resulting in decreased expression of the target gene (Farnham, 2009; Stormo, 2000).

TFBS conservation across species indicates their functional significance, which can be identified using experimental techniques such as electrophoretic mobility shift assays (EMSAs) or chromatin immunoprecipitation (ChIP), as well as computational methods like motif scanning and discovery. The availability of whole-genome sequence data and bioinformatics tools has further enhanced TFBS identification.

Characterizing TFBS has revealed vital information about gene regulation mechanisms and facilitated the development of gene expression manipulation tools. Targeting specific TFBS allows for precise gene expression control, which has significant implications in gene therapy and synthetic biology. Furthermore, identifying TFBS has helped decipher complex regulatory networks governing gene expression, enhancing our understanding of normal development and disease.

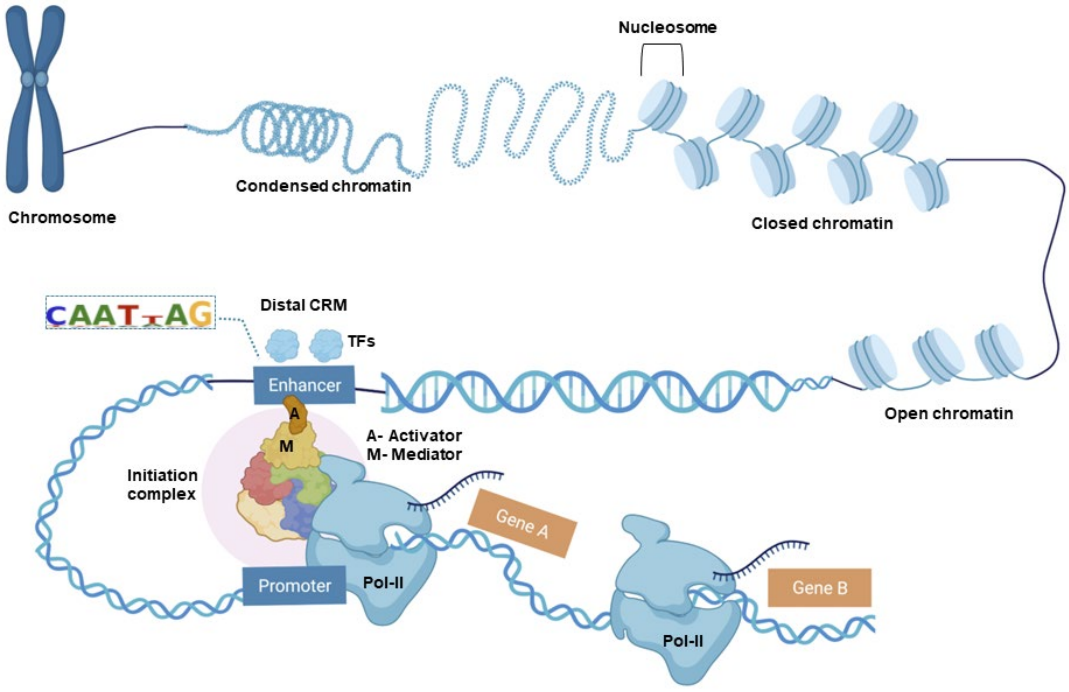


Figure 1.1: Gene expression and its regulation. *Genomic DNA, packaged into the nucleus with histone proteins called nucleosomes, undergoes the process of gene expression. This process involves the conversion of DNA-encoded information into RNA, followed by the production of functional proteins. Gene expression is regulated at different levels in the genome, with transcriptional regulators playing a crucial role in conferring tissue and cell-specific functions. Transcription Factors (TFs) initiate gene transcription by binding to specific DNA nucleotide sequences (motifs), recruiting RNA Polymerase II and cofactors. They regulate the expression of specific target genes, playing a vital role in development and differentiation. Activator proteins bind to enhancer sequences, bending the DNA and bringing it closer to the promoter site. Other TF proteins and cofactors join the activator to form an initiator complex that binds to the gene promoter and recruits RNA Polymerase II, initiating transcription.*

1.1.4 Combinatorial Transcription Factor Binding

Combinatorial transcription factor binding refers to the phenomenon where multiple TFs bind to DNA in a coordinated and synergistic manner to regulate gene expression. This mechanism allows for a high degree of regulatory specificity and diversity, as different combinations of TFs can generate different gene expression patterns.

Combinatorial TF binding can occur in several ways. One way is through the cooperative binding of multiple TFs to a single TFBS. In this case, the binding of one TF can increase the affinity of another TF for the DNA sequence, leading to a synergistic effect on gene expression. Another way is through the binding of TFs to adjacent TFBSs, which can create composite elements that have unique regulatory properties that cannot be achieved by either TF alone.

1.1.5 Cell-type specificity of Transcription factor binding sites

The two primary factors that contribute to cell-type specificity are the differential expression of TFs and the chromatin state of the DNA. For example, a transcription factor that is expressed at high levels in liver cells may exhibit no expression in heart cells, leading to distinctive profiles of TF binding sites in these two cell types (Gaulton et al., 2010). Furthermore, the chromatin state can differ among various cell types, this can impact the availability of TF binding sites. As an instance, a TF binding site that is readily accessible in liver cells may be obstructed in heart cells owing to the disparities in chromatin structure.

Several studies have demonstrated that the binding of TFs to DNA is highly dependent on the cell type (Cusanovich et al., 2014). Research also suggests that cell-specific transcription factors play a key role in the differentiation process of neurons by regulating transcriptional cascades and guiding neuronal specification, migration and circuit assembly (Greig et al., 2013).

1.1.6 Cis-regulatory modules

Cis-regulatory modules (CRMs) are functional units of DNA composed of multiple cis-regulatory elements (CRE) that act together to regulate gene expression. They can be either promoters, insulators or enhancers and depending on the nature of the transcription factor associated with it, CRMs can activate or repress gene expression. Combinatorial binding refers to the phenomenon where multiple TFs bind to a CRE in a specific sequence and spatial configuration, allowing for precise regulation of gene expression. Combinatorial control of CRMs explains the complexity of gene expression patterns during animal development.

Recent research has demonstrated the importance of CRMs in a variety of biological processes, including development, homeostasis, and disease. Additionally, alterations in CRM activity have been shown to play a critical role in development and tissue differentiation (Boyer et al., 2005).

Cell-type specific gene expression is a complex process that involves the interaction of regulatory elements with specific activator proteins to initiate transcription of a gene. For instance, **Figure 1.2** illustrates two neuronal cell types, A and B, and their corresponding gene structures. In cell type A, the activator proteins for gene A recognize the enhancer element, thereby triggering the transcription of gene A while blocking the expression of gene B. In contrast, in cell type B, the activator proteins for gene B bind to the enhancer element and initiate the transcription of gene B while suppressing the expression of gene A. This mechanism ensures that only the appropriate genes are expressed in a specific cell type, thereby allowing for the differentiation and specialization of cells in a tissue or organism.

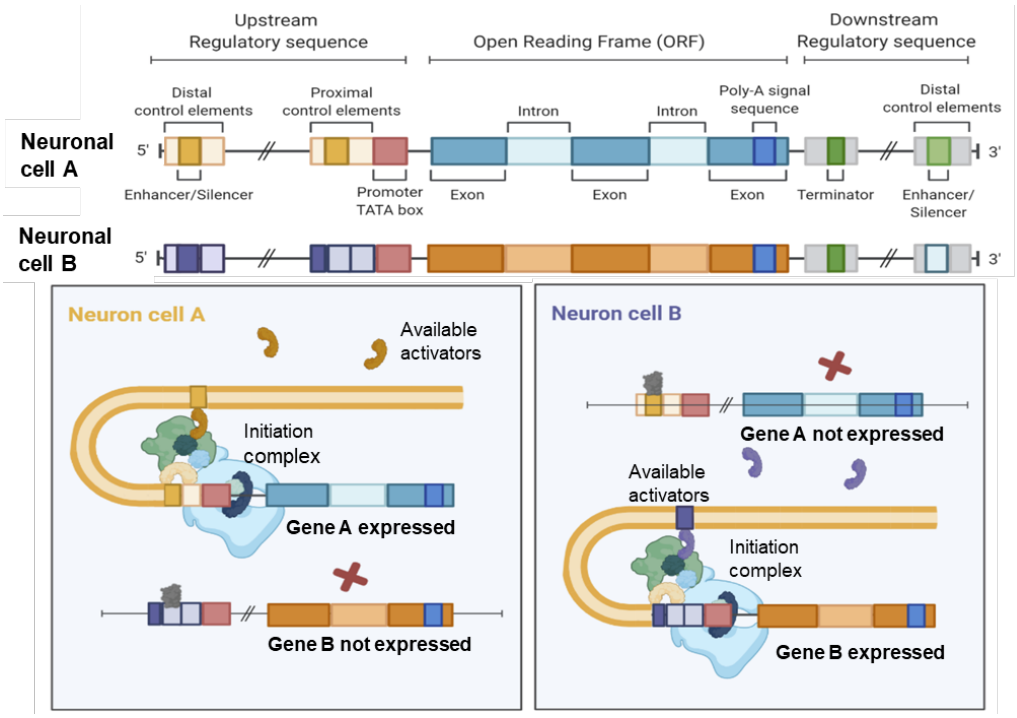


Figure 1.2: Cell-type specific gene expression*. Here is an example of cell-type specific gene expression. Given are two neuronal cell types A and B and their gene structure depicting presence of distal and proximal control elements, open reading frame and downstream regulatory element. Within a cell, specific activator proteins are available which bind to control elements and results in cell type-specific expression. In cell type A, activators of gene A recognize the enhancer and initiates transcription of gene A, whereas expression for gene B is blocked/turned off. and vice-versa. (Image adapted from Biorender illustration by Mirko Scrivano)*

1.1.7 Chromatin accessibility and Nucleosome positioning

Genomic DNA is a long stretch of molecule packaged into nucleus with histone proteins called nucleosome. Chromatin refers to the DNA-histone complex that forms the structural basis of the chromosome. Chromatin accessibility is the ability of transcription factors and other regulatory proteins to access specific regions of DNA within the chromatin structure, while nucleosome positioning refers to the placement of histone proteins on DNA. These two factors are intimately linked and together influence gene expression patterns. Regions of DNA that are accessible to regulatory proteins are often characterized by a depletion of nucleosomes, while inaccessible regions are characterized by tightly packed nucleosomes. The capacity of TFs to bind DNA depends

on their interaction with nucleosome which can inhibit the binding by obstructing the motif, hence TFs can bind only to the nucleosome-depleted (nucleosome free) regions (Zhu et al., 2018). The specific positioning of nucleosomes on DNA also affects accessibility to regulatory proteins, as nucleosomes can either block or facilitate access to regulatory elements (Jiang & Pugh, 2009).

This provides second layer of regulation by profiling chromatin accessibility as the chromatin needs to be in an accessible state to allow for the binding of transcriptional machinery. Chromatin structure is regulated by ATP-dependent remodelling complexes that can mobilize nucleosome to facilitate access of the transcription apparatus and its regulators to DNA.

1.1.8 Transcriptional control in Development

Transcriptional control is a complex and dynamic process that plays a critical role in the development of multicellular organisms. It involves the regulation of gene expression, which is essential for cell fate determination and differentiation into specialized cell types. Mapping of chromatin marks and transcription factor binding proteins at the genome level has the potential of identifying various subtypes of regulatory elements. While the expression of many genes is affected successively by all these mechanisms, the control of developmental complexity by means of spatially differential gene expression is executed by transcription factors and their sequence-specific interactions with the regulatory genome. Therefore, the primary key to the informational process of development lies in the control system regulating the expression of genes encoding transcription factors (Peter & Davidson, 2015).

Recent advances in genome-wide technologies, such as chromatin immunoprecipitation sequencing (ChIP-seq), assay for transposase-accessible chromatin using sequencing (ATAC-seq) and RNA sequencing (RNA-seq), have enabled researchers to identify the genomic locations of transcription factor binding sites, chromatin accessibility and the expression patterns of genes across various developmental stages and cell types (Buenrostro et al., 2015). These studies have

provided new insights into the regulatory networks that control gene expression during development and how these networks are disrupted in disease.

1.2 Regulatory Principles of Neural cell subtype specification

1.2.1 *Drosophila* Neurogenesis

Drosophila melanogaster, commonly known as the fruit fly, has been used as a model organism in the study of genetics and developmental biology for over a century. Its small size, short life cycle, and easy maintenance make it a convenient organism to work with, and its genetic similarity to humans has led to important discoveries in a variety of fields (Bier, 2005; Rubin et al., 2000). One area of research that has benefited greatly from the use of *Drosophila* as a model organism is the study of neurogenesis, or the process by which neurons are generated and differentiated. *Drosophila* neurogenesis begins with the subdivision of lateral neurogenic ectoderm into columnar domains along the dorsoventral axis as shown in **Figure 1.3** (Cowden & Levine, 2003; Von Ohlen & Doe, 2000), followed by the formation of proneural clusters by delamination. Neuroblasts then undergo a series of divisions to generate mother ganglion cells which further gives rise to neurons and glia (Broadus et al., 1995). The early embryonic ectoderm undergoes patterning and is subdivided into the anterior-posterior (AP) and dorsoventral (DV) axes. The AP axis defines segments and provides positional significance within each segment whereas the DV axis is subdivided by columnar genes (vnd, ind and msh) giving rise to Central Nervous System (CNS). Both spatial and temporal mechanisms are responsible for distinct sets of neuroblasts. Spatial patterning indicates neuronal diversity generated by precursors ensuing asymmetric cell divisions giving rise to ordered sequences of cells based on their identity.

Neuroblast division is tightly regulated by several signalling pathways and transcription factors. For example, the Notch signalling pathway plays a critical role in regulating neuroblast self-renewal and differentiation by controlling the asymmetric distribution of cell fate determinants during mitosis (Knoblich, 2008). The proneural genes, which encode basic helix-loop-helix transcription factors, are also essential for neuroblast specification and differentiation (Powell & Jarman, 2008). During the course of *Drosophila* neurogenesis, a number of important cellular processes occur that contribute to the formation of functional neural circuits. These include axon guidance,

synapse formation, and myelination. For example, the growth cone guidance receptor Dscam plays a crucial role in ensuring proper targeting and connectivity of axons in the developing *Drosophila* brain (Hattori et al., 2008). Similarly, the glial-specific transcription factor glial cells missing (gcm) is required for proper differentiation and myelination of glial cells in the peripheral nervous system (Lin & Goodman, 1994).

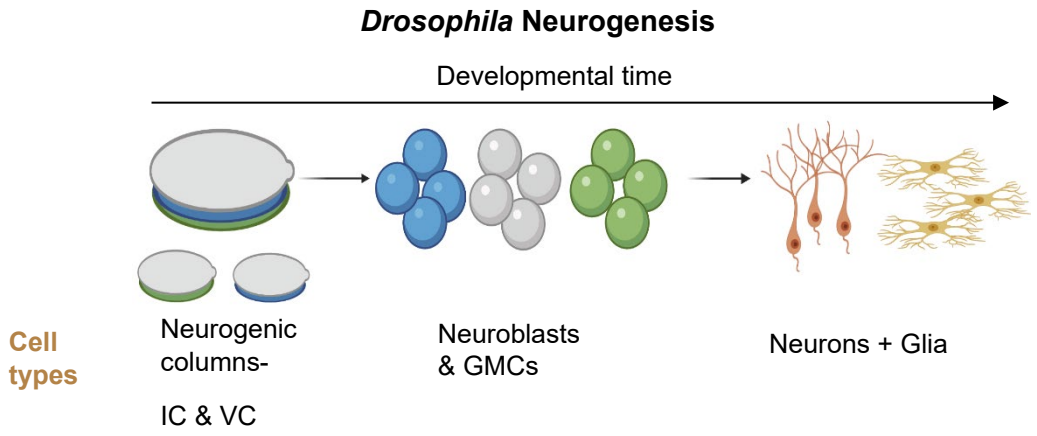


***Figure 1.3: Drosophila Neurogenesis**. The process of neurogenesis begins with the division of lateral neurogenic ectoderm into columnar domains Intermediate Columnar domain (IC) and Ventral Columnar domain (VC) along the dorsoventral axis. These columns contain clusters of cells that express specific combinations of transcription factors and signalling molecules, which are critical for the development of specific neuronal cell types. The formation of these columnar domains is a crucial step in the process of neurogenesis as it sets the stage for the subsequent formation of proneural clusters from Ganglion Mother cells (GMCs) and the generation of neurons and glia.*

1.2.2 Embryonic stages of *Drosophila* and their significance in neural development

Embryogenesis in *Drosophila* is an intricate process that can be divided into fourteen distinct stages, each of which is characterized by specific morphological and molecular events. These stages occur over a period of approximately 24 hours at 25°C and are tightly regulated by a complex network of genetic and environmental factors. The initial stages of embryogenesis are characterized by rapid nuclear division without cell division, resulting in the formation of a multinucleate syncytium. This syncytium then undergoes cellularization, a process in which the syncytium is divided into individual cells.

As development proceeds, the cells of the embryo begin to differentiate, giving rise to distinct body segments and tissues. Finally, the fully formed larva emerges from the eggshell and begins its life as a free-living organism. The precise timing and regulation of each of these developmental stages is critical for the proper formation of the adult fly and provides a powerful system for studying the genetic and molecular mechanisms underlying embryonic development.

Figure 1.4 illustrates overview of the stages of development that begins with cleavage, which is a rapid process of cell division that starts immediately after fertilization. During the first four stages of embryogenesis, the nucleus of the fertilized egg undergoes 13 rapid divisions, resulting in the formation of a large number of nuclei that become arranged in a single layer beneath the egg surface (Nüsslein-Volhard & Wieschaus, 1980). Cell membranes are then formed around these nuclei, leading to the formation of the cellular blastoderm, a homogeneous cellular sheet surrounding the central yolk (St Johnston, 2002).

During gastrulation, which occurs during stages 6-8 of embryogenesis (2-4 hours), cells within the polar caps and the mid-ventral part of the blastoderm undergo invagination, resulting in the formation of the three germ layers: endoderm, mesoderm, and ectoderm (Keller, 2002). Most of the cells that remain at the surface of the blastoderm represent the ectoderm, while the invaginating cells form the endoderm (anterior and posterior midgut rudiments) and the mesoderm. A narrow mid-dorsal partition of the blastoderm gives rise to the amnioserosa, a thin membrane that covers the germ band dorsally (Campos-Ortega et al., 1997). During stages 9-11 (4-6 hours, 6-8 hours), the germ band continues to elongate, and the ectoderm begins to split up into numerous different organ primordia, including the foregut and hindgut, CNS, and epidermis (Campos-Ortega et al., 1997). The closure of the dorsal gap between the two lateral epidermal sheets, also known as dorsal closure, occurs in stages 12-13 (8-10 hours). Following the completion of dorsal closure, the cells of the embryo continue to differentiate (18-22 hours), giving rise to distinct tissues and body segments **(Figure 1.4)**.

The precursors of the central nervous system (CNS) in *Drosophilaer* are derived from specialized ectodermal regions, the neurogenic regions, which give rise to neuroblasts, the primary precursor cells of the CNS. The ventral neurogenic region (VNE) generates the neuroblasts of the ventral nerve cord, while the procephalic neurogenic region (PNE) produces the brain and adjacent optic lobe anlage (Hartenstein & Campos-Ortega, 1984). The neurogenic regions also contain epidermal precursors that give rise to the ventral and head epidermis. Neuroblast delamination from the ectoderm occurs in three waves in the ventral neurogenic region, starting at stage 9, and the full complement of neuroblasts is segregated by stage 11 (Crews, 2019; Hartenstein & Campos-Ortega, 1984).

Overview of the stages of development

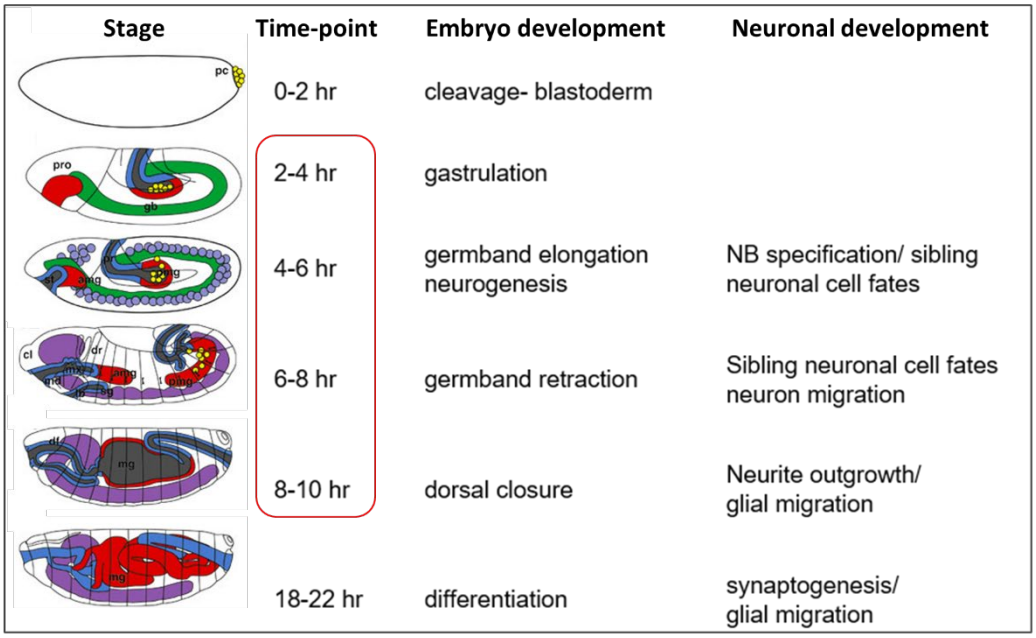


Figure 1.4: Stages of embryonic development in Drosophila*. Embryonic development is represented by several distinct stages starting from cleavage to Blastoderm, followed by gastrulation, germband elongation and differentiation. Each distinct stage is characterized by 2-hour window specifying time-points ranging between 0-22 hours after egg laying (AEL) and plays a critical role in the development of the nervous system. During early stages, neuroblast specification occurs at 4-6 hr AEL, followed by the development of sibling neuronal cell fates and neuron migration at 6-8 hr AEL. Later stages such as Neurite outgrowth and glial formation occur at 8-10 hr AEL. The time-points highlighted in red were the focus of this study. (Adapted from Hartenstein, 1993)*

1.2.3 Signalling Pathways Involved in Neural Cell Subtype Specification

In *Drosophila*, neural cell subtype specification is regulated by a complex interplay of signalling pathways. The Notch, Wnt, Hedgehog, and BMP signalling pathways are known to be involved in specifying various neural cell types. These signalling mechanisms and their roles in differentiation, proliferation, and neuronal development are widely studied, noting their conservation from flies to mammals, which highlights the deep evolutionary roots of these pathways.

The Notch signalling pathway is highly conserved from invertebrates to vertebrates including humans and regulates cell fate decisions in many tissues, including the nervous system. Notch signalling is involved in specifying various neural cell types, such as glial and neuronal subtypes. The pathway works through the interaction between transmembrane Notch receptors and their ligands, which leads to the activation of downstream transcription factors. These include the Suppressor of Hairless (Su(H))/C-promoter binding factor 1 (CBF1) and the Enhancer of split (E(spl)) family of basic helix-loop-helix (bHLH) transcription factors (Artavanis-Tsakonas et al., 1999).

The Wnt signaling pathway, involved in several aspects of neuronal development is highly conserved from early metazoans to complex organisms including humans. It is responsible for regulating cell proliferation, fate determination, and differentiation. During nervous system development, Wnt signalling plays a role in specifying various neural cell types, such as dopaminergic neurons and oligodendrocytes. The pathway involves the interaction between Wnt ligands and Frizzled receptors, leading to the activation of downstream transcription factors such as the T-cell factor/lymphoid enhancer factor (TCF/LEF) family of transcription factors (Chenn & Walsh, 2002).

Hedgehog signalling is another conserved pathway across invertebrates and vertebrates that plays a crucial role in regulating cell proliferation and differentiation during development. In the nervous system, Hedgehog signalling has been identified as an essential player in the specification of neuronal and glial cell types, axon guidance, and synapse formation. The pathway operates through the interaction between Hedgehog

ligands and Patched receptors, leading to the activation of the Gli family of transcription factors (Gli1, Gli2, and Gli3) (Dahmane & Ruiz Altaba, 1999).

BMP (bone morphogenetic protein) signalling is another important pathway conserved from flies to human that plays a crucial role in neural cell fate determination during development. The BMP pathway is involved in various aspects of neural development, including the specification of distinct neural cell types. This includes the specification of astrocytes and oligodendrocytes. BMP signalling is initiated by the interaction between BMP ligands and BMP receptors, leading to the activation of downstream transcription factors such as the Smad family of transcription factors (Mabie et al., 1999). Overall, these signalling pathways play crucial roles in neural cell subtype specification during *Drosophila* nervous system development. The interplay between these pathways and their downstream effectors determines the fate of neural progenitor cells and ultimately contributes to the formation of a functional nervous system.

1.2.4 Neural specific TFs / TF selectors in determination of unique neural subtypes

The central nervous system of *Drosophila* is an ideal resource for investigating the temporal specification of cell fate during neurogenesis due to its abundance of relevant transcription factors and the availability of numerous techniques for manipulating and analysing individual progenitor cells and their progeny. This has been demonstrated by various studies (Cleary & Doe, 2006; Grosskortenhaus et al., 2005; Isshiki et al., 2001; Kanai et al., 2005; Maurange et al., 2008; Novotny et al., 2002; Pearson & Doe, 2003, 2004).

During *Drosophila* neurogenesis, the specification of unique neural subtypes is tightly regulated by the activity of neural-specific transcription factors (TFs) and TF selectors. These TFs and selectors act as key regulators of gene expression, controlling the activation or repression of specific genes in different neural cell types. For example, the TFs Even-skipped (Eve), Hunchback (Hb), and Krüppel (Kr) are expressed in specific subsets of neuroblasts, and their activity is required for the generation of distinct neural subtypes in the developing *Drosophila* embryo (Doe et al., 1991).

- ***Ase (Achaete-Scute complex)***- The Achaete-Scute complex is a group of genes that encode bHLH transcription factors. Ase is one of the four genes in this complex and is responsible for the development of sensory organs in *Drosophila* (Skeath and Carroll, 1991). Ase is expressed in the progenitor cells that give rise to sensory organ precursor (SOP) cells, which then differentiate into various sensory cell types (Brand et al., 1993).
- ***Sens (Senseless)-*** Sens is a transcription factor that is involved in the development of sensory organs in *Drosophila*. It is expressed in the SOP cells that give rise to sensory cells and is required for the differentiation of these cells into various sensory cell types (Nolo et al., 2000).
- ***Insv (Insensitive)-*** It is involved in the regulation of Notch signalling in *Drosophila*. It is expressed in the cells that give rise to the peripheral nervous system and is required for the differentiation of these cells into various neuronal subtypes (Golowasch et al., 2009).
- ***H (Hairless)-*** It is involved in the regulation of Notch signalling in *Drosophila*. It acts as a negative regulator of the pathway and is required for the proper development of various tissues, including the wing and eye (Barolo et al., 2000).
- ***Sc (Scute)-*** It plays a role in sensory organ development in *Drosophila*. It is expressed in sensory organ precursor cells and is required for the proper differentiation of sensory organs, including mechanosensory bristles and chemosensory organs (Cubas et al., 1991). Sc regulates the expression of other transcription factors involved in sensory organ development, including Atonal and Senseless (Brand et al., 1993).
- ***Ttk (Tramtrack)-*** It is a transcription factor that plays a role in *Drosophila* development, including the development of the nervous system, gut, and genitalia (Melnick et al., 1993). It is also involved in the regulation of cell proliferation and apoptosis (Klinedinst & Bodmer, 2003). Ttk has been shown to interact with other transcription factors, including grainyhead and zinc-finger homeodomain proteins, to regulate gene expression during development (Klinedinst & Bodmer, 2003; Melnick et al., 1993).

- ***Seq (Sequoia)-*** It plays a role in cardiac development in *Drosophila*. It is expressed in the developing heart and is required for the proper differentiation of cardiac cells. Seq regulates the expression of several genes involved in cardiac development, including tinman and pannier (Qian et al., 2007).
- ***Ham (Hamlet)-*** This factor plays a crucial role in determining the fate of neurons in the peripheral nervous system and olfactory receptor neurons. Additionally, ham is involved in the maturation of intermediate precursor cells in the type II neuroblast lineages of the larval brain.
- ***Fer1 (48 related 1)***- It encodes a basic helix-loop-helix (bHLH) and is primarily expressed in mature neurons and is essential for maintaining the identity of these cells. The protein product of the N gene likely acts as a regulator of Fer1, which functions downstream of N to maintain neuronal identity (FlyBase Gene Snapshot: Fer1, 2018)
- ***Elba (Early Boundary Activity)-*** Elba belongs to the GAGA family of transcription factors, which are characterized by a conserved DNA-binding domain known as the BTB/POZ domain. It is involved in the regulation of several important developmental processes, including segmentation and germ cell formation. It also interacts with other transcription factors and chromatin remodelling proteins to modulate gene expression. Elba1, Elba2, and Elba3 are three isoforms of the Elba generated from the same gene through alternative splicing and differ in their N-terminal regions. Elba1 is the longest isoform, while Elba2 and Elba3 are shorter isoforms.
- ***Scrt (Scratch)-*** It is a transcription factor that contains zinc finger C2H2 motifs. It is primarily involved in the inhibition of genes that promote the differentiation of non-neuronal cell types during development (Roark et al., 1995). This regulatory function helps ensure proper cell fate determination during embryonic development in *Drosophila* (FlyBase Gene Snapshot: Scratch (scrt), 2019).
- ***Chn (Charlatan)-*** It is a C2H2 zinc-finger transcriptional factor that plays a pivotal role in the regulation of multiple genes. Chn is known to function in diverse cell types such as sensory neurons, photoreceptors, blood cells, muscle precursors, and intestinal precursors (FlyBase Gene Snapshot: Charlatan (chn), 2018).

- ***gtSu(H) (Suppressor of Hairless)-*** It is involved in the Notch signalling pathway and plays a critical role in the regulation of gene expression during development, including cell fate determination and differentiation. In the absence of Notch signalling, gtSu(H) is bound by a co-repressor complex containing the protein Hairless, which prevents gtSu(H) from activating target genes. Upon activation of the Notch pathway, the intracellular domain of the Notch receptor is released and binds to gtSu(H), leading to the release of the Hairless co-repressor complex and the subsequent activation of target genes (FlyBase Gene Snapshot: Suppressor of Hairless [Su(H)], 2021).

1.3 Overview of Sequencing Technologies for Studying TF Binding, Chromatin Accessibility, and Gene Expression

1.3.1 Introduction to Sequencing Technologies

Next-generation sequencing (NGS) technologies, also known as high-throughput sequencing, have revolutionized genomics research in recent years. These sequencing methods have made it possible to generate vast amounts of DNA sequence data at unprecedented speed and low cost, enabling comprehensive analysis of genomes, transcriptomes, and epigenomes (Goodwin et al., 2016). NGS methods typically involve the fragmentation of DNA or RNA into smaller pieces, which are then sequenced in parallel using various platforms. These platforms can generate millions to billions of short reads, which are subsequently assembled into longer contiguous sequences or analysed individually to provide valuable information on genome structure, gene expression, and chromatin accessibility (Wang et al., 2009). One of the major advantages of NGS is its ability to generate large amounts of sequence data from a single sample. This has enabled researchers to perform genome-wide analyses and identify genomic features such as genetic variation, gene expression, and chromatin accessibility that are critical for understanding the underlying biology of various organisms and diseases (Metzker, 2010). It has also facilitated the study of transcription factor binding, chromatin accessibility, and gene expression by providing high-throughput methods for measuring these events on a genome-wide scale. The power of NGS in genomics research is exemplified by numerous studies that have characterized the neural transcriptome and epigenome, identified regulatory elements, and revealed the dynamics of transcription factor binding and chromatin accessibility during neural development.

1.3.2 Bulk vs Single-cell Sequencing

Bulk sequencing is a method of sequencing in which DNA or RNA from a group of cells is extracted, amplified, and sequenced as a single sample. This approach has been widely used in genomics research, and it provides a cost-effective way to analyse the average genetic information of a population of cells (Wang et al., 2009). It has been

instrumental in identifying genetic variants associated with diseases, understanding the genetic basis of complex traits, and characterizing the diversity of microbial communities. However, bulk sequencing has several limitations. First, it provides only an average representation of the genetic material of the cells in the sample, thereby masking the heterogeneity that exists among individual cells (Suvà & Tirosh, 2019). Second, it may overlook rare cell types or subpopulations that may be biologically important. Finally, it may not reveal the full extent of genomic variation within a population of cells.

Single-cell sequencing, on the other hand, is a newer technology that allows the analysis of the genetic material of individual cells. This approach has opened new avenues of research by enabling the identification of rare cell types, the study of cell lineage and differentiation, and the discovery of new cell types (Macosko et al., 2015; Suvà & Tirosh, 2019). Single-cell sequencing can be performed on DNA or RNA, and it can reveal the full extent of genomic variation within a population of cells. However, single-cell sequencing is more expensive than bulk sequencing, and it requires more complex data analysis pipelines to interpret the large amount of data generated. Moreover, it can be technically challenging due to the low amounts of genetic material present in individual cells and the need to maintain the integrity of the genetic material during the sequencing process.

1.3.3 Chromatin Immunoprecipitation sequencing: ChIP-seq and ChIP-nexus

As mentioned in 1.1.8, ChIP-seq is a commonly used method for locating genomic regions where transcription factors (TFs) bind to DNA. This technique involves cross-linking TFs to DNA, isolating the TF-bound DNA fragments using immunoprecipitation, sequencing the enriched DNA fragments, and finally mapping these DNA sequences back to the reference genome to identify the genomic regions bound by the TF of interest. ChIP-seq has several advantages over earlier techniques for identifying TF binding sites. For instance, ChIP-seq allows for the identification of genome-wide binding sites of a TF in a single experiment, and it can detect both strong and weak binding sites. Additionally, ChIP-seq can be used to identify both direct and indirect targets of a TF (Spitz & Furlong, 2012).

ChIP-seq approach involves in vivo crosslinking of DNA-protein complexes, followed by fragmentation of the samples and treatment with an exonuclease to remove unbound oligonucleotides. The DNA-protein complex is then immunoprecipitated using antibodies specific to the protein of interest. Subsequently, the DNA is extracted, purified, and sequenced, allowing for the generation of high-resolution sequences that identify the sites where the protein binds to DNA (**Figure 1.6A**). ChIP technology has undergone significant advancements, from ChIP-polymerase chain reaction (ChIP-PCR) for single locus detection to ChIP-ChIP microarray and ChIP-seq for high-throughput sequencing. ChIP-exo has emerged as a powerful variant of ChIP-seq, offering ultra-high resolution by improving spatial arrangement of TFs and reducing noise. ChIP-nexus, another variant of ChIP-exo, introduces a circular ligation step to enhance complexity (Figure 1.5). In 1.4.2, a comprehensive elaboration is presented, focusing on deep-learning methods within the context of TFBS prediction.

ChIP-nexus, or Chromatin Immunoprecipitation Experiments with Nucleotide resolution through Exonuclease, Unique barcode, and Single ligation, is an advanced genomic technique designed to improve the mapping of transcription factor binding sites at nucleotide resolution *in vivo*. It enhances the conventional ChIP-seq method by incorporating exonuclease treatment to precisely trim DNA fragments bound to TFs, unique barcodes to distinguish individual DNA fragments during sequencing, and a single ligation step to minimize biases.

Studies have demonstrated that ChIP-nexus outperforms existing ChIP protocols, providing a more detailed view of transcription factor binding landscapes. It can accurately identify binding sites within enhancers containing multiple binding motifs and enables analysis of *in vivo* binding specificities (He et al., 2015). While ChIP-nexus offers valuable advantages in resolution and specificity, one significant limitation is the increased cost and time associated with ChIP-nexus compared to conventional ChIP-seq. Hence, ChIP-nexus may not be as commonly used as traditional ChIP-seq due to the practical challenges it presents.

Single-cell genomics has revolutionized our understanding of cellular heterogeneity and gene regulation by enabling the study of individual cells' molecular profiles. While single-cell techniques, such as scRNA-seq and scATAC-seq, have gained popularity across various omics fields, scChIP-seq remains relatively less explored, with only a few studies conducted so far [references]. The technical challenges associated with scChIP-seq arise from the lower amounts of DNA associated with individual protein-DNA interactions in single cells, making it harder to obtain sufficient data for meaningful analysis. Moreover, the complexity of the protocol and the need for specialized bioinformatic tools for data analysis have presented challenges to its widespread adoption.

Analyzing scChIP-seq data comes with several challenges. The inherent sparsity of scChIP-seq data, resulting from the low amounts of DNA material per cell, poses difficulties in distinguishing true signals from background noise. Technical noise and batch effects introduced during the experimental workflow require careful data normalization and analysis. Moreover, the cell-to-cell heterogeneity inherent in single-cell assays makes it challenging to accurately define cell populations and identify cell-specific binding patterns. The sparse data generated in scChIP-seq, especially with certain methods like Drop-ChIP, hinders the accurate identification of chromatin features and may impose analysing thousands of cells to obtain reliable results. Another limitation is the requirement for costly microfluidic devices in some scChIP-seq methods, which may limit their widespread use due to budget constraints. Furthermore, specific cell input requirements may render scChIP-seq less suitable for samples with very low cell numbers, such as preimplantation embryos, raising concerns about the broader applicability of these methods in such cases. The complexity of scChIP-seq experimental protocols introduces technical challenges in sample preparation, library construction, and data analysis. This complexity may result in variations and biases that need to be carefully accounted for during the experimental design and data interpretation. Additionally, comparing scChIP-seq data to conventional bulk ChIP-seq can be challenging due to differences in data resolution and sensitivity, requiring the use of appropriate statistical methods for accurate comparisons (Ma & Zhang, 2020).

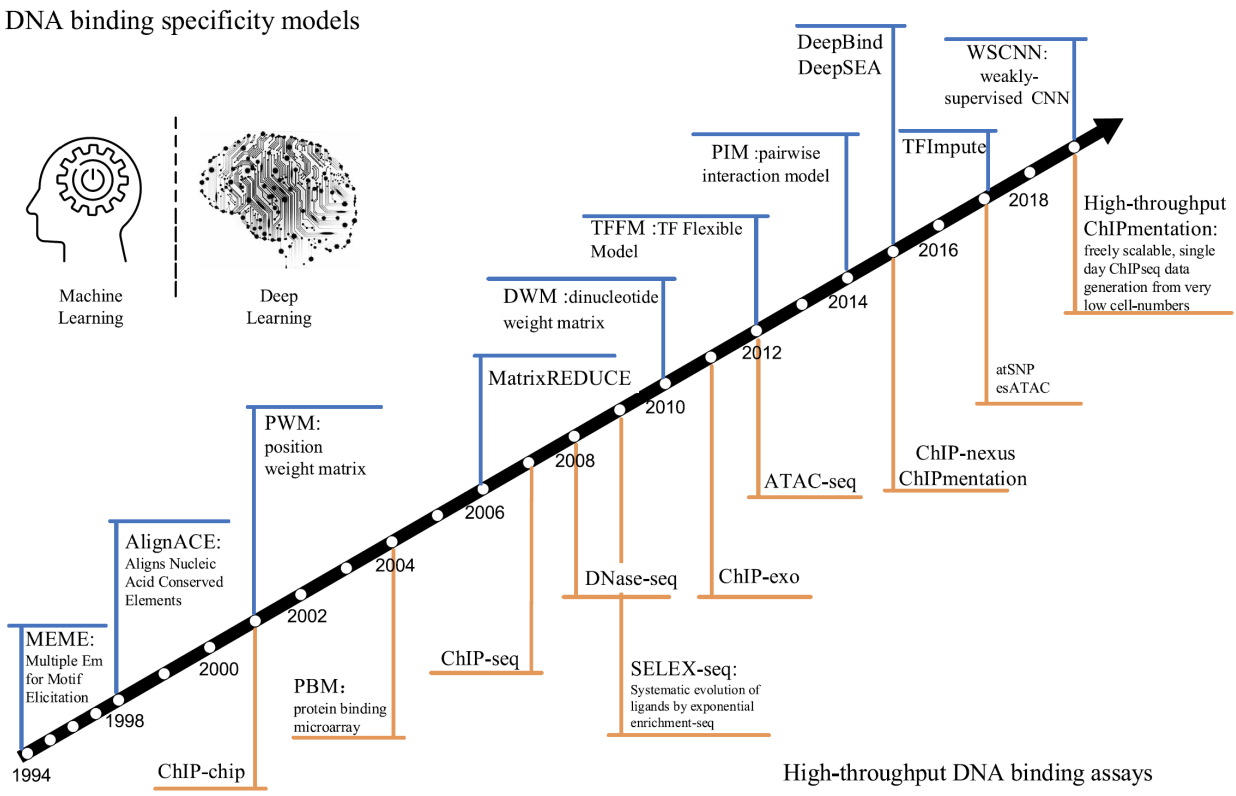


Figure 1.5: Timeline of high-throughput DNA binding assays and computational models*. High-throughput techniques have allowed for complex models to be developed through machine learning for identifying transcription factor binding sites. While these models can fit noise and bias better than simple PWM models, they may not always perform well on independent in vivo data. In vitro methods like SELEX and PBM have been developed as cheaper and faster alternatives to ChIP-seq, but they rely on sequence data and may not reflect in vivo binding specificity. Notably, the model algorithm shifted from machine learning to a deep learning approach with the advent of ChIP-seq. (Image courtesy: (Zeng et al., 2020))*

1.3.4 Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) and scATAC-seq

ATAC-seq is a sequencing protocol designed to analyse and map the accessibility of chromatin throughout the genome (Buenrostro et al., 2013). The key step in this technique is the in vitro insertion of Tn5 to a mixture containing DNA molecules and adapters, which fragments the genome and attaches adapters to the fragments simultaneously. When the chromatin is more accessible, the likelihood of transposition is higher compared to when the chromatin is less accessible. As a result, the amplified fragments obtained after the PCR step are enriched in the regions where the chromatin was accessible (Song & Crawford, 2010) (**Figure 1.6B**). ATAC-seq offers several advantages over alternative methods, such as DNase-seq and FAIRE-seq, in the study of chromatin accessibility. It requires only a small amount of initial material and can be performed in a single day, making it a highly efficient method. Furthermore, ATAC-seq provides higher resolution compared to other techniques, allowing the identification of regulatory elements at the single-nucleotide level. This increased resolution has made ATAC-seq an essential tool for investigating chromatin accessibility in a variety of contexts, including developmental processes, disease, and cellular differentiation. Additionally, ATAC-seq has been utilized to identify regulatory elements that are linked to specific diseases, such as cancer (Simon et al., 2012).

Single-cell ATAC-seq detects open chromatin in individual cells. With data being sparse, obtaining information from single cells assists in the understanding of determinants of cell-to-cell chromatin variation. This scATAC-seq approach utilizes combinatorial cellular indexing to analyse chromatin accessibility in thousands of individual cells per assay, providing a highly scalable and comprehensive method for epigenetic profiling (Cusanovich et al., 2015). By avoiding the need for compartmentalization of cells, this technique enables the simultaneous analysis of a large number of cells without compromising data quality or resolution. The method involves the isolation of nuclei, which are then tagged with barcoded Tn5 transposases in bulk using microfluidic wells. After pooling, a small subset of tagged nuclei is redistributed into a second set of wells, and a second barcode is introduced during PCR amplification to

distinguish cells and fragments from the first well. The resulting fragments are then pooled for library preparation and sequencing, allowing for the identification of chromatin accessibility patterns in each individual cell as shown in **Figure 1.6C** (Cao et al., 2017; Cusanovich et al. 2015).

1.3.5 Precision Nuclear Run-On Sequencing (PRO-seq)

Standard RNA-seq is a widely used technique for measuring the steady-state levels of total RNA in a sample, encompassing both mature and nascent transcripts present at a specific time. This approach provides high sensitivity and a comprehensive view of the transcriptome, making it well-suited for identifying differentially expressed genes between experimental conditions. However, standard RNA-seq lacks temporal resolution and is unable to capture transcriptional dynamics over time.

In contrast, nascent RNA-seq methods, such as Precision nuclear run-on-sequencing (PRO-seq), focus on selectively capturing newly synthesized RNA molecules, providing a snapshot of ongoing transcriptional activity at a given moment. Similarly, Global run-on-sequencing (GRO-seq) operates at a genome-wide level and captures nascent RNA transcripts, allowing for the identification and quantification of actively transcribed genes and enhancers throughout the entire genome. These nascent RNA-seq techniques offer higher temporal resolution, enabling the study of transcriptional dynamics and regulatory mechanisms in real-time. Nascent RNA-seq, such as PRO-seq, is especially valuable for identifying active enhancers, promoters, and regulatory elements, shedding light on the dynamic regulation of gene expression. It also facilitates the investigation of transcription factor binding dynamics and the intricate processes governing transcription initiation and elongation.

Transcription of coding as well as non-coding RNAs by RNA polymerase involves interaction with several transcription factors for a controlled and directed recruitment of polymerase, which further guides the initiation, elongation and termination (Core et al., 2008). Efficient mapping of these transcription factors along with nucleosomes and their modifications has been achieved by ultra-high throughput sequencing such as ChIP-seq. It suggests the presence of RNA polymerase II in higher amounts at the 5’ end of

eukaryotic genes. However, it remains unclear whether the polymerase is bound to the promoter or involved in transcription. Accumulation of RNA polymerase II at the 3' end is proposed to facilitate RNA processing and termination of transcription. This signifies that transcription can be regulated at the level of recruitment, initiation, and elongation with the promoter-proximal pausing as a rate-limiting step in regulation (Core & Lis, 2008). A comprehensive and quantitative snapshot of transcription can be achieved due to the ability of nascent RNA-seq to measure the density of RNA polymerase across the genome. Collection of these snapshots by the regulatory switches unveil those genes which show primary/secondary response to signals, providing useful insights of their regulation (Adelman & Lis, 2012).

GRO-seq maps the position, amount and orientation of transcriptionally engaged RNA polymerase genome-wide. In GRO-seq, bromouridine-labelled nascent RNAs are affinity purified and analysed by high throughput sequencing. Bromouridine enables RNA polymerase to add multiple nucleotides to the nascent RNA, thereby leading to a resolution of the order of tens of bases. However, molecular mechanisms of transcriptional regulation are better understood at a base-pair resolution, which can be achieved by PRO-seq (Mahat et al., 2016). PRO-seq is a modification of GRO-seq wherein biotin-labelled ribonucleotide triphosphates (NTPs) are used as a substrate for the nuclear run-on reaction for efficient affinity purification of nascent RNAs. The incorporation of first biotin base by Pol II prevents further transcript elongation. RNA containing biotin nucleotide is enriched by affinity purification using streptavidin-coated magnetic beads followed by adaptor ligation (**Figure 1.6D**). It allows for the identification of the last incorporated NTP when sequenced from 3' end, thus revealing the exact location of the active site of Pol II engaged with nascent RNA (Kwak et al., 2013). Both GRO-seq and PRO-seq calculate the transcription rates and allows assessment of polymerase pausing and elongation; the data of these methods show sharp peaks around the TSS in both sense and antisense directions. The main advantage of using these assays is that they measure the production of new RNA directly as it is transcribed rather than the concentration of mature RNAs provided by bulk RNA-seq (Mahat et al., 2016). Therefore, immediate changes in the transcription can be reported in the time course of minutes instead of hours.

The PRO-seq profile of a gene exhibits several features like higher read density at the promoter-proximal pause site compared to the gene body representing the paused polymerases; uniform read density across exons and introns, the inclusion of reads exceeding polyadenylation site and divergent reads at the promoter of mammalian genes exhibiting divergent transcription. PRO-seq holds several advantages over other sequencing methods as it provides base-pair resolution and strand-specific information of global RNA polymerase. Additionally, it reveals transcriptional stages and quantifies the expression rates in a single assay.

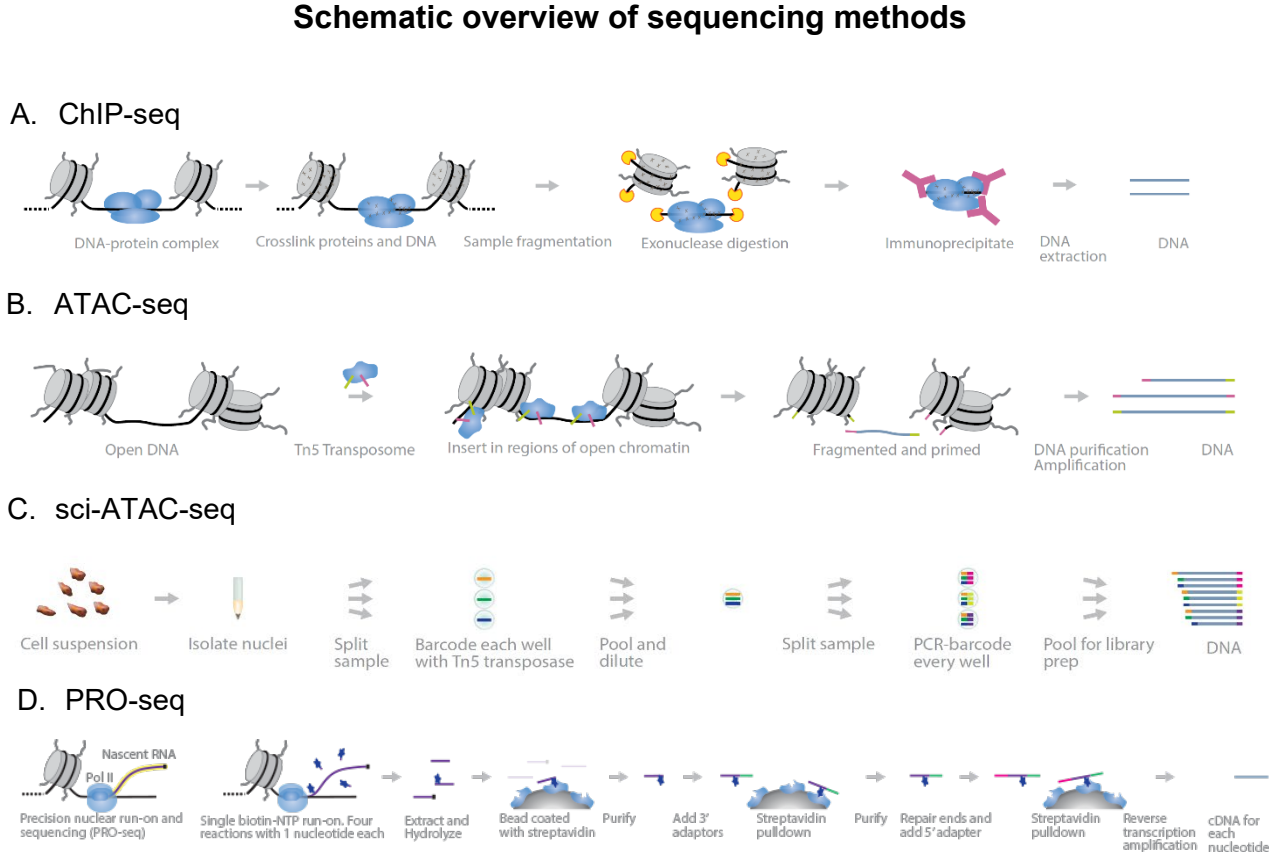


Figure 1.6: Schematic overview of sequencing methods used in this study. *A) ChIP-seq, B) ATAC-seq, C) sci-ATAC-seq and D) PRO-seq. Image source: Illumina, Inc. (2023)*

Combination of these sequencing techniques can be applied to gain insights about gene regulatory elements and their regulatory activities can be predicted using machine learning techniques based on sequence information. Computational approaches for the detection of TF binding sites can be broadly classified into two categories: (1) Machine learning methods and (2) Deep-learning-based methods.

1.4 Computational approaches for the TF binding site identification

1.4.1 Traditional machine learning methods for TF binding sites detection

Predicting TFBSs accurately is essential for understanding gene regulation and deciphering the complex network of transcriptional control. Over the years, various computational approaches have been developed for TFBS prediction, ranging from traditional machine learning (ML) to deep learning (DL) techniques. Here, I review the progress made in TFBS prediction, shedding light on the advantages of DL over traditional ML, and discussing the application of DL architectures in predicting combinatorial TF binding. Early approaches for TFBS prediction relied on handcrafted features derived from DNA sequences, such as k-mer frequencies, position weight matrices (PWMs), and DNA structural properties. Numerous studies have been conducted to improve TFBS prediction using traditional machine learning methods. Early works focused on extracting handcrafted features from DNA sequences to represent TF binding patterns. For instance, a study (Stormo & Hartzell 1989) used PWMs to model TF binding preferences and successfully identified TFBSs in bacterial genomes. Another study (Blanchette et al. 2002) employed position-specific scoring matrices (PSSMs) to detect TFBSs in eukaryotic genomes. While these methods achieved some success, they often struggled to capture the full complexity of TF binding, especially when dealing with combinatorial TF binding events.

Machine learning, a subfield of artificial intelligence (AI), emphasizes creating algorithms and models capable of learning from data to make informed predictions or decisions. The evolution of machine learning has played a pivotal role in enhancing the detection of transcription factor (TF) binding sites. Machine learning techniques for identifying TF binding sites generally fall into two categories: supervised and unsupervised learning. In supervised learning, models are trained using labelled datasets, where sequences containing TF binding sites (TFBSs) are marked as positive examples, and those without are designated as negative examples. Once trained, these models can predict the likelihood of a sequence segment being a TFBS based on its features. Notably, supervised learning algorithms such as support vector machines (SVMs), random forests,

and neural networks have been extensively utilized to differentiate sequences into TFBSs or non-TFBSs categories [references].

Although these algorithms have rendered valuable insights, they face challenges in discerning intricate sequence patterns and the context-driven interactions intrinsic to TF binding. Additionally, they often rely on manually-selected features, which may not always capture the full complexity of TF-DNA interactions.

1.4.2 Deep Learning Methods for TF binding sites detection

Deep learning (DL) is a sub-class of machine learning considered as a descendent of classical artificial neural networks (ANN) comprising multiple computational nodes arranged in form of series of layers. It has made impressive recent advances in the field of functional genomics ranging from the prediction of sequence-specificity of DNA/RNA binding proteins to DNA methylation and control of splicing (Alipanahi et al., 2015; Zhou & Troyanskaya, 2015). DL methods surpass traditional machine learning in performance by autonomously identifying relevant features from input data, thus enhancing prediction accuracy. There are four major classes of neural networks: fully connected, convolutional, recurrent and graph convolutional. With the unprecedented rise in the generation of big data in medicine, such as Convolutional Neural Networks (CNN) have been used to solve biological prediction problems like the classification of gene expression using histone modification data, prediction of chromatin accessibility and prediction of protein subcellular localization from sequence information (LeCun et al., 2015; Angermueller et al., 2016; Li et al., 2023; Vaz & Balaji, 2021). CNN is a multi-layer perceptron consisting of one or more convolution layers used to extract different features of the input, rectification layers making the output nonlinear, pooling layers to prevent overfitting, and fully connected layers to combine local features into global features and for calculating scores.

Deep learning, with its ability to automatically learn hierarchical features from raw data, has revolutionized bioinformatics, including TFBS prediction. Convolutional neural networks (CNNs) have been employed in DeepBind to accurately predict TF binding specificities from DNA sequences, outperforming traditional PWM-based approaches

(Alipanahi et al., 2015). Similarly, Zhou and Troyanskaya applied deep learning-based sequence models to predict the effects of noncoding variants, showcasing the power of deep learning in capturing complex sequence patterns (Zhou and Troyanskaya, 2015).

In recent years, various approaches have been developed to determine transcription factor binding sites using deep learning technology as mentioned in 1.3.3 **(Figure 1.5)**. Some pioneering studies are based on the application of CNN in the genomic context from DNA sequences alone. Examples include DeepSEA (Zhou & Troyanskaya, 2015), DeepBIND (B. Alipanahi et al., 2015), Basset (Kelley et al., 2016), Basenji (Kelley et al., 2018), DanQ (Quang & Xie, 2016) and BPNet (Avsec et al., 2021). Other methods such as FactorNet (Quang & Xie, 2019), Anchor (Li et al., 2019) and Leopard (Li & Guan, 2021) utilize chromatin accessibility along with the sequence to infer TF binding locations. However, these DL methodologies have limitations, particularly in their lack of cell-type-specific data, constraining their definition of binding profiles across various cell types.

The U-NET architecture strands one of the most popular and state-of-art convolutional network architectures for image segmentation. This U-shaped architecture is obtained by combination of a specific encoder-decoder arrangement connected via a bridge. The encoder block provides down sampling and does feature extraction and the decoder block does up sampling hence resulting in localization of objects. The input sequence is passed through a series of encoder blocks and each block contains multiple convolution layers followed by an optional batch normalization layer to normalize the output of previous layer on basis of batch's mean and standard deviation. It is then followed by a ReLU activation layer which adds non-linearity to the network and provides better generalization of the data. Next, a max-pooling layer is added to reduce the spatial dimensions of the input while maintaining the number of channels. Skip connections act as shortcut connections that directly connect as input to the decoder layer preventing information loss during transpose convolutions and therefore provide additional information which aids the network to learn better representation. To continue the flow of information, a bridge is used to connect the encoder and the decoder network. The decoder block takes the abstract representation of the input and uses transpose convolution to upscale it. This is followed by concatenation with the output of the corresponding skip layer feature map from the

encoder block to prevent information loss. Next, two convolution layers (same padding) with ReLU activation are used to further increase the depth of the network. The last decoder uses 1*1 convolutional layer with sigmoid activation to obtain the desired size of the output (Ronneberger et al., 2015). U-NET architecture provides fast and precise prediction even with limited training data and reduces the computational cost by decreasing the number of trainable parameters.

TransUNet is a recent extension of the UNet architecture that incorporates self-attention mechanisms inspired by the Transformer model. Transformers are a type of deep learning architecture that has been widely used in natural language processing (NLP) tasks, such as language translation and text summarization. Transformers were first introduced in the paper "Attention is All You Need" by (Vaswani et al., 2017) as an alternative to recurrent neural networks (RNNs) for processing sequential data. The Transformer architecture replaces the traditional RNN approach with a self-attention mechanism that allows the model to attend to different parts of the input sequence. Self-attention enables the model to capture long-range dependencies in the sequence more efficiently and has been shown to outperform RNN-based approaches in several NLP tasks. The success of Transformers in natural language processing has inspired their application in other domains, such as bioinformatics. Transformers have been used for tasks such as protein structure prediction, DNA sequence classification, and gene expression prediction (Angermueller et al., 2016; Cui et al., 2019; Sapoval, 2022).

In light of these advancements, our proposed study is to harness more advanced deep learning methods, particularity TransUNet architecture, to predict TF combinatorial binding sites and cell-type specific gene expression profiles in early embryonic stages of *D. melanogaster*. Our approach will integrate various regulatory elements, including TF binding motif patterns and chromatin accessibility.

1.5 Rationale and Objectives of the study

To summarise, the regulation of gene expression is a complex process, which involves various regulatory elements such as TFs and chromatin accessibility. Understanding the combinatorial binding of TFs and their impact on cell-type specific gene expression is crucial for unravelling the molecular mechanisms underlying cell-type specificity during development and disease. In this context, deep learning methods can be used to predict TF binding sites (TFBS) from DNA sequence information and chromatin accessibility data. Specifically, deep learning models can capture complex interactions between TF motifs and predict combinatorial binding sites with high accuracy. Moreover, these models can be implemented at single-cell resolution to identify cell-type-specific regulatory networks and facilitate a deeper understanding of gene regulation during cell-type development. The integration of TF binding, chromatin accessibility, and deep learning methods can provide insights into the complex interactions between different regulatory elements in gene regulation during cell-type development. In this study, we aimed to predict TF combinatorial binding sites and cell-type specific gene expression profiles in early embryonic stages of *D. melanogaster* using deep learning approaches. By integrating DNA sequence information and bulk-ATAC sequencing data, we developed a deep learning model for predicting TF combinatorial binding sites. We also applied the bulk model for prediction of TF sites at single-nucleotide resolution utilizing sc-ATAC seq data. Additionally, we built a deep learning model for predicting cell-specific gene expression profiles by combining TF binding and chromatin accessibility data.

By employing advanced deep learning techniques for TF binding prediction, this study aims to uncover fundamental principles of gene regulation that have broader implications for developmental biology and human health. Discovering the regulatory mechanisms governing cell fate decisions during early development can provide valuable insights into potential therapeutic targets for diseases. Dysregulation of TF binding and gene expression during development can lead to developmental disorders and other diseases, making these insights crucial for future therapeutic interventions and drug development. This study represents a significant advancement in the field of TF binding site prediction for several key reasons:

Predicting Combinatorial TF Binding: One of the key advancements is the ability to predict combinatorial TF binding. TFs often work in combination to regulate gene expression, and their interactions can lead to complex regulatory networks. By employing deep learning models, we can capture the combinatorial binding patterns of TFs more accurately, providing insights into the cooperative and synergistic effects of TFs in gene regulation. Understanding these interactions is crucial for unravelling the complexity of gene regulatory networks in cellular processes.

Predicting Single-Cell TF Binding: Another important advancement is the extension of the bulk model to predict TF binding at single-nucleotide resolution. Single-cell analysis allows us to delve into the heterogeneity within a population of cells and identify cell-type-specific regulatory events. By applying our deep learning model to single-cell data, we gain a finer level of resolution, enabling us to explore the dynamic changes in TF binding across individual cells during early embryonic development.

Using Advanced DL Architectures: This study utilizes more sophisticated deep learning architectures, such as U-NET and TransUNet, which are specifically designed for image segmentation tasks. By adapting these architectures to analyse the spatial patterns of TF binding in the context of chromatin accessibility, we can enhance the accuracy and interpretability of our predictions. These advanced architectures allow us to capture the spatial context of TF binding, providing insights into the three-dimensional organization of the chromatin and its impact on TF binding.

Focus on *Drosophila* Early Embryo: Studying TF binding in the early embryo of *Drosophila* is of immense importance for multiple reasons. First, *Drosophila* has long served as a valuable model organism for developmental biology due to its well-characterized genetics and experimental tools. The early stages of embryo development are critical for establishing the body plan and cell fate determination, making it a key window to understand the regulatory mechanisms underlying development. Second, the relatively simple and well-defined cell types in early embryos enable the identification of cell-type-specific TF binding events with greater precision. Third, Drosophila early embryos provide an opportunity to study TF binding and gene regulation in a highly dynamic and rapidly changing developmental context. Fourth, and importantly, our

collaboration has granted us access to *Drosophila* early embryo TF chip-seq and ATAC-seq data. This not only aligns with our shared research interests but also facilitates the swift validation of our predictions.

The rationale for studying TFs in *Drosophila* early embryos lies in the potential to uncover fundamental regulatory principles that can be extended to other developmental processes and organisms. Deciphering the regulatory mechanisms during this crucial developmental stage can shed light on conserved principles of gene regulation and provide insights into the evolution of developmental programs. Furthermore, understanding TF binding in the early embryo can have broader implications for human health, as many regulatory elements and mechanisms are conserved across species. Thus, our study aims to contribute to the understanding of gene regulation during early embryonic development and may pave the way for future investigations into the regulatory networks that underlie complex multicellular organisms.

Specific aims of the project were:

1. To develop a deep learning model to predict TF binding sites for single TF and multiple TFs from DNA sequence and bulk-ATAC-seq data.
2. To apply the bulk model for prediction of TF sites at single nucleotide resolution.
3. To predict the actual downstream gene expression by combining TF binding and chromatin accessibility.

Chapter 2: MATERIALS AND METHODS

2.1 Materials

2.1.1 Description of the dataset used in the study.

ChIP-seq: The collaborating lab of **Dr. Qi Dai from Stockholm University** generously provided ChIP-seq data for approximately fourteen neuronal-specific TFs expressed in *Drosophila* embryonic CNS, which were collected at two time windows (2 to 6 hours and 6 to 12 hours) after egg laying. These neuronal-specific TFs are Su(H), Scute (Sc), Charlatan (Chn), Asense (Ase), Senseless (Sens), Insensitive (Insv), Early boundary activity (Elba), Scratch (Scrt), Shaven (sv/D-Pax2), 48 Related 1 (Fer1), Hamlet (Ham), Tramtrak (ttk/ttk69), Hairless (H) and Sequoia (Seq). The detailed information about the data used in this study was listed in **Table 2.1** and the TFs with their motif representation was depicted in **Table 2.2.**

ChIP-nexus: ChIP-nexus data was provided by collaborators (**Dr. Qi Dai from Stockholm University)** for four TFs in early *Drosophila* embryos: Insensitive (Insv), and Early boundary activity 1/2/3 (Elba1, Elba2, Elba3). This data captures the binding patterns of these TFs with high resolution, allowing for precise identification of their binding sites within the genome (Ueberschär et al., 2019).

ATAC-seq: The bulk ATAC-seq data was also provided by Dr. Qi Dai and included samples of wild type and mutant insensitive (wtInsv) and Elba 1,2,3 (wtElba1, wtElba2, wtElba3).

Bulk RNA-seq: The bulk RNA-seq data was also provided by Dr. Qi Dai and included samples of wild type known as yellow white (yw) and Elba 1/2/3 (Elba1, Elba2, Elba3). Yellow and white (yw) are well-known genes in Drosophila melanogaster, commonly known as the fruit fly. These genes are associated with eye colour in the fruit fly and play a crucial role in the synthesis and transport of pigments responsible for eye coloration. The Yellow gene is involved in the production of pigments that give colour to the eyes

and cuticle of the fruit fly. The White gene is associated with the transport of pigments within the developing eye cells of the fruit fly. The White gene encodes a protein that acts as a transporter to move pigment precursors from the cytoplasm into the developing pigment granules within the eye cells. These data were utilized for predicting gene expression.

PRO-seq: The PRO-seq data used in this study was obtained from a collaboration with Dr. Qi Dai's lab and was derived from previous work conducted by my supervisor. This dataset specifically pertained to the 2-4 hour window of wild-type *Drosophila* embryo (Ueberschär et al., 2019). PRO-seq data accurately captured real-time transcriptional activity during this critical developmental phase. The gene body expression refers to the quantification of gene expression levels specifically within the gene body region as transcription occurs. We utilized this dataset to obtain true labels for prediction of actual downstream gene expression.

Sci-ATAC-seq: The sci-ATAC-seq (single-cell combinatorial indexing assay for transposase-accessible chromatin using sequencing) data was obtained through published study (Cusanovich et al., 2018), the authors profiled chromatin accessibility of 20,000 single cells from *Drosophila melanogaster* embryos at three different timepoints after the laying of eggs- a) 2-4 hours, b) 6-8 hours and c) 10-12 hours. These time points correspond to landmark stages in the embryo development process starting from several multipotent cells after 2-4 hours of egg laying, with major lineages in mesoderm and endoderm being specified after 6-8 hours and majority of cells undergoing terminal differentiation after 10-12 hours. The raw ATAC seq and DNase-seq data were available through NCBI GEO (accession GSE101581) and ArrayExpress (E-MTAB-5999).

Sc-RNA-seq: The sc-RNA-seq data was obtained by recently published study (Calderon et al., 2022) where the authors obtained gene expression profiles of individual cells throughout *Drosophila* embryonic development. They collected 41 time points of embryonic development, ranging from 2 to 16 hours post-fertilization, and isolated and sequenced the RNA of approximately 100,000 individual cells.

We therefore had sequencing data from different techniques **(Table 2.1)** with the same time points which could be utilized to predict TF binding at single nucleotide resolution and cell-type specific gene expression of embryonic neurogenesis in *Drosophila melanogaster*.

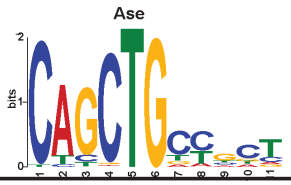
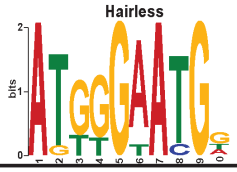
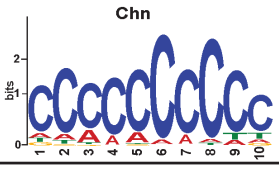
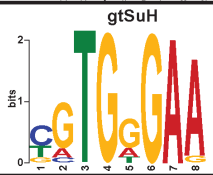
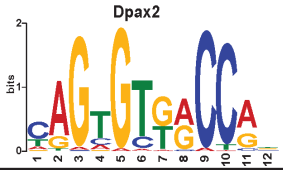
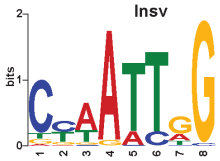
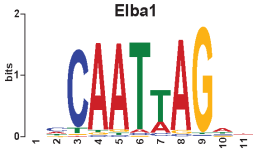
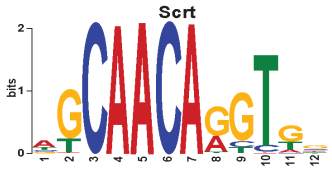
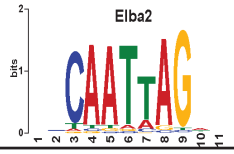
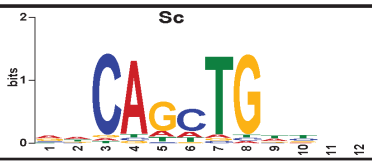
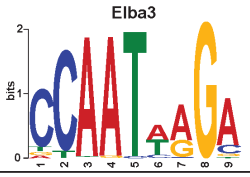
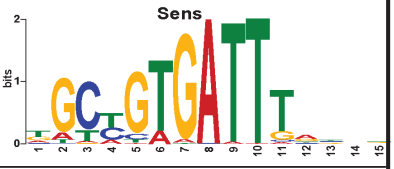
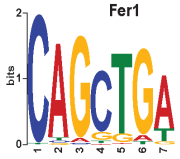
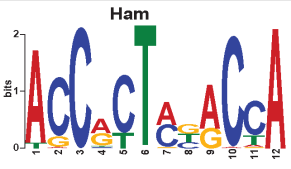
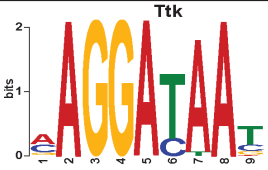
Table 2.1: Summary of sequencing datasets utilized in the study.

ChIP-seq	ChIP antibody/TF	Time-point	Mammalian homolog
	Insensitive (Insv)	2.5-6h,6.5-12h	BEND5, BEND6
	Hamlet (Ham)	2.5-6h,6.5-12h	Evi1, Prdm16
	Hairless (H)	2.5-6h,6.5-12h	HR
	Sequoia (Seq)	2.5-6h,6.5-12h	Prdm10
	Tramtrack (Ttk69/Ttk)	2.5-6h,6.5-12h	ZBTB17
	Suppressor of Hairless (SuH) & gtSuH	2.5-6h,6.5-12h	CBF1/Rbpjk
	49 related 1 (Fer1)	2.5-6h,6.5-12h	FER1L5
	Asense (Ase)	2.5-6h,6.5-12h	Asl3
	Charlatan (chn)	2.5-6h,6.5-12h	ZNF462
	Scratch (Scrt)	2.5-6h,6.5-12h	Scratch1
	Shaven (Sv/Dpax2)	2.5-6h,6.5-12h	Pax2
	Senseless (Sens)	2.5-6h,6.5-12h	Gfi-1
	Early boundary activity protein 1,2,3 (Elba)1,2,3	2.5-6h,6.5-12h	BEND5, BEND6

ChIP-nexus	TF	Time-point	
	Insensitive (Insv)	2.5-6.5 h	
	Elba 1	2.5-6.5 h	
	Elba 2	2.5-6.5 h	
	Elba 3	2.5-6.5 h	
Bulk ATAC-seq	**TF**	**Time-point**	
	Insensitive (Insv)	2.5-6.5 h	
	Mock (IgG)	2.5-6.5 h	
	Wild type (WT)	2.5-6.5 h	
	Elba 1	2.5-6.5 h	
	Elba 2	2.5-6.5 h	
	Elba 3	2.5-6.5 h	
Bulk RNA-seq	**TF**	**Time-point**	
	Yellow while (yw)	2.5-6.5 h	
	Elba 1	2.5-6.5 h	
	Elba 2	2.5-6.5 h	
	Elba 3	2.5-6.5 h	
PRO-seq	**TF**	**Time-point**	
	Insensitive (Insv)	2.5-6.5 h	

	Elba 1	2.5-6.5 h	
	Elba 2	2.5-6.5 h	
	Elba 3	2.5-6.5 h	
sciATAC-seq	**Tissue**	**Time-point**	
	Whole embryo	2-4 h	
	Whole embryo	6-8 h	
	Whole embryo	10-12h	
scRNA-seq	**Tissue**	**Time-point**	
	whole embryos	2-4h	
	whole embryos	4-6h	
	whole embryos	6-8h	
	whole embryos	8-10h	
	whole embryos	10-12h	
	whole embryos	12-14h	
	whole embryos	14-16h	
	whole embryos	16-18h	
	whole embryos	18-20h	

Table 2.2: Transcription factors and their corresponding motif logos.

TF Name	Motif LOGO	TF Name	Motif LOGO
Asense (Ase)		Hairless (H)	
Charlatan (Chn)		Suppressor of H (gtSuH)	
Shaven (sv/D-Pax2)		Insensitive (Insv)	
Early boundary activity 1 (Elba1)		Scratch (Scrt)	
Early boundary activity 2 (Elba2)		Scute (Sc)	
Early boundary activity 3 (Elba3)		Sensless (Sens)	
48 Related 1 (Fer1)		Sequoia (Seq)	
Hamlet (Ham)		Tramtrack (Ttk)	

2.1.2 Data pre-processing

The raw data for scATAC-seq was downloaded from the Sequence Read Archive using SRA toolkit using the accession ID SRR5837698. The initial steps of high-throughput sequencing data analysis were standardized across all different techniques. For each raw file used as input (single/paired-end reads in fastq format), a quality check was performed using FASTQC (FASTQ Quality Check v0.11.9) (Andrews, 2010). The quality check provided information about the quality of the sequencing data, such as base quality, sequence content, and overrepresented sequences. Next, adapters were trimmed using Trimmomatic (Bolger et al., 2014) or Cutadapt (Martin, 2011) if required. Adapters were trimmed to eliminate bias and negative effects on downstream analysis. The trimmed reads were aligned to the latest BDGP Release 6 of the *D. melanogaster* genome assembly (dm6) by STAR (Dobin et al., 2013) or Bowtie2 aligners (Langmead & Salzberg, 2012). The aligned data was converted into BAM format, which is a compressed binary file that contains sequence alignment data. Post-alignment quality check and filtering was carried out using Samtools (Li et al., 2009) and Picard (Wysoker et al., 2013). Samtools is a suite of tools that allows manipulation of SAM (Sequence Alignment/Map) format files, while Picard is a set of command-line tools that performs various tasks on BAM/SAM files. The post-alignment quality check and filtering included checking for proper alignment, removal of duplicate reads, and filtering out reads with low mapping quality.

The core or downstream analysis included peak calling, peak annotation, differential expression, and de-novo motif discovery. Peak calling identified regions of the genome that were enriched for reads and were associated with transcription factor binding sites, histone modifications, and open chromatin regions. Peak annotation involved associating peaks with genes and functional annotations. *De-novo* motif discovery involved identifying DNA sequences that were enriched in the vicinity of peak summits, which were indicative of transcription factor binding motifs.

ChIP-seq and ChIP-nexus

To ensure the quality of the data obtained from ChIP-seq and ChIP-nexus experiments, additional quality checks and analyses were conducted. For instance, I used the multiBamSummary and plotCorrelation functions from deepTools (Ramírez et al., 2014) to calculate Spearman correlations between different TF ChIP-seq bam files. I also employed the plotFingerprint function to generate cumulative read coverage profiles of ChIP-seq reads with replicates, which allowed to diagnose the consistency among replicates. In ChIP-seq and ChIP-nexus experiments, TF binding sites are identified by the accumulation of sequence reads at specific genome loci. The regions of high sequencing read density against input are considered as peaks where protein interacts with DNA. I performed peak calling using both Model-based Analysis of ChIP-Seq (MACS2) (Gaspar, 2018) and Genome-wide Event finding and Motif discovery (GEM) (Guo et al., 2012) peak callers to compare the number of peaks obtained by both methods. GEM provided high-resolution motif discovery in addition to the binding event. For peak calling, I used IgG (mock) control files for specific time-points. I further used HOMER (Hypergeometric Optimization of Motif EnRichment) (Heinz et al., 2010) to annotate peaks and to find motifs within the ChIP-seq peak regions. *De-novo* motif discovery performed using HOMER gave the specific motif pattern for each transcription factor. Motifs were stored in the form of a text like a position weight matrix (PWM) derived from nucleotide sequences. This PWM was used to run Find Individual Motif Occurrences (FIMO) (Grant et al., 2011) from MEME-suite (Bailey et al., 2009) to scan the peak regions for all the occurrences of the TF motif.

I used Bioconductor packages such as ChIPseeker and ChIPpeakAnno to visualize annotated genomic features and functional enrichment of target genes. To analyse the feature distribution of ChIP-seq peaks, the genomic coordinates of the peaks were annotated with respect to the reference genome, and the number of peaks overlapping each genomic feature was counted. Overlapping regions or the intersections between peaks from different TFs were visualized using the Intervene tool (Khan & Mathelier, 2017).

Bulk ATAC-seq

For the ATAC-seq data, post-alignment quality check was done to evaluate the quality of ATAC-seq reads. Picard was used to plot fragment size distribution which generates periodical peaks corresponding to nucleosome free regions within the first 100 bps, followed by mononucleosome at ~200bp, di-nucleosome at ~400bp and tri-nucleosome at ~600bp if successful (Roadmap et al., 2015). Fragments within first 100bp must be enriched around the TSS of genes. A Bioconductor package called ATACseqQC (Ou et al., 2018) was used to evaluate ATAC-seq specific quality measures and to generate exploratory plots. For the core analysis, the peak calling was done with MACS2 to identify the open chromatin or the accessible regions (Zhang et al., 2008). Peak calling does not directly explain the underlying mechanism of TF occupancy; hence footprint analysis is required to better understand the motif usage in binding of TFs. The factorFootprints function from ATACseqQC used TF motif in form of PWM as input to calculate the accumulated coverage for the binding sites with genome-wide occupancy of the specified TF (Ou et al., 2018). Finally, the heatmap showing signal distribution around TSS was plotted to visualize the open and closed chromatin states.

sci-ATAC-seq

Similar to bulk ATAC-seq, additional quality metrics were calculated for sci-ATAC data to determine the fragment size distribution. The processed data provided by the authors in form of sparse matrix for different time points (2-4h, 6-8h, 10-12h) was used to reproduce their results such as clustering using densityClust (Welch et al., 2019), identification of differentially accessible sites between clusters of cells and determining the fate of cell-types with pseudotemporal ordering of cells (Pseudotime analysis) using Monocle (Qiu et al., 2017). Additional analysis was performed using Seurat package (Butler et al., 2018).

To obtain cell-type specific information, I downloaded the raw and processed files from the companion website provided by authors to perform downstream analysis on the

scATAC-seq data. This dataset served as the basis for identifying cell-type specific peaks, with a particular focus on neuronal cell types.

I employed scATAC-pro tool to further process and analyse the data. It offers various functionalities, including the ability to split identified clusters into cluster-specific bigwig files. The first step was to preprocess the scATAC-seq data quality control checks, data normalization, and removing any potential batch effects. These preprocessing steps ensured that the data was of high quality and ready for downstream analysis. Next, I used the clustering algorithm within scATAC-pro to identify distinct cell populations within the dataset. The algorithm grouped cells based on their chromatin accessibility profiles, allowing us to discern individual clusters representing different cell types. Each cluster contained cells with similar chromatin accessibility patterns, indicating shared regulatory features. Once the clusters were identified, I utilized scATAC-pro to split the identified clusters into cluster-specific bigwig files. A bigwig file is a commonly used file format that contains genome-wide signal tracks, representing the level of chromatin accessibility at each genomic position. By splitting the clusters into individual bigwig files, I obtained high-resolution information about chromatin accessibility specific to each cell type. These cluster-specific bigwig files were crucial for subsequent analyses.

sc-RNA-seq

For scRNA-seq, the authors provided processed files and scripts to reproduce the analysis. I extracted the relevant data for each time point from https://shendure-web.gs.washington.edu/content/members/DEAP_website/public/ and used it for preparation of input files for the model. I used Seurat to get the average gene expression of each gene per cluster and obtained the log normalised expression values across all the cell-types.

PRO-seq

For PRO-seq, we used previously published data from our lab and collaborators (Ueberschär et al., 2019). We used gene body expression for each gene for training and testing the M2 (regression) model.

Bulk RNA-seq

For bulk RNA-seq, we utilized previously processed BED files containing gene expression values for *wild-type sample of Drosophila gene* yw (yellow white) to prepare labels. Limma-voom function was used to normalize gene expression followed by differential gene testing. For the purpose of predicting gene expression, we gathered RNA-seq dataset that encompasses the canonical transcript of genes (transcripts with longest CDS). This selection ensures that the gene expression values are linked to unique and non-overlapping promoter regions. Specifically, we focused on the upstream 1000 nucleotides and downstream 200 nucleotides of the Transcription Start Site (TSS) to define the promoter region, known to play a pivotal role in gene regulation. Additionally, the dataset was enriched with transcription factor peaks within the TSS regions, denoted by the identifier “peakTSS”. These peaks further contribute valuable information about the interaction between transcription factors and promoter regions.

2.1.3 Preparation of input files and true labels

In this study, two main files were required as input to the model. First, the bulk ATAC-seq signal coverage at single-base level and the second was one-hot encode matrix of reference genomic DNA sequence. Signal coverage was obtained by converting BAM formatted files of ATAC-seq data to bedGraph format using the bamCoverage function of deepTools with parameters such as normalization using read per genomic content (RPGC) or 1X normalization and binSize as 1. The effectiveGenomeSize parameter was also required for 1X normalization. It refers to the part of the genome that is mappable and excludes the large stretches of NNNN. The effective genome size was selected according to the read length and organism as described in the deepTools documentation
(https://deeptools.readthedocs.io/en/latest/content/feature/effectiveGenomeSize.html).
The blacklist file for dm6 assembly (in bed format) was provided to exclude these regions from analysis. Also, uncharacterized/unknown, random, extra and mitochondrial chromosomes were ignored while normalizing (ignoreForNormalization) to avoid

computational bias. Finally, the bedGraph file consisting of chromosome name, start position, end position, and binary labels was converted into bigwig format using the ‘bedGraphToBigwig’ utility from UCSC tools. This allowed the bed file to be visualized as a wiggle track in the UCSC Genome Browser (Karolchik et al., 2003).

For labelling, ChIP-seq peaks were used to scan for TF motifs within the peak regions using FIMO. The ‘getfasta' function of bedtools (Quinlan & Hall, 2010) was used to extract sequences of peak regions (obtained from MACS2 peak file in bed format) from the reference genome. The extracted sequence file was then used to run FIMO with the specific TF motif as the position weight matrix (PWM). This step provided all the occurrences of the TF motif with their specific positions within the peak regions. Next, the peak file was modified to use the motif length and position as positive labels (labelled as 1), and the remaining genomic positions were labelled 0.Within a 2000-base pair span, the nucleotides containing the TF motif (with its specific length) were marked as '1,' while the adjacent nucleotides lacking the motif were marked as '0.'

The DNA sequence or nucleotides (A,C,G,T) were encoded into a binary numeric matrix using a standard method called one-hot encoding. This method assigns a value of 1 to the nucleotide if it is present and 0 to the remaining three. For example, A is encoded as [1,0,0,0], C as [0,1,0,0], G as [0,0,1,0] and T as [0,0,0,1]. The promoter regions of genes with 1000 bp upstream and 1000 bp downstream of the Transcription Start Site (TSS) were used as the training set.

2.1.4 Description of Software and Hardware used

We trained our model on the state-of-the-art Nvidia DGX A100 system provided by the Australian National Computational Infrastructure. This cutting-edge system features eight high-performance A100 GPUs, which enabled us to train our model rapidly and efficiently.

To build and train our deep learning model, we utilized the TensorFlow 2.8.0 (Abadi et al., 2016) framework, which is a highly efficient and widely used platform for implementing neural networks. In addition, we leveraged the capabilities of TensorFlow Addons 0.16.1 and SciPy 1.9.3 (Virtanen et al., 2020) to evaluate the performance of our

models, while Matplotlib 3.6.1 (Hunter, 2007) and Seaborn 0.12.1 (Waskom et al., 2020) were used for data visualization, enabling us to generate clear and informative plots and graphs. In addition to the software and packages mentioned earlier, we also utilized a combination of R (Ihaka & Gentleman, 1996), Bioconductor packages (Gentleman et al., 2004) and Linux-based tools for data pre-processing and downstream analysis as mentioned in **Table 2.3**.

Table 2.3: List of Software and tools employed in this study.

S.No	Software	Purpose of Usage	Package source
1	R	used for statistical analysis and data visualization	(Ihaka & Gentleman, 1996)
2	FASTQC	quality control of high-throughput sequencing data	(Andrews, 2010)
3	Samtools	manipulating and processing SAM/BAM files	(Li et al., 2009)
4	Deeptools	processing BAM files and normalization	(Ramírez et al., 2014)
5	HOMER	motif discovery and functional annotation of ChIP-seq peaks	(Heinz et al., 2010)
6	MEME-Suite	discovery and analysis of sequence motifs in data	(Bailey et al., 2009)
7	FIMO	for searching for sequence motifs in DNA sequences	(Grant et al., 2011)
8	MACS2	peak caller to identify regions of the genome that are enriched for protein binding	(Zhang et al., 2008)
9	UCSC Genome Browser	visualizing and analysing genomic data	(Karolchik et al., 2003)
10	GEM Peak Caller	peak calling and Motif identification	(Guo et al., 2012)
11	Picard	quality control and duplicate removal	(Wysoker et al., 2013)
12	Trimmomatic	quality control and trimming of high-throughput sequencing data	(Bolger et al., 2014)
13	Cutadapt	removing adapter sequences from data	(Martin, 2011)
14	bedtools	manipulating and analyzing genomic data in the BED format	(Quinlan & Hall, 2010)
15	Intervene	analysis of overlapping genomic regions, including identification of shared and unique regions and visualization of overlaps.	(Khan & Mathelier, 2017)
16	ATACseqQC	quality control and analysis of ATAC-seq data	(Ou et al., 2018)
17	scATAC-pro	processing and analyzing single-cell ATAC-seq data	(Yu et al., 2020)
18	Seurat	analysis of single-cell RNA-seq data	(Butler et al., 2018)

19	Monocle	trajectory analysis and visualization of gene expression dynamics in single-cell data	(Qiu et al., 2017)
20	Cicero	Calculating gene-activity scores	(Pliner et al., 2018)

2.2 Methods

2.2.1 Model building strategy and architecture.

Given the availability of data both at the bulk and single-cell levels, we devised a two-step approach to effectively predict transcription factor binding sites (TFBS) across varying resolutions. Our methodology exploits on ChIP-seq data, utilizing DNA sequences and bulk ATAC-seq profiles as inputs to train the model and predict TFBS at bulk level and single nucleotide resolution.

In the initial step, we employed ChIP-seq peaks as labels to create a robust model. DNA sequence information and the coverage signal from bulk ATAC-seq data were utilized as inputs. The Trans-U-Net architecture, known for its adeptness in capturing intricate relationships within biological data, was harnessed to build this model. By implementing a classification model (M1) and a regression model (M2), we obtained comprehensive insights into TFBS and gene expression levels, respectively. This step provided a solid foundation for predictive capabilities at the bulk level.

Expanding on this foundation, the second step involved the extension of our model to predict TFBS at the single-nucleotidel resolution. This step leveraged processed single-cell ATAC-seq (scATAC-seq) data, which had undergone meticulous preprocessing steps. This processed scATAC-seq data, along with cell-type information extracted from the same dataset, served as inputs for the previously developed bulk model. This strategy allowed us to predict TFBS in a single-cell context, unveiling cell-type-specific regulatory interactions and contributing to a deeper understanding of gene regulation dynamics.

The Trans-UNet architecture utilized self-attention mechanisms from the Transformer network to learn the relationship between the input and output of the network. This self-attention mechanism allowed the model to attend to different parts of the input when making predictions for the output. The architecture used in this project was composed of several key components, including Squeeze and Excitation (SE) Encoder Blocks, Transformer Encoder Blocks, SE Decoder Blocks and SE Block. Similar to the UNet architecture, the encoder and corresponding decoder at the same level in the

model had a skip connection (SC). This allowed the decoder to utilize the concatenation of the upsampled feature map with the corresponding higher resolution encoder feature map at that level, which helped to preserve spatial information and improve segmentation accuracy.

SE Encoder Blocks and SE Decoder Blocks incorporated Squeeze and Excitation blocks, which were used to capture channel-wise dependencies in the input data. This improved the model's ability to attend to important features and ignore irrelevant ones, leading to improved segmentation performance. Transformer Encoder Blocks were used to capture long-range dependencies in the input data. These blocks are commonly used in natural language processing tasks but have also shown promise in image segmentation tasks. The SE Block was used to capture dependencies between channels **(Figure 2.1)**.

Extension to Gene Expression Prediction: Building on our robust model architecture designed for Transcription Factor Binding Sites (TFBS) prediction, we enhanced our predictive scope to predict gene expression levels. Our approach was twofold:

1. We implemented a regression model that utilized both PRO-seq and RNA-seq as response variables to predict gene expression. This adaptation was accomplished by adjusting the final layer of our DL model to “linear”.
2. We established a classification model to discern between high and low gene expression levels.

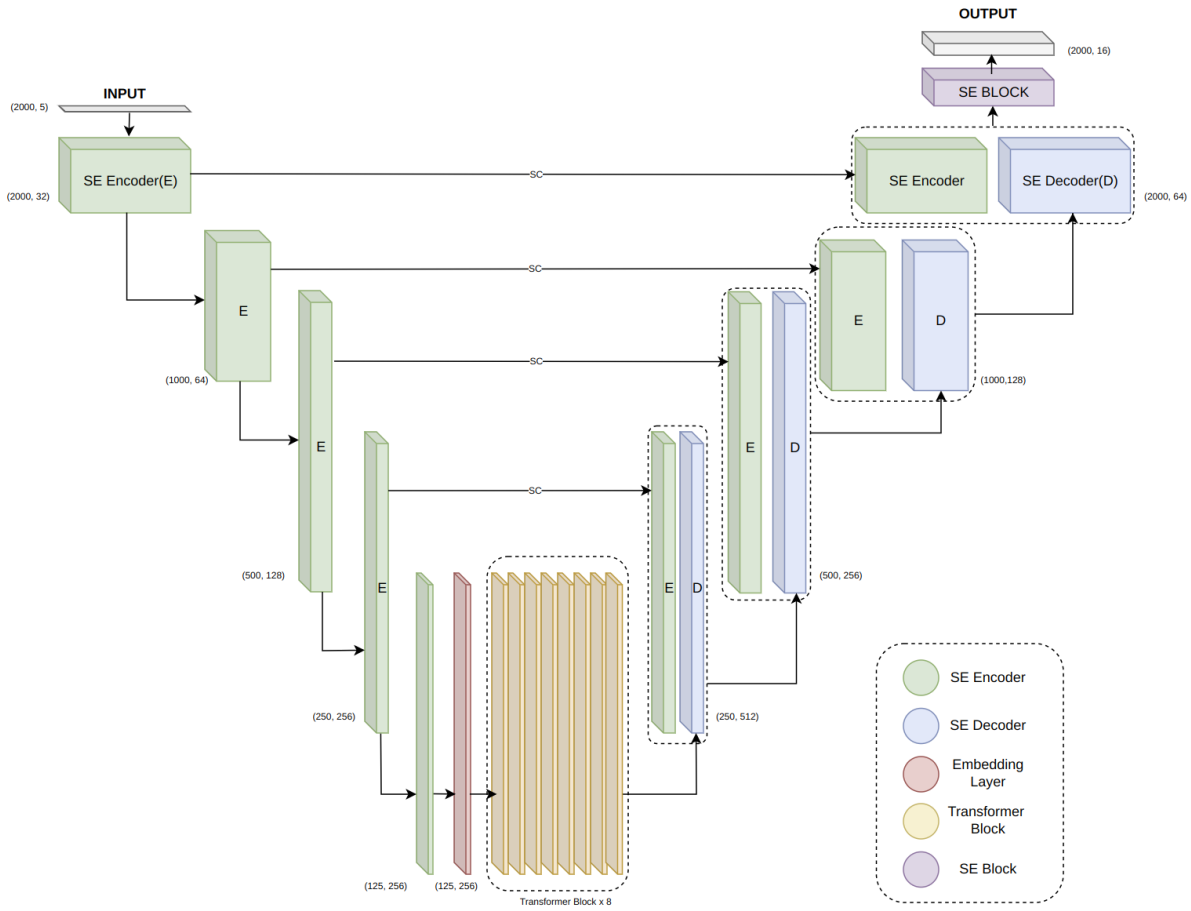


Figure 2.1: Deep learning model architecture. *The architecture was designed to extract informative features from genomic data inputs. It incorporates several key components, including Squeeze and Excitation (SE) Encoder Blocks, Transformer Encoder Blocks, SE Decoder Blocks and SE Block and has U-shaped arrangement of these layers. Like UNet, encoder and decoder blocks at each level are connected through a skip connection (SC) to maintain spatial information.*

2.2.2 Description of model training process and hyperparameter tuning

We first trained the model on individual TFs (single TF model) and then trained it for all the TFs together (multi-TF model). During the training process, the model was trained on the training dataset (85%) using the Adam optimizer with a learning rate of 0.001. We used a batch size of 32 and trained the model for 100 epochs. During training, we monitored the model's performance on a validation dataset and saved the best-performing model based on the validation loss. We also used early stopping to prevent overfitting, stopping training if the validation loss did not improve after 10 epochs. The

model was trained until the loss was converged or was halted with early stopping method to stop training once the model performance stops improving.

2.2.3 Evaluation metrics used to assess the performance of the model

By using multiple evaluation metrics, the strengths and weaknesses of the model can be identified and addressed, leading to better overall performance. The area under the curve (AUC) of the receiver operating characteristic (ROC) and precision-recall (PR) curve are commonly used evaluation metrics for classification models. The AUC-ROC is a measure of how well the model can distinguish between the positive and negative classes, while the PR-AUC is a measure of how well the model can identify the positive class. The AUC-ROC is calculated by plotting the true positive rate (TPR) against the false positive rate (FPR) at different classification thresholds. The TPR is the ratio of correctly classified positive samples to the total number of positive samples, while the FPR is the ratio of incorrectly classified negative samples to the total number of negative samples. A higher AUC-ROC indicates better performance, with a value of 1 indicating perfect performance.

The PR-AUC is calculated by plotting the precision against the recall at different classification thresholds. Precision is the ratio of correctly classified positive samples to the total number of predicted positive samples, while recall is the ratio of correctly classified positive samples to the total number of actual positive samples. A higher AUC-PR indicates better performance, with a value of 1 indicating perfect performance.

In addition to the AUC-ROC and PR-AUC, other metrics were also calculated to evaluate the performance of the model. The F1 score is the harmonic mean of precision and recall, with a value between 0 and 1, where a higher value indicates better performance. The Matthews correlation coefficient (MCC) is a measure of the correlation between the predicted and actual classes, with a value between -1 and 1, where a higher value indicates better performance. Finally, accuracy is a measure of the proportion of correctly classified samples out of the total number of samples. The AUC-ROC and PR-

AUC were calculated using the 'tf.keras metrics'. Additionally, the F1 score, MCC and accuracy were calculated using 'sklearn.metrics'.

R-squared (R^2) is a common evaluation metric used for regression models. It measures the proportion of the variance in the dependent variable that is explained by the independent variables in the model. The R^2 value ranges from 0 to 1, where a value of 1 indicates that the model fits the data perfectly, and a value of 0 indicates that the model does not explain any of the variability in the dependent variable. R^2 is calculated by dividing the explained variance (SS_{reg}) by the total variance (SS_{tot}), where:

SS_{reg} is the sum of squared differences between the predicted values and the mean of the dependent variable, and

SS_{tot} is the sum of squared differences between the actual values and the mean of the dependent variable.

The formula for R^2 is:

$$R^2 = 1 - (SS_{reg} / SS_{tot})$$

A higher R^2 value indicates that the model is a better fit for the data, as it explains more of the variance in the dependent variable.

2.2.4 Model interpretation methods.

To further explore the biological insights gained from our trained model, we utilized saliency maps to visualize the motifs that were most influential in the model's predictions. Saliency maps are a commonly used visualization technique that highlights the important features in the input data that contribute to the model's predictions. They are generated by computing the gradient of the output with respect to the input data and visualizing the regions of the input that have the highest gradient values. These regions are indicative of the motifs that have the greatest impact on the model's prediction.

To generate saliency maps, we used a technique known as input-masked gradients, which combines the gradient values with the input sequences to visualize the

segments that have the greatest influence on the model's prediction. This approach has been previously used to interpret the relationship between input and prediction in machine learning models (Eraslan et al., 2019).

2.2.5 Strategies for downstream analysis of predicted values

Given the challenge of obtaining TF binding information in single cells through the combination of ChIP-seq and ATAC-seq, direct methods for validating our prediction scores were not feasible. Therefore, we resorted to employing a few indirect approaches to ensure that our predictions of TF binding sites in single cells were as accurate as those in bulk data. Hence, alternative methods for validation of the prediction scores were explored. To validate the accuracy of our predictions, we conducted an assessment of enriched sequence motifs in the predicted TF regions and investigated whether these motifs aligned with the known binding profiles of the neuronal TFs. We utilized the HOMER tool to conduct motif enrichment analysis.

Furthermore, to assess the robustness of our model, we subjected the predicted regions from our single-cell model to several downstream processing methods, including clustering of cell types and calculation of TF activity scores. Additionally, we compared the results of our single-cell model to those obtained from the original scATAC-seq data to evaluate the consistency of cell type clustering across different modalities. This comprehensive analysis allowed us to further validate the accuracy and reliability of our single-cell model for predicting TF binding sites.

To accomplish these tasks, we utilized the scATAC-pro tool (Yu et al., 2020), which is a computational tool that enables the joint analysis of chromatin accessibility data (such as ATAC-seq) and TF binding data (such as ChIP-seq) at the single-cell level, as well as the integration of gene expression data from single-cell RNA sequencing. However, since this tool currently only supports human and mouse datasets, we locally adapted the scripts to accommodate the analysis of *Drosophila* genome and neuronal TF motifs. By customizing the scripts in this way, we were able to extend the utility of scATAC-pro to our specific research question, facilitating the comprehensive and robust analysis of our single-cell multi-omics data. Clustering analysis was employed to group similar data

points or entities based on specific characteristics. In the context of single-cell genomics, it aimed to group cells with similar gene expression profiles to gain insights into cell types and states. The Seurat package was utilized to perform clustering using the Louvain Algorithm. Motif analysis aimed to identify short DNA sequences indicative of regulatory elements or transcription factor binding sites. The chromVAR tool was utilized to calculate z-scores and deviations, revealing motif enrichment or depletion across different cell clusters. This provided insights into motifs that were more or less active in specific cell populations. Next, differential motif enrichment was performed. A barcode-cluster mapping was created to associate cell clusters with specific motifs. By comparing motif presence across clusters, differential motif enrichment analysis identified motifs statistically enriched in certain clusters, shedding light on potential regulatory networks associated with distinct cell populations. To investigate the co-occurrence patterns of transcription factor motifs in genomic regions, a motif co-occurrence correlation heatmap was generated. This heatmap offers insights into potential cooperative interactions between TFs and aids in understanding their collaborative roles in gene regulation. The process was executed using the following steps. Data from various TFs, along with their corresponding genomic coordinates, were collected and concatenated into a single DataFrame while maintaining the information about the TF each region belongs to. Duplicate entries were removed to ensure accuracy in subsequent analyses. A pivot table was constructed from the combined data to form a motif presence matrix. This matrix captured the presence (1) or absence (0) of each TF motif in genomic regions. The correlation matrix was computed from the motif presence matrix. Each cell of the correlation matrix indicated the degree of correlation between the presence/absence patterns of two TF motifs.

For the reanalysis of scATAC data from published study (Calderon et al., 2022), the bigwig files for neural clusters were downloaded from https://shendure-web.gs.washington.edu/content/members/DEAP_website/public/ATAC/revision/bigwigs/ .The bigWig file, which contained the original data, was first converted into a bedGraph format. This format allowed for easier manipulation and analysis of the data. The values in the bedGraph file were then normalized to a range of 0-1. This normalization process involved finding the minimum and maximum values in the file and scaling all the values

accordingly. Next, unwanted chromosomes such as ChrM and other unrecognised chromosomes were removed from the normalized bedGraph file. This step was necessary to remove any potential noise from the data. Finally, the filtered bedGraph file was converted back into a bigWig format. This format is more compact and efficient for storing large genomic datasets. The resulting bigWig file was then used as an input for the multi-TF model to predict TF binding sites across various cell types.

Furthermore, the code extracts DNA and single-cell ATAC-seq data values from the respective BigWig files based on genomic coordinates found in the reference BED file containing genome coordinates of 1000 flanking regions for gene promoters. The DNA sequences are one-hot encoded using the method outlined in the "data_utils.py" script from the GitHub repository. These encoded sequences are then combined with the ATAC-seq data to create batches (genome_seq_batch). Each batch contains the merged DNA and ATAC-seq data for specified genomic regions. In deep learning, a batch refers to a subset of the training dataset that is used to train a neural network during one iteration. When training a model, the dataset is often too large to be processed all at once due to memory constraints. Instead, the dataset is divided into smaller batches, and the model parameters are updated after each batch. For example, Mini-Batch Gradient Descent: Training with batches is commonly referred to as mini-batch gradient descent. This is a compromise between stochastic gradient descent (which uses one sample at a time) and batch gradient descent (which uses the entire dataset). Mini-batch gradient descent helps to make the training process faster and more stable. In summary, batches are used in deep learning to efficiently train models on large datasets by processing smaller subsets of the data at a time.

The code to run the model and all the associated analysis is present on the GitHub repo here- https://github.com/jiayuwen/scTFNet/

Chapter 3: RESULTS AND DISCUSSION

3.1 Integration of Transcriptional Regulation and Chromatin Accessibility: Insights into Gene Expression Dynamics

3.1.1 ChIP-seq data analysis reveals TF binding and motif occurrence at TSS.

In this section, I evaluated the quality and consistency of different neuronal TF ChIP-seq samples through various exploratory analyses as described in Methods. To determine the overall similarity and consistency between different neuronal TFs at two different time points, I assessed the overlap between TF peaks at 2.5 to 6.5 hours and 6.5 to 12 hours after egg laying. Venn diagrams were used to visualize the number of overlaps in related TF, and three-way and four-way Venn diagrams were computed for this purpose. Several TFs are known to exhibit cooperativity, where their binding to DNA is dependent on the presence of other TFs. For example, the TFs Ham, Insv have been shown to exhibit cooperative binding with each other, and their binding sites often overlap in vivo. Similarly, there is evidence to suggest that Insv and Elba, may also exhibit cooperativity. It is known that Insv and Elba co-localize to the same genomic regions in *Drosophila* embryos, suggesting that they may interact with each other to regulate gene expression. Additionally, Insv was shown to recruit Elba to its binding sites, and mutations in Insv that disrupted this interaction led to defects in embryonic patterning. These cooperativity relationships play a crucial role in the regulation of gene expression, as they allow for more precise and coordinated control of gene expression patterns.

To evaluate the quality and consistency of the ChIP-seq replicates, a fingerprint plot was used, which provides a useful visual representation of the overlap between replicates. The plot shows the cumulative fraction of reads that are common between two replicates, as a function of the fraction of reads in each replicate. As shown in **Figure 3.1.1A**, the results indicated a strong consistency among the replicates, with a high degree of overlap between the replicates for all TFs. This suggests that the replicates are of high quality and can be used confidently in model training and downstream analyses.

However, a slight variation was observed between Seq at 6.5-12h and Seq at 2.5-6.5 time-points, which may be attributed to biological variability or technical factors. Despite this variation, the overall consistency between replicates was found to be high, indicating that the ChIP-seq data is of high quality and suitable for further analysis. The high consistency between replicates is also an indication that the experimental protocol was performed with high precision and accuracy.

Figure 3.1.1B provides an example of TFs exhibiting cooperative binding, as demonstrated by the number of overlapping binding regions among four TFs: Ham, Insv, Hairless, and gtSuH. These TFs are known to exhibit cooperative behaviour. In addition, another group of potentially cooperating TFs, including Ase, Fer1, gtSuH, Sens, and Seq, displayed a considerable number of overlapping peaks, with 6971, 3388, 8041, 3698, and 9753 overlapping peaks, respectively.

The feature distribution of ChIP-seq peaks was analysed by annotating the genomic coordinates of the peaks with respect to the reference genome and counting the number of peaks overlapping each genomic feature (as outlined in the methods section). Visualizing intersections of multiple genomic region sets can become challenging as the number of sets increases beyond four. To illustrate this, I generated an UpSet plot for the intersections of ChIP-seq peaks for the two timepoints separately. The interactions were ranked by frequency and the resulting plot is shown in **Figure 3.1.2**.

Furthermore, the pairwise fraction of overlap was calculated to identify the intersection of peaks between all TFs, as depicted in **Figure 3.1.1C**. The pairwise fraction of overlap analysis allows us to determine the degree of similarity between the peaks of each TF at the two different time-points. The results shows a high degree of overlap between the peaks of certain TFs, particularly among those known to exhibit cooperativity, such as Insv and Elba, Hairless and Seq, Ase and Seq, Ham and gtSuH, among others. These pairs demonstrate significant correlation with the correlation coefficient (percentage overlap) ranging between 0.71 to 0.87 as shown by brown colour of pie-charts. No significant correlations were observed between Fer and other TFs with the correlation coefficient of only 0.07-0.15 depicted by pale yellow colour of pie chart.

These findings suggest that the binding of these TFs is closely related and potentially working together to regulate gene expression during neural development. These exploratory analyses provide insight into the quality and consistency of the different neuronal TF ChIP-seq samples, enabling a more comprehensive understanding of the data. It is important to note that the observed degree of overlap between the TF peaks could also be attributed to shared motifs or binding preferences. Therefore, further analysis such as motif analysis could help us understand the underlying mechanism of cooperativity between these TFs. Nevertheless, the results obtained from this pairwise overlap analysis provide valuable insights into the regulatory relationships between different TFs and highlight potential cooperative interactions that could be further explored.

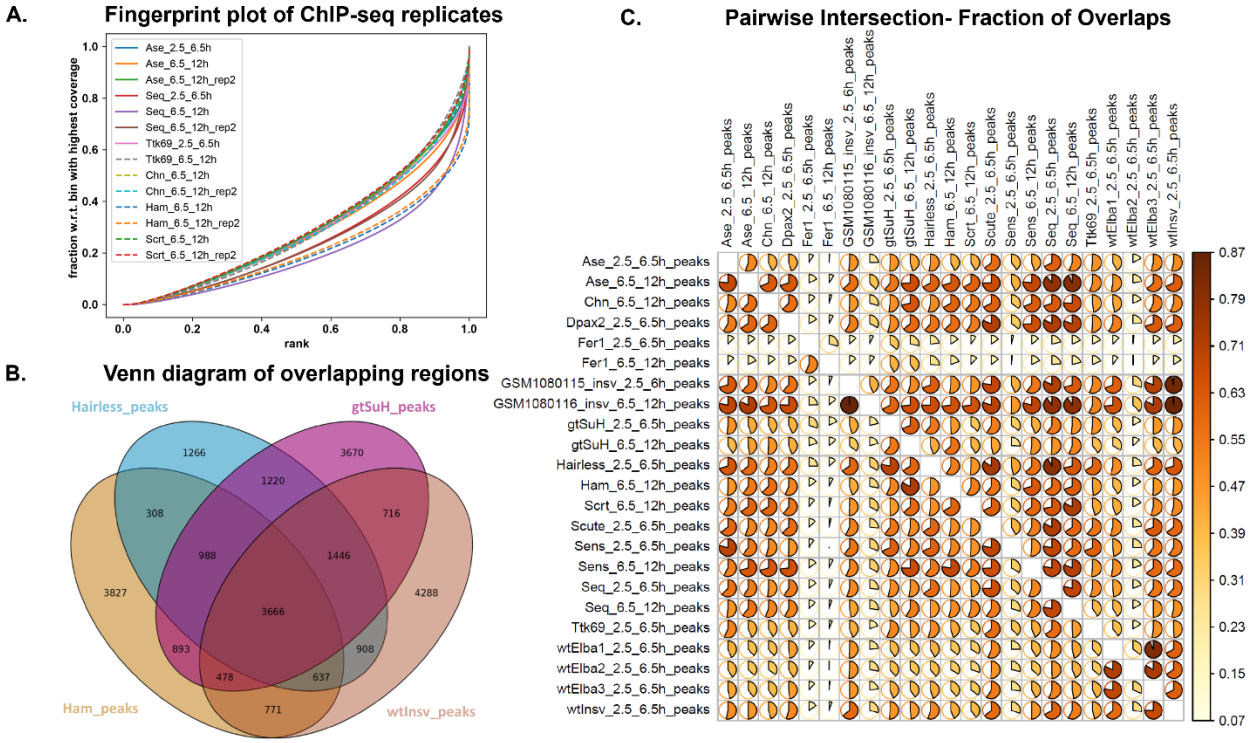


Figure 3.1.1: ChIP-seq analysis of PNS specific TFs. *A.) Fingerprint plot for ChIP-seq replicates at different time-point. This plot validates the ChIP-seq data and examines the quality and similarity between the samples and their replicates. Strong consistency observed among replicates for all TFs, indicating high quality. Slight variation between Seq 2.5-6.5(red solid line) and 6.5-12h (purple) time-points noted, possibly due to biological or technical factors. B) a four-way Venn diagram*

showing the number of overlapping peaks between four correlated TFs- Hairless, gtSuH, Hamlet, Insensitive. This Venn Diagram with Hairless (blue, 1,266 peaks), gtSuH (purple, 3,670 peaks), Hamlet (yellow, 3,827 peaks), and Insensitive (red, 4,288 peaks) illustrates the cooperative binding behaviour of these TFs. Notably, the central area where all four TFs intersect comprises 3,666 shared peaks and indicates a significant overlap, highlighting the complex interplay and shared regulatory regions among them. C.) Pairwise interaction of fraction of overlap between all ChIP-seq samples. The interaction plot demonstrates the overlapping rate between peaks of each sample and the pie chart depicts the percentage of overlapping peaks. It showcases a notable overlap/correlation among TF peaks, especially between cooperative pairs like Insv and Elba, Hairless and Seq, Ase and Seq, Ham and gtSuH depicted by brown colour and ranging between 0.71 to 0.87, suggesting a synergistic role in gene regulation during neural development. Conversely, minimal overlap is observed between Fer and other TFs depicted by pale yellow and ranging from 0.07 to 0.15.

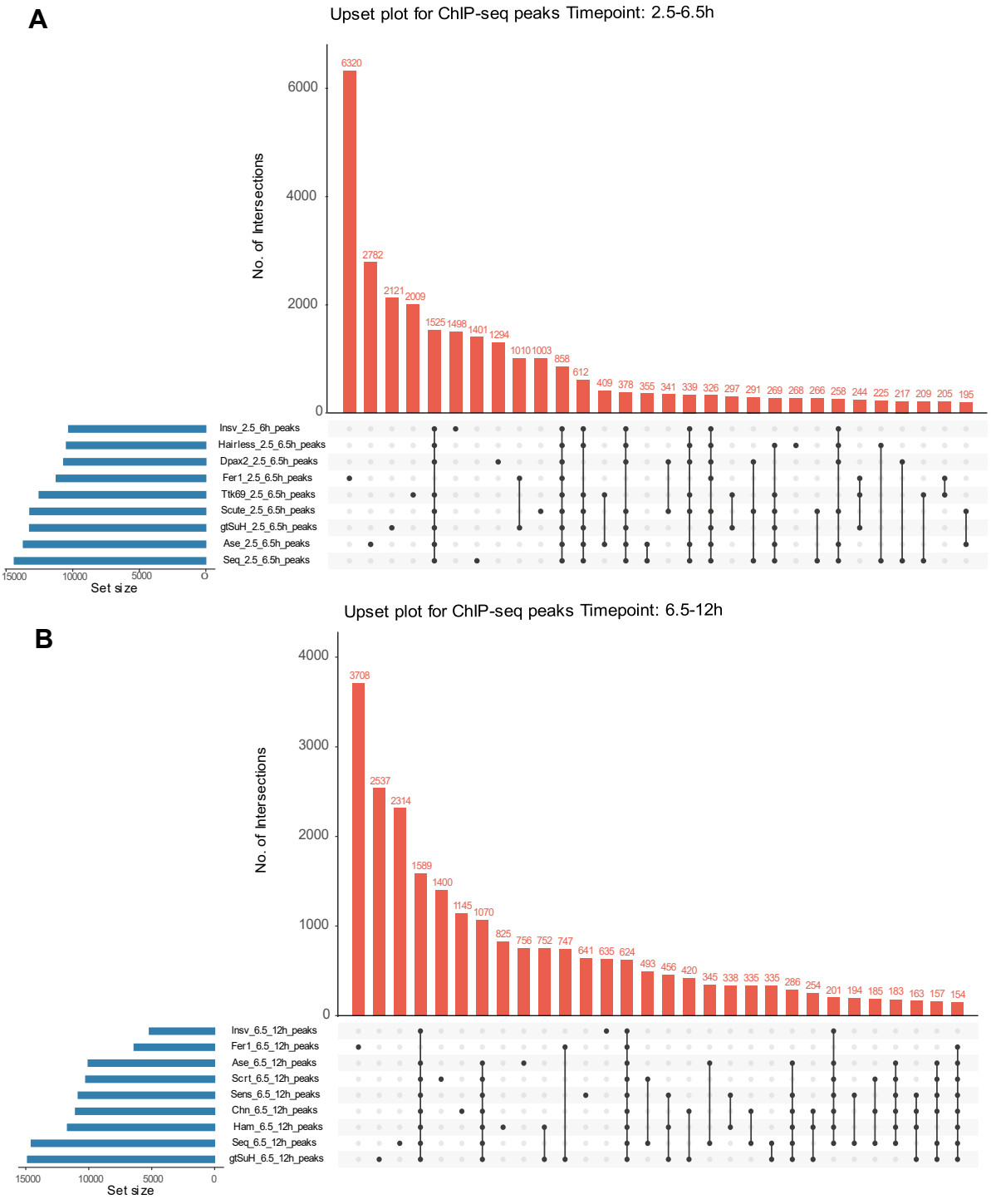


Figure 3.1.2: Upset Plot of overlapping regions at different timepoints. *The UpSet plot shows the number of intersections for samples at Time-point: A. 2.5 to 6.5 hours AEL. B. 6.5 to 12 hours AEL.*

The average profile plot of ChIP-seq samples centred at transcription start sites (TSSs) provides a visualization of the distribution of reads across the genome relative to the TSSs. It is a useful tool to assess the specificity and quality of the ChIP-seq data. In this study, an average profile plot was generated for all the ChIP-seq peaks of the samples at both the timepoints as shown in **Figure 3.1.3**. The plot shows the average signal intensity of all the peaks centred at the TSSs. The x-axis represents the genomic distance relative to the TSSs, while the y-axis represents the peak count frequency or the number of times a peak occurs at a particular genomic region, which helps in visualizing the distribution and density of peaks across the genome.

The average profile plot can reveal the presence of broad peaks, indicating the binding of transcription factors to the entire gene region, or sharp peaks, indicating specific binding to a particular region of the gene. It can also identify the presence of noise or background signals, which can affect the accuracy of downstream analysis. In **Figure 3.1.3**, the average profile plot clearly depicts the binding patterns of various TFs, with most showing binding at the TSS and Fer1 exhibiting a peak slightly before the TSS. This information suggests that the ChIP-seq data was of good quality overall, enabling accurate detection of TF binding patterns. It also indicates that the TF binding is predominantly located near TSSs. However, it should be noted that the number of binding sites for different TFs varied across all samples. For instance, in the figure **3.1.3A**, the number of binding sites for Elba1, Elba2, Elba3 and Insv were significantly lower compared to other TFs. Overall, the average profile plot provided a useful summary of the ChIP-seq data.

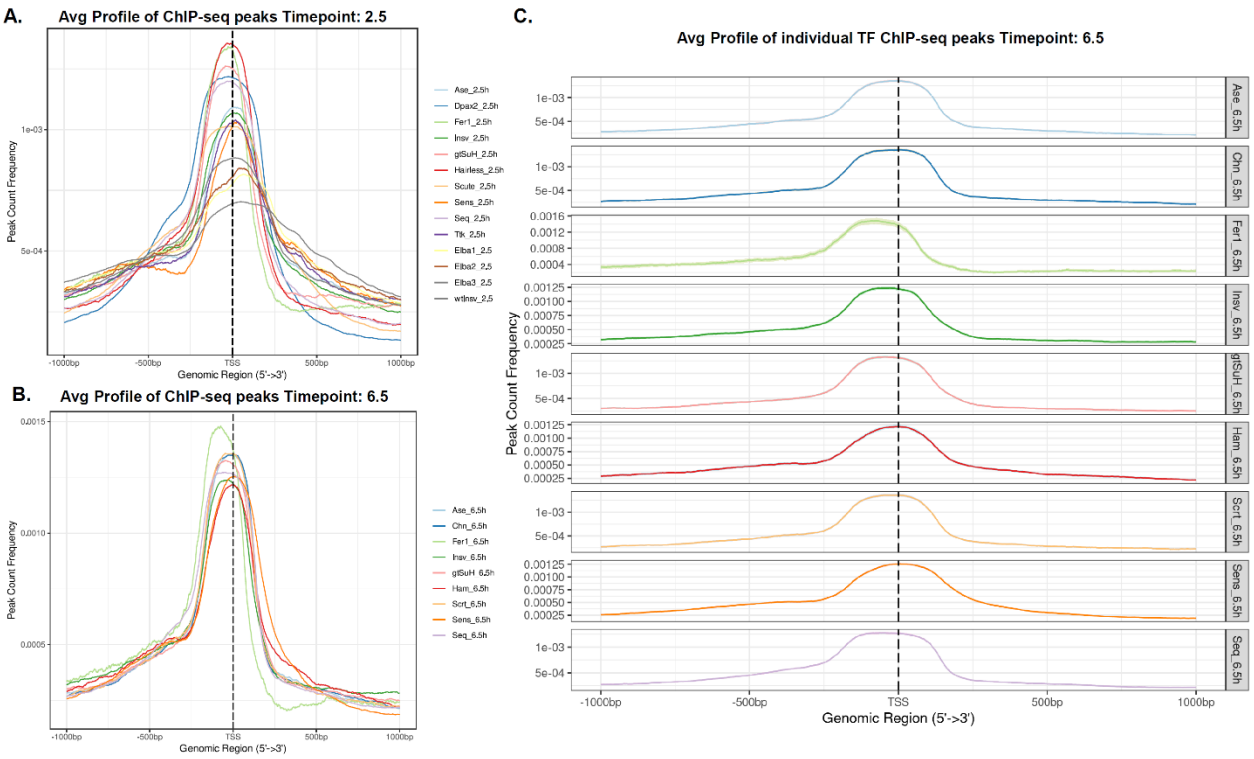


Figure 3.1.3: Average Profile plot for all the ChIP-seq peaks of samples at timepoint*- A. 2.5-6.5 hours AEL. B. 6.5-12 hours AEL. C. Average profile plot for individual TFs at 6.5 hours timepoint. The plot shows the average signal intensity of all the peaks centred at the TSSs. The x-axis represents the genomic distance relative to the TSSs, while the y-axis represents peak count frequency.*

The downstream analysis of ChIP-seq samples was performed to gain insight into the regulatory mechanisms of the neural development (see Methods). The distribution of ChIP-seq peaks across genomic features can provide insights into the functional roles of transcription factors. For instance, the peaks may be enriched in promoter regions, enhancers, introns, or exons, indicating the potential regulatory mechanisms involved. The feature distribution analysis can also reveal any biases or artifacts in the data. To analyse the feature distribution of ChIP-seq peaks, the genomic coordinates of the peaks were annotated with respect to the reference genome, and the number of peaks overlapping each genomic feature was counted. The distribution was then be visualized using a bar plot as shown in **Figure 3.1.4A**. It was found that the majority of peaks were located in promoter regions, followed by intronic and intergenic regions. However, Fer1 displayed a distinct pattern of peak distribution with nearly equal numbers of binding

peaks observed in promoter and distal intergenic regions. This observation indicates that Fer1 may play a unique regulatory role in controlling gene expression in both proximal and distal regulatory regions. The distinct feature of peak distribution in Fer1 highlights the importance of studying individual TFs and their binding patterns to better understand their regulatory functions.

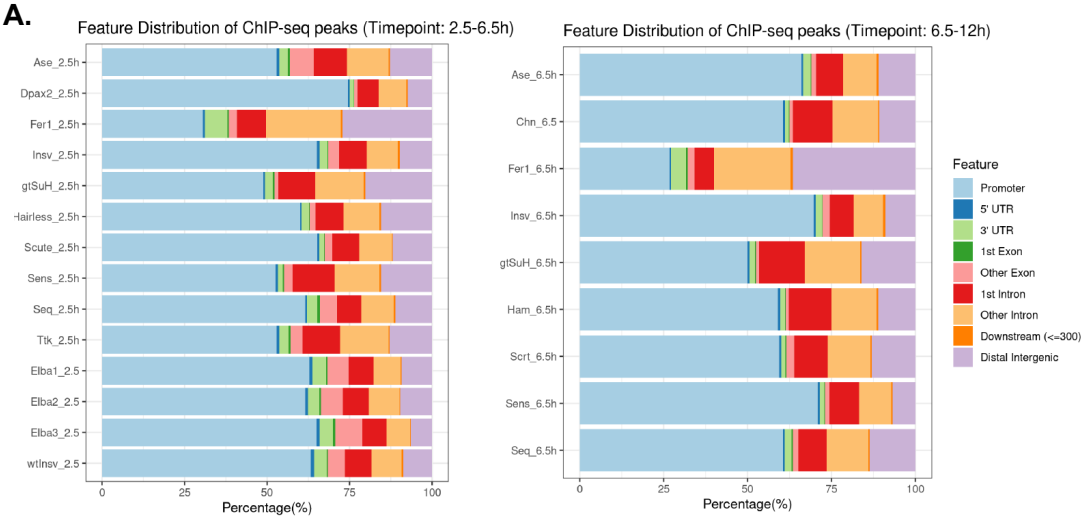
Next, the outcome of peak calling with GEM is displayed in **Figure 3.1.4B**, the resulting HTML output provided a detailed summary of the identified peaks, including the count of binding events, PWM motifs and their logo, and read and spatial distribution of the motifs. As shown in **Figure 3.1.3B**, we present the GEM peak calling results for the wtInsv sample, which includes the motif logo for the TF and the read distribution plot with all the detected motif occurrences. The identified motifs and their read distribution patterns provide important insights into the binding preferences and potential regulatory functions of the TF. This analysis provides insights into the binding preferences of the TFs and the potential target genes they regulate.

To gain further insight into the biological processes associated with the identified TFs, gene ontology analysis was performed (see methods). The hierarchical clustering of enriched terms was visualized using Jaccard's similarity index in the form of a tree plot, as shown in **Figure 3.1.4C**. This plot allows for the identification of biological processes that are enriched and their hierarchical relationships. The analysis revealed the enrichment of important pathways such as “neuron projection morphogenesis”, “axon development”, “cell morphogenesis involved in differentiation”, and “cell fate commitment”.

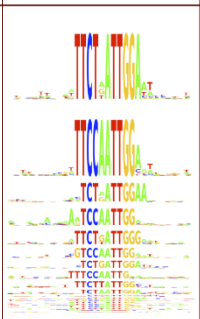
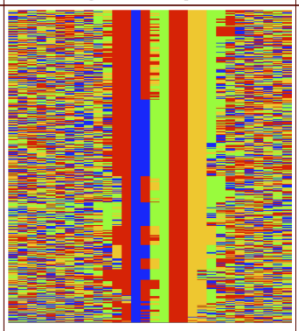
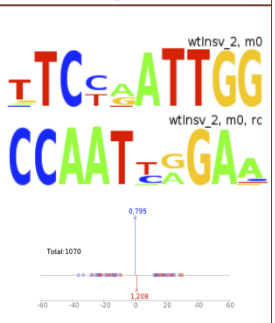
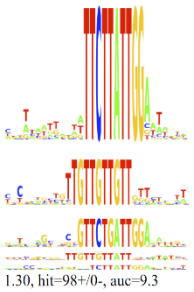
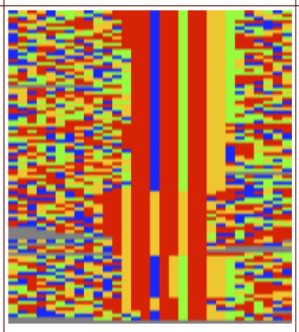
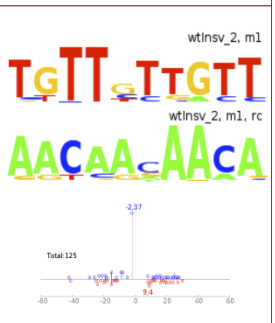
Lastly, **Figure 3.1.5** illustrates examples of TF co-binding/cooperativity using the genomic tracks for enriched peaks of correlated TFs at specific genomic locations. The genomic tracks shown in **Figure 3.1.5A** illustrate the co-binding of four TFs- Insv, Elba1, Elba2, and Elba3 at the gene *CG12811* on Chr 3R of the *Drosophila* genome. The peaks of these correlated TFs indicate their binding sites, and the genomic features highlight the co-binding of all four factors at the gene of interest. The co-binding of these factors suggests their cooperativity in regulating the expression of *CG12811*.

Similarly, in **Figure 3.1.5B**, three TFs- Ttk69, Scute, and Dpax2 show co-binding at gene *Sps2* on Chr 2L of the *Drosophila* genome. The enriched peaks of these TFs indicate their binding sites, and the genomic tracks highlight their co-binding at the gene of interest. The co-binding of these factors suggests their cooperativity in regulating the expression of *Sps2*. Some TFs frequently co-bound with each other and play a major role in TF cooperativity and gene regulation by either enhancing or suppressing the binding of other factors. These findings provide valuable insights into the mechanisms underlying gene regulation during neural development and target genes of these TFs.

A.



B

Motif ID	High scoring k-mers of the KSM motif	KSM motif	Aligned bound sequences	PWM and spatial distribution
m0	K-mer / Offset / Pos Hit / Neg Hit / HGP -TTCTNATTGGA / -7 / 127 / 1 / -40.6 -TTCCAATTGG / -7 / 110 / 0 / -35.9 --TCCAATTGGA / -6 / 105 / 1 / -33.7 --TCTNATTGGAA / -6 / 100 / 0 / -32.5 --TCNAATTGGANT / -6 / 99 / 1 / -31.6 --TCTGATTGGA / -6 / 77 / 1 / -24.7 --TCTAATTGGA / -6 / 75 / 0 / -24.4 -TTCNNATTGGANT / -7 / 72 / 0 / -23.5 --TCNAATTGGANA / -6 / 68 / 0 / -22.6 ATTCTNATTGG / -8 / 66 / 1 / -21.3	 1.10, hit=455+/4-, auc=24.5		 8.53/16.86, hit=559+/53-, auc=28.2
m1	K-mer / Offset / Pos Hit / Neg Hit / HGP --TTCTTATTGG / -7 / 56 / 0 / -18.2 ---TCTTATTGGA / -6 / 38 / 0 / -12.7 TGTTGTTGTT / -9 / 21 / 0 / -7.3 -TTTCTTATTG / -8 / 21 / 1 / -6.6 -GTTCTGATTG / -8 / 13 / 0 / -5.1	 1.30, hit=98+/0-, auc=9.3		 8.88/15.35, hit=290+/149-, auc=9.8

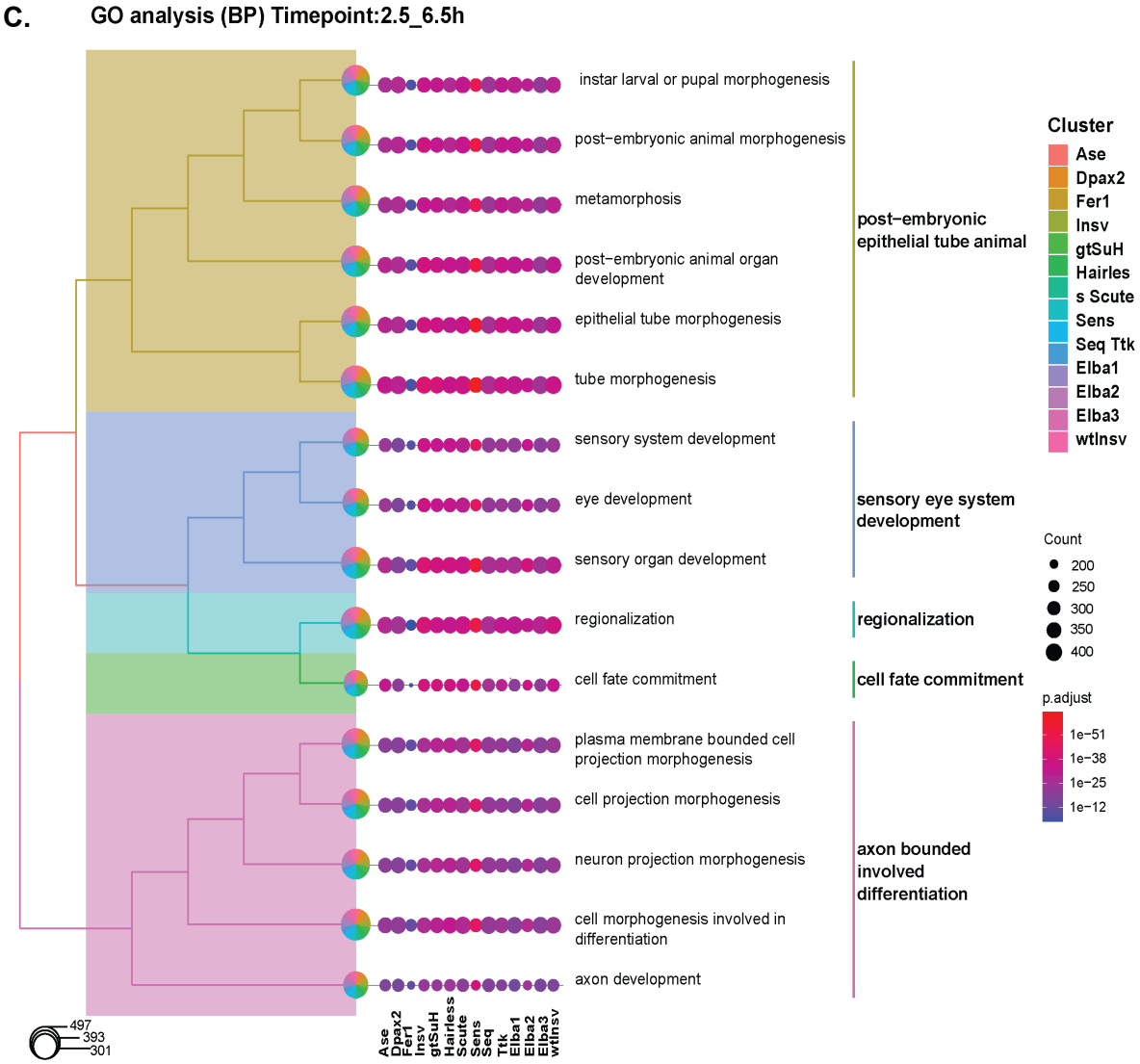


Figure 3.1.4: Downstream analysis of ChIP-seq samples. *A. Feature distribution plot for all the ChIP-seq peaks of samples at timepoint-2.5-6.5 hours and 6.5-12 hours. B.) GEM peak calling outcome displaying Motifs, high scoring K-mers of each motif id, heatmap of aligned sequences at TSS and position weight matrix with spatial distribution. C. Treeplot of Gene Ontology Analysis of TFs. The plot shows the hierarchical clustering of enriched terms by using Jaccard's similarity index.*

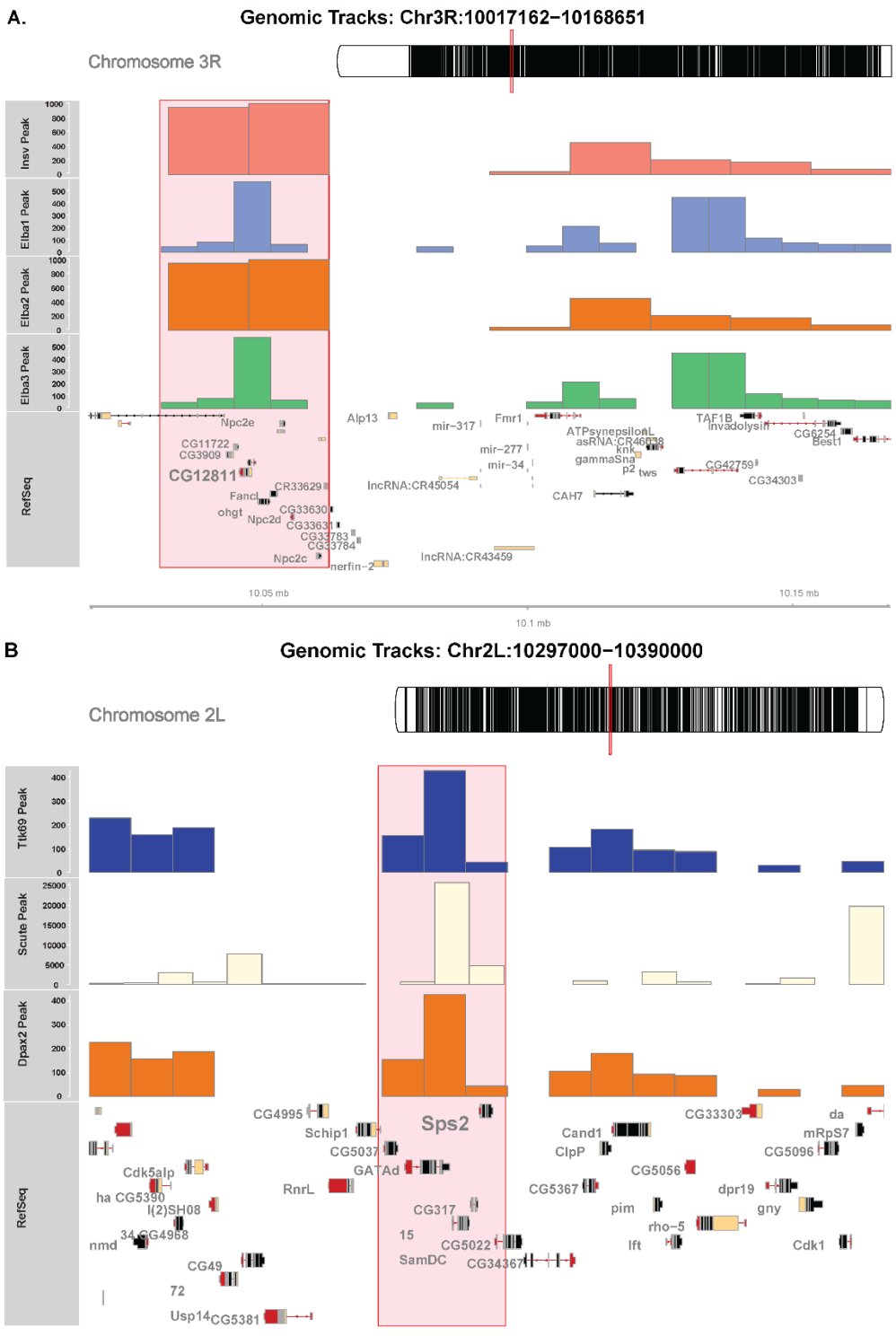


Figure 3.1.5: Examples of TF co-binding/cooperativity displayed by genomic tracks of specific locations. *A. Visualization of co-binding showing genomic tracks for enriched peaks of four correlated TFs, Insv, Elba1, Elba2 and Elba3 along with the genomic features highlighting co-binding of all 4-factors at gene CG12811 on Chromosome 3R. B. Co-binding of TFs Ttk69, Dpax2 and Scute at Sps2 gene on Chromosome 2L.*

3.1.2 Bulk ATAC-seq data analysis reveals TF binding correlates with accessible chromatin and shows TF motif occupancy at TSS.

The quality control and downstream analysis of bulk-ATAC-seq data were performed to ensure the reliability and accuracy of the results. Nucleosomes are the basic units of chromatin, consisting of DNA wrapped around a core of histone proteins. Nucleosome-free regions are regions of DNA that lack nucleosomes and are more accessible to TFs and other proteins involved in gene regulation. Mono-nucleosomes, on the other hand, are regions where a single nucleosome occupies the DNA. These regions are less accessible to transcription factors and other proteins involved in gene regulation compared to nucleosome free regions. In ATAC-seq, the Tn5 transposase preferentially inserts into nucleosome free regions, resulting in a higher frequency of sequencing reads in these regions. Therefore, the presence of nucleosome free regions and mono-nucleosomes can be inferred from the fragment length distribution and TSS enrichment plots of ATAC-seq data. The fragment length distribution of the ATAC-seq samples were analysed, and three distinctive peaks were observed, indicating good quality of the ATAC-seq data. An example of bulk ATAC-seq analysis of wtInsv TF is shown in **Figure 3.1.6**. In the insert size histogram **(Figure 3.1.6A)**, the first peak (~50bp) corresponds to Tn5 transposase insertion into nucleosome-free regions, the second peak (~200bp) indicates insertion of Tn5 around a single nucleosome, and a small third nucleosome peak was seen at ~400bp.

To further assess the quality of the data, the TSS enrichment score was calculated as the ratio between the aggregate distribution of reads centred on TSSs and those flanking the corresponding TSSs. The TSS enrichment plot showed that nucleosome-free fragments were enriched at TSSs, whereas mono-nucleosome fragments were depleted at TSSs but enriched at flanking regions **(Figure 3.1.6B, C)**.

To infer the footprints of the transcription factor, the pattern/occurrence of Tn5 transposase cutting sites around the Insv-binding sites were analysed **(Figure 3.1.6D)**. This analysis revealed distinct regions that were protected from Tn5 transposase

fragmentation, suggesting the presence of Insv-bound regions that resist cleavage by the enzyme. These regions are referred to as footprints and are indicative of protein-DNA interactions. The occurrence of these footprints around the Insv-binding sites suggests that the protein is bound to specific DNA sequences in the genome, which are protected from cleavage. This observation is consistent with the known function of Insv as an insulator protein that plays a role in regulating gene expression by organizing chromatin structure and blocking the spread of heterochromatin. Hence, binding of the transcription factor Insv was significantly associated with accessible chromatin.

The open and closed chromatin states were visualized through a heatmap displaying the signal distribution around TSS, as detailed in the methods section (**Figure 3.1.6E**). The open chromatin regions are associated with active transcription and the closed chromatin regions with transcriptional repression. Specifically, we found that the majority of TF binding sites were located within open chromatin regions, as demonstrated by the heatmap of signal distribution around TSS **(Figure 3.1.6E**).

Our results show that transcription factors often work together, enhancing the precision and control of gene regulation. This cooperation between TFs suggests they interact to fine-tune the expression of key developmental genes.

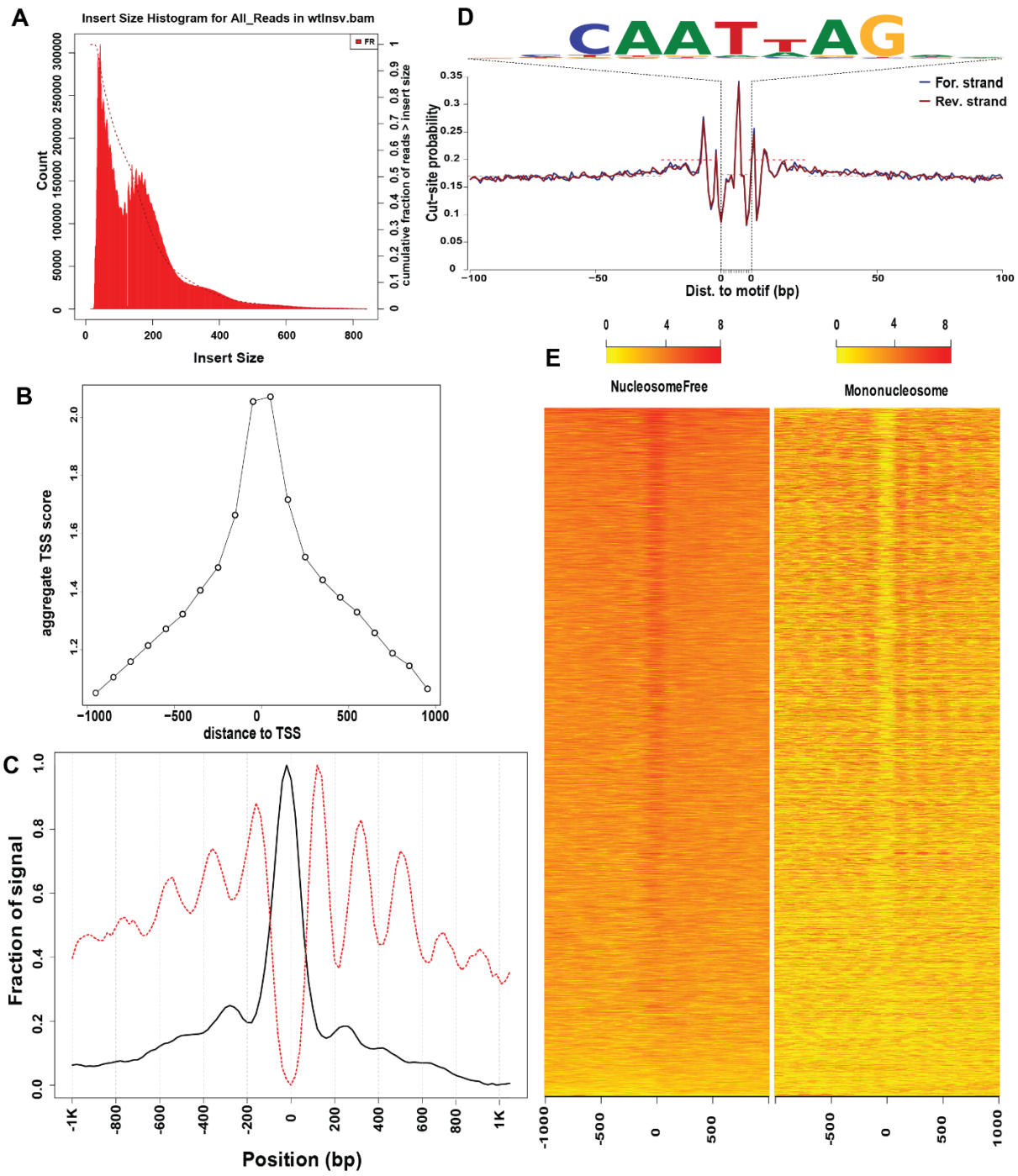


Figure 3.1.6: Bulk ATAC-seq data shows TF motif occupancy at Transcription start site. *A. Histogram of Insert Sizes: This panel shows a histogram of the distribution of insert sizes for the ATAC-seq reads. The insert size is the distance between the two ends of a DNA fragment that has been sequenced. This information is useful for quality control and for estimating the fragment length distribution. B. TSS Enrichment Score: This panel shows the enrichment of ATAC-seq signal around transcription start sites (TSS). The TSS enrichment score is calculated by comparing the number of ATAC-seq reads that fall within a certain distance from the TSS to the number of reads that fall outside of that distance. A high TSS enrichment score indicates that the chromatin around the TSS is more accessible. C. Coverage Curve of*

Nucleosome Position: This panel shows the ATAC-seq signal density as a function of the distance from the centre of nucleosomes. It shows nucleosome free fragments are enriched at TSS (Black line) whereas mono-nucleosome fragments (red dotted line) depleted at TSS but enriched at flanking regions. D. Footprint analysis. Genomic regions bound by transcription factors/insulators are locally protected from Tn5 transposase fragmentation. The pattern/occurrence of Tn5 transposase cutting site around Transcription factor binding site of Insv are shown to infer the footprints of TF. E. Heatmap of nucleosome position. The heatmap shows signal distribution around TSS to clearly visualize the open and closed chromatin states

3.2 Accurate Prediction of TF binding sites at bulk level

3.2.1 Schematic representation of the prediction model

In this section, we describe an overview of our prediction model as depicted in **Figure 3.2.1.** We primarily focused on the “bulk model” development, a core part of our prediction approach. This model is capable of forecasting both individual and combinatorial TF binding sites based on DNA sequence on promoter regions and ATAC-seq coverage signals. These inputs are essential as they provide crucial information about the chromatin accessibility landscape, allowing the model to discern the accessibility of TF binding sites. The utilization of the Trans-UNet architecture enhances the model's ability to learn complex interactions between TF motifs and predict TF binding sites with higher accuracy (See methods for Trans-UNet architecture as described in 2.2.1). Within this architecture, we have developed two sub-models, M1 and M2, each serving unique purposes.

M1 is a classification model specifically tasked with predicting TF binding sites, providing valuable insights into the combinatorial binding patterns of TFs, which are crucial in gene regulation. We used TF ChIP-seq peaks as true labels, representing known TF binding sites, for the model training. To validate our model and gain biological insights, we used saliency maps to visualize predictive motifs. These maps, commonly derived from gradient backpropagation, highlight influential features in the input data.

M2 is a regression model that focuses on predicting gene expression levels. We utilized both PRO-seq data and RNA-seq data as our ground truth. By comparing the predictions from our regression model with the expression values from PRO-seq and RNA-seq data, we evaluated the model's ability to accurately predict downstream gene expression levels from promoter sequences and chromatin accessibility.

In the second part of the workflow, we will apply our bulk model to predict TF binding sites by integrating sc-ATAC-seq data. This allowed us to explore cell-type-specific regulatory events at the single-cell resolution. From the sc-ATAC-seq data, we obtained cell-type-specific peaks, which represented regions of open chromatin and

potential TF binding sites unique to each cell type. These cell-type-specific peaks served as critical inputs for the bulk model, enabling us to predict TF binding patterns at the single-cell level.

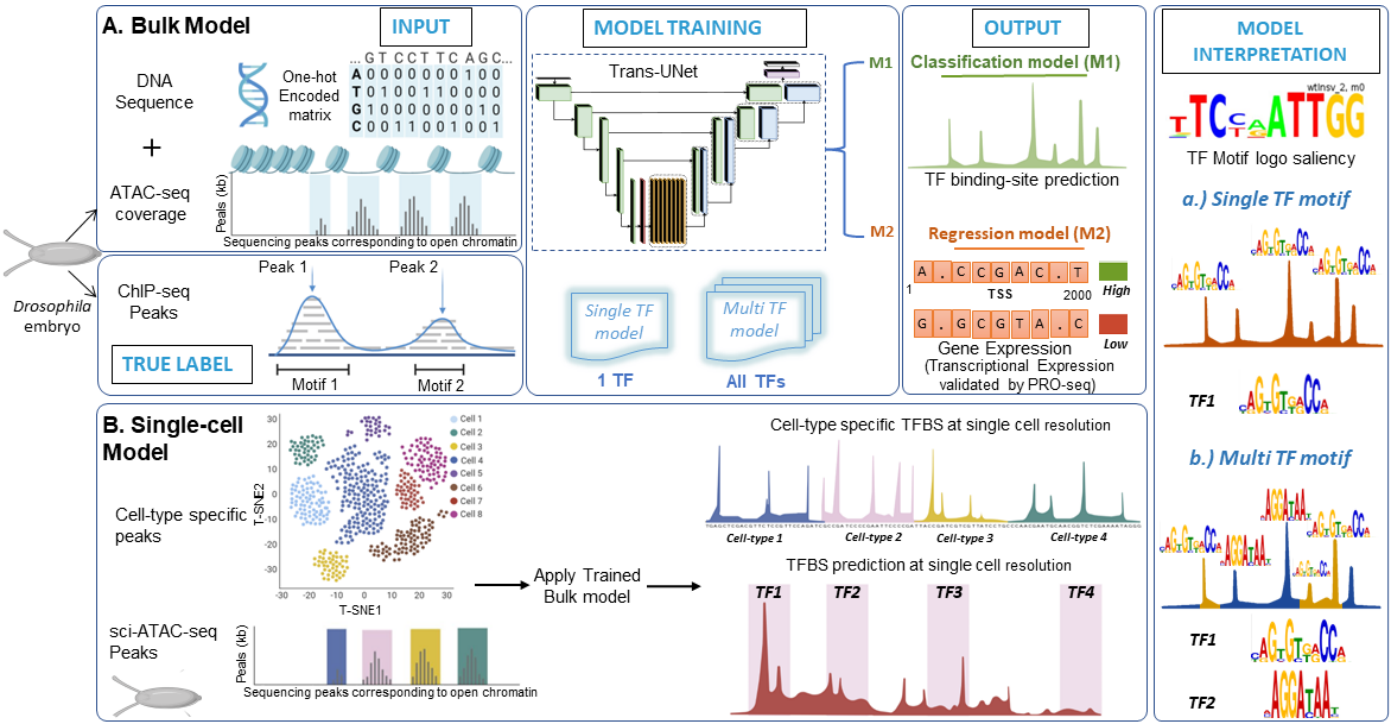


Figure 3.2.1: Schematic overview of project design*. The project is broadly divided into two parts* ***A.) Bulk model*** *which takes Dm6 reference DNA sequence represented as one hot encoded matrix and ATAC-seq coverage normalised to 1X resolution as input to the model. The ChIP-seq peaks denoting binding sites of TF were used as True labels. For model training, Trans-UNet architecture was used to build two models- Classification model or M1 to predict TFBS and Regression model or M2 to predict gene expression. For the M1 model, two approaches were implemented, first to predict binding site of single TF and second to predict binding sites of multi-TFs. The model was interpreted using Motif logo saliency.* ***B) Single-cell model,*** *here single cell ATAC-seq data was pre-processed to extract cell-type specific peaks for input into the bulk model. As a result, cell-type specific binding sites were predicted at single nucleotide resolution.*

3.2.2 Model Performance Evaluation

To ensure comprehensive coverage of the TF binding landscape, we employed a two-phase strategy: training the model on individual TFs to create a single TF model, and subsequently training the model on all the TFs together to create a multi-TF model. In the initial phase, we began by training the model on individual TFs (single TF model) that allowed us to fine-tune the model's predictions specifically for each TF's binding sites. By focusing on individual TFs, we aimed to capture the unique binding preferences and patterns associated with each TF. This step enabled the model to learn intricate relationships between the DNA sequences and the binding events of each TF, leading to refined predictions that were tailored to the specific characteristics of each TF's binding behaviour. After achieving satisfactory performance with the single TF models, we proceeded to the second phase of our training strategy. In this phase, we adopted a holistic approach by training the model on all the TFs collectively, referred to as the multi-TF model. This encompassing model considered the diverse binding behaviours and preferences of all TFs simultaneously. By doing so, we aimed to capture potential interactions and dependencies between TFs, contributing to a more comprehensive understanding of the regulatory landscape.

In our study, we used several evaluation metrics to assess the performance of our models. These metrics are crucial in determining the accuracy and effectiveness of machine learning models. One such metric is the area under the curve (AUC), which is commonly used to evaluate the performance of binary classification models. The AUC measures the ability of the model to distinguish between positive and negative samples, with a value of 1 indicating perfect classification and a value of 0.5 indicating random guessing. In our study, we calculated the AUC for each TF separately and compared the performance of the single-TF model and multi-TF model at two different timepoints

The results of our study demonstrate the effectiveness of our model in accurately predicting TF binding sites in both the phases. Among the TFs, wtElba3 and Scute showed the highest performance with an AUC of 0.997 in the single-TF model and 0.999 in the multi-TF model. Similarly, wtElba1 and Dpax2 also exhibited high accuracy with AUROC values of 0.99 and 0.986 in the single-TF model, respectively. On the other hand,

the TF with the lowest score was Fer1 with an AUC of 0.877 for the single-TF model and 0.975 for the multi-TF model **(Figure 3.2.2, 3.2.3)**.

Another important evaluation metric is the precision-recall curve (PRC), which provides a visual representation of the trade-off between precision (positive predictive value) and recall (sensitivity) at different classification thresholds. The area under the PRC (AUPRC/PR_AUC) is a useful metric for imbalanced datasets, where the number of positive samples is much smaller than the number of negative samples. In our study, we also calculated the PR_AUC for each TF separately and compared the performance of the single-TF model and multi-TF model. **Figure 3.2.2** reveals that most TFs achieved higher PR_AUC scores in the multi-TF model, except for Fer1, which had lower scores of 0.21 for the multi-TF model and 0.53 for the single-TF model. For the remaining TFs, the PR_AUC ranged from 0.79 to 0.98, with the highest values obtained for Dpax2, Scute, wtElba3, wtElba1, and wtInsv. These results were consistent with the AUC scores mentioned earlier, indicating that the multi-TF model improved the prediction accuracy of the TF binding sites compared to the single-TF model.

We also used the F1 score, which is a harmonic mean of precision and recall, to evaluate the overall performance of our models. The F1 score provides a single measure of the model's ability to correctly classify positive and negative samples. The F1 score ranges from 0 to 1, where a score of 1 indicates perfect precision and recall, and a score of 0 indicates poor performance. In our study, the overall F1 score for most of the TFs ranged between 0.72 to 0.95, which was consistent with the AUC and PR_AUC scores for the multi-TF models. However, we observed that the F1 score for Fer1 was significantly low at 0.55 for the single-TF model and 0.08 for the multi-TF model. This poor performance may be attributed to the fact that majority of Fer1 binding sites were located in the intergenic region rather than in promoters, leading to a lack of specific sequence features that can aid in accurate prediction. As shown in Figure **3.1.4A**, approximately only 28% and 25% of the total peaks were in the promoter region for 2.5-6.5- and 6.5-12-hours' time points respectively.

We also observed that the F1-score of the single-TF model for Sens was 0, whereas the multi-TF model achieved an F1-score of 0.82. This improvement in

performance may be attributed to the fact that the multi-TF model can capture the combinatorial effects of multiple TFs, leading to improved prediction accuracy. Notably, thirteen out of fourteen TFs displayed substantial increases in AUC, PR_AUC, and F1 scores in the multi-TF model, with Fer1 serving as an exception—exhibiting increased AUC but diminished PR_AUC and F1 scores. These collective findings suggest that multi-TF model significantly enhanced the predictive strength of our approach.

In conclusion, our study utilized multiple evaluation metrics to assess the performance of our model in predicting TF binding sites. The substantial AUC, PR_AUC, and F1-score values for most of the TFs underscore our model's the effectiveness and reliability. Furthermore, the comparison between single-TF and multi-TF models demonstrated the advantages of using a multi-TF model for predicting binding sites accurately. These results suggest that our multi-TF model can be a powerful tool for predicting TF combinatorial binding sites and can have important applications in understanding gene regulation mechanisms. We also observed some TFs, such as Fer1 and Sens, underperformed. This could be due to the limited availability of training data for these TFs, as well as their unique binding characteristics.

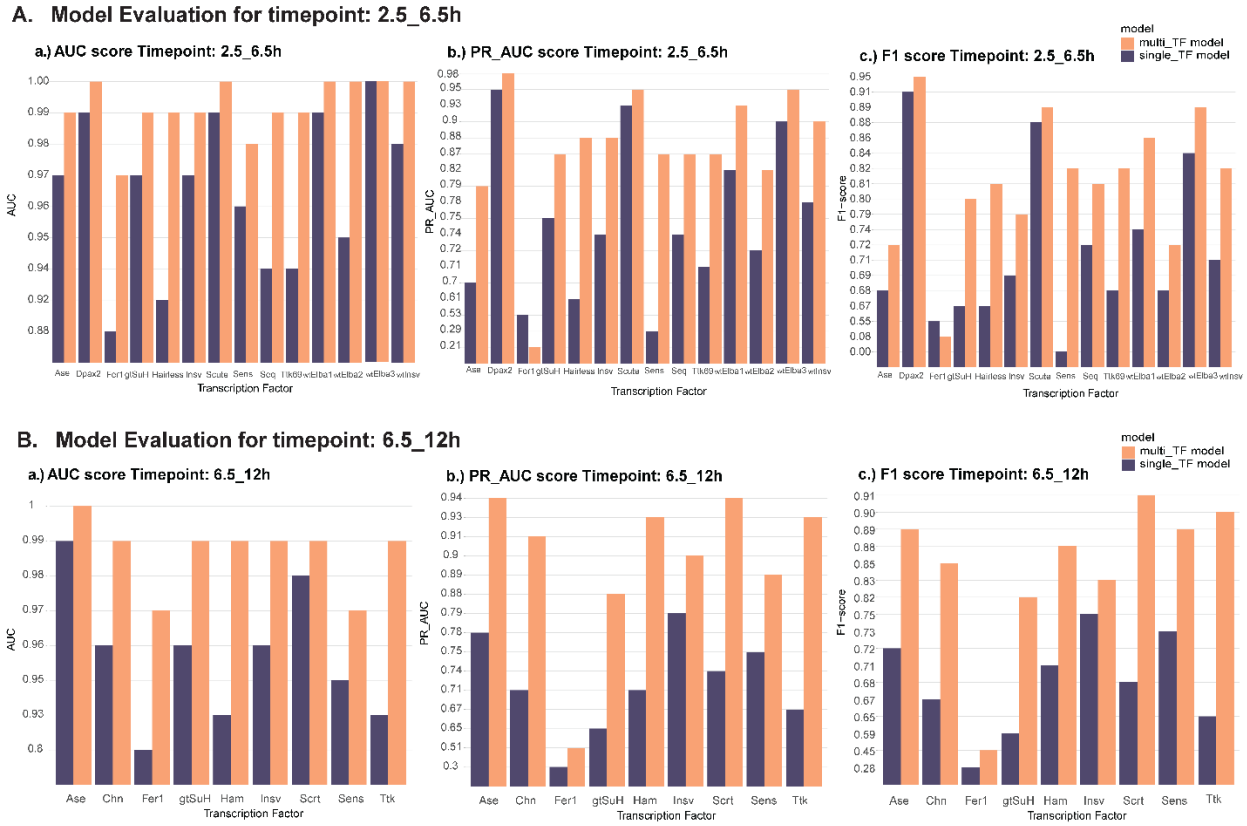


Figure 3.2.2: Evaluation metrics of the classification model (M1). *A. Model Evaluation for Timepoint 2.5-6.5 Hours Using AUC, PR_AUC, and F1-Score: This panel shows the performance of the classification model (M1) in predicting transcription factor binding activity at timepoint 2.5-6.5 hours AEL. The evaluation metrics used are the area under the receiver operating characteristic curve (AUC), the area under the precision-recall curve (PR_AUC), and the F1-score. The two histograms depict the evaluation scores for each transcription factor using the multi-TF model (in orange) and the single TF model (in dark blue). A higher score indicates better model performance. The AUC, PR_AUC and F1-score ranges from 0 to 1, with a value of 0.5 indicating that the model is performing no better than random chance, and a value of 1 indicating perfect performance, where all positive samples are correctly identified, and all negative samples are correctly classified as negative. The multi-TF model predicts the activity of multiple transcription factors simultaneously, while the single TF model predicts the activity of each transcription factor independently. B. Model Evaluation for Timepoint 6.5-12 Hours using the Same Metrics as A: This panel shows the performance of the classification model (M1) in predicting transcription factor binding at timepoint 6.5-12 hours AEL using the same evaluation metrics as panel A.*

3.2.3 Model performance of single vs muti TF model

To evaluate the performance of the single vs multi-TF model, we employed a training set of 14 neuronal TFs. To observe the difference between the performance of both models, we plotted AUROC curve **(Figure 3.2.3)** which is a graphical representation of the performance of a classification model across different classification thresholds. It plots the true positive rate against the false positive rate for various classification thresholds. The closer the curve is to the top left corner of the graph, the better the performance of the model. The area under this curve represents the probability that the model can distinguish between positive and negative samples. For the multi-TF model, the AUROC values for all TFs when predicted together were consistently high, suggesting that the model was able to accurately predict binding sites for multiple TFs simultaneously. Additionally, the performance of the multi-TF model was also consistent across the two timepoints, indicating that the model was robust and reliable. This suggests that the multi-TF model was able to capture more complex interactions between TFs and DNA sequences, leading to improved predictions of TF binding sites. The single-TF model, on the other hand, may have been limited by its inability to capture such interactions, resulting in lower prediction accuracy. However, it is worth noting that the performance of the single-TF model was still relatively high, with AUROC values ranging from 0.877 to 0.999 **(Figure 3.2.3)** and PR_AUC values of majority of TFs ranging from 0.7 to 0.95. This suggests that the single-TF model can still be useful in certain scenarios, such as when only one specific TF is of interest or when computational resources are limited. Overall, our results highlight the strengths and weaknesses of each model and suggest that a multi-TF model can be a powerful tool for accurately predicting TF binding sites, while a single-TF model can be a viable alternative in certain situations.

There could be several reasons why the multi-TF model outperformed the single-TF model. One possible reason is that the multi-TF model benefits from the shared information among different TFs, leading to a better overall prediction. This is because different TFs may have similar binding patterns and regulatory functions, and incorporating information from multiple TFs can provide a more comprehensive understanding of the regulatory landscape. Additionally, the multi-TF model may also

have a better ability to distinguish true binding events from background noise or non-specific binding events, as it considers the overall binding patterns across different TFs. Finally, the increased amount of training data available for the multi-TF model may have also contributed to its improved performance.

Single-TF Model:

Strengths: The single-TF model can be useful for studying the specific binding properties of individual transcription factors. This model can help identify the exact binding sites of a single TF, allowing for more focused analysis of regulatory mechanisms.

Weaknesses: The single-TF model can be limited in its ability to accurately predict binding sites of a TF, especially if the TF is part of a complex regulatory network. Additionally, if a TF has a low number of available training samples, the model's performance can suffer.

Multi-TF Model:

Strengths: The multi-TF model can capture the interactions and dependencies between different transcription factors, providing a more comprehensive understanding of regulatory networks. This model can also better handle noisy and incomplete data, as it can use information from multiple TFs to make more accurate predictions.

Weaknesses: The multi-TF model may be less interpretable than the single-TF model, as it can be difficult to tease apart the contributions of individual TFs to the final prediction. Additionally, the model's performance can suffer if there are limited available training samples for some of the TFs.

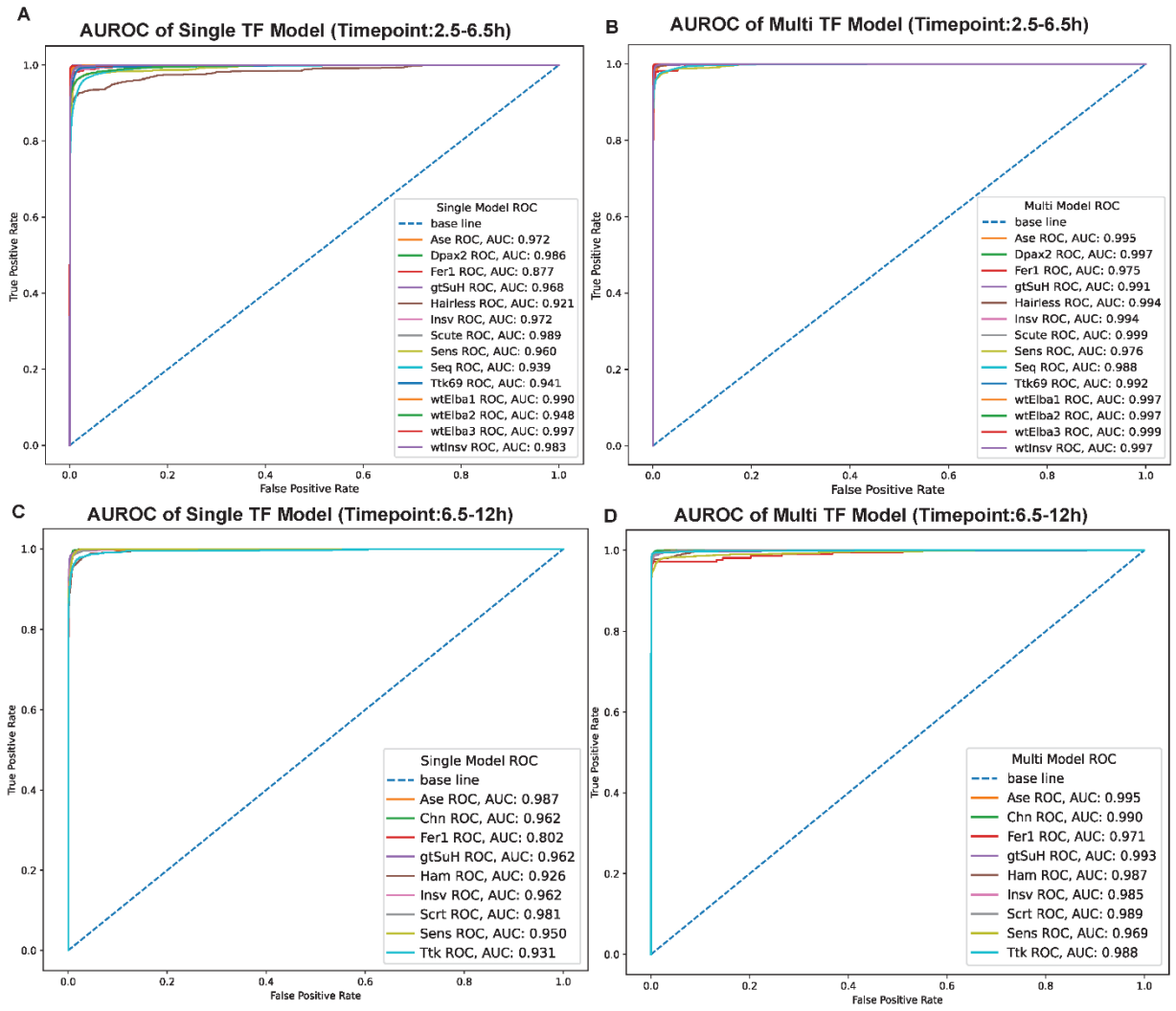


Figure 3.2.3: Area under the receiver operating characteristic curve (AUROC) for single and multi-TF models at both timepoints. *A. shows the performance of the classification model (M1) at timepoint 2.5-6.5 hours AEL in terms of the area under the receiver operating characteristic curve (AUROC) for each TF individually. The X-axis represents the false positive rate, while the Y-axis represents the true positive rate. The diagonal blue dotted line represents the baseline performance, which is achieved by random guessing. Each TF is depicted by a different colour, and the values of AUC are closer to 0.9, indicating excellent learning by the model. B. Similar to A but for multi-TF model or AUROC of all TFs when predicted together. C. Similar to A but for Timepoint 6.5-12 Hours. D. Similar to B but for timepoint 6.5-12 hours AEL.*

3.2.4 Benchmarking model performance

Next, to establish a rigorous benchmark for our model's performance, we conducted a comprehensive comparison of evaluation metrics across three distinct DL architectures:

1. TransUNet: Our proposed architecture, TransUNet, which combines the strengths of transformer-based models with UNet. This hybrid approach aims to capture both local and global contextual information, crucial for accurate prediction of complex regulatory interactions.
2. UNet- An adaptation of a recently published UNet-based approach by Li and Guan (2021), which has a Convolutional Neural Network (CNN) implementation compared to transformer in our model.
3. SE-UNet- Implementation of a UNet architecture enhanced with Squeeze-and-Excitation (SE) blocks. These blocks enable the network to automatically emphasize informative features while suppressing less relevant ones, potentially leading to improved representation learning.

Table 3 presents a comprehensive performance comparison of three deep learning architectures applied to the prediction of TF binding sites. The comparison is conducted for various TFs using multiple evaluation metrics. The metrics include Area Under the Curve (AUC), Precision-Recall Area Under the Curve (PR AUC), F1-score, Dice Coefficient, Matthews Correlation Coefficient (MCC), and loss (outlined in detail in the Methods section).

Table 3 reveals that across different TFs, TransUNet generally shows competitive or improved performance compared to both SE-UNet and UNet. For gtSuH, the PR AUC increases from 0.58 (UNet) to 0.71 (TransUNet), indicating a notable improvement in the model's ability to balance precision and recall. Moreover, the AUC rises from 0.88 UNet to 0.98 TransUNet, signifying enhanced overall performance in distinguishing between positive and negative instances. These improvements suggest that TransUNet is better at identifying true positive binding events while minimizing false positives. The results for Ttk69 show a relatively minor difference in PR AUC (UNet - 0.580, TransUNet - 0.561) and F1-score (UNet - 0.545, TransUNet - 0.601) between UNet and TransUNet. However, TransUNet demonstrates a substantial improvement in AUC (UNet - 0.882, TransUNet - 0.967), highlighting its effectiveness in classifying binding sites.

This indicates that the TransUNet architecture, with its integration of transformer-based components, has an advantage in capturing complex relationships within the data.

The results vary across TFs, with some TFs exhibiting notably higher AUC and PR AUC scores, indicating effective binding prediction. This variation suggests that the regulatory behaviour of TFs might differ, with some being more predictable based on sequence and accessibility data. Or they might possess large number of binding sites for the training.

The MCC score provides insights into the robustness of the models, accounting for true positives, true negatives, false positives, and false negatives. Higher MCC scores suggest better model performance in distinguishing between different classes, indicating robustness in binding site prediction. The architectures' loss values are an indicator of how well the models converge during training. Lower loss values generally imply that the models are effectively learning from the data. While TransUNet shows a higher loss in some cases, this is not necessarily indicative of poorer performance as other metrics need to be considered together. The variation in performance across different TFs highlights the complex nature of regulatory interactions and the need for tailored approaches to predict their binding patterns effectively.

Key Observations:

1. TransUNet consistently exhibits competitive performance across all evaluated metrics, with strong AUC and PR AUC scores indicating effective classification capabilities.
2. SE-UNet showcases competitive results in terms of AUC and PR AUC, though it occasionally lags behind TransUNet and UNet in other metrics such as F1 score and Dice Coef.
3. UNet demonstrates consistent results, often being intermediate between TransUNet and SE-UNet in most metrics.

Hence, the TransUNet architecture stands out as a powerful choice, consistently offering a balance between classification performance and other evaluation metrics. While SE-UNet performs well in certain areas, it may be less robust in capturing certain aspects of the predictions. The traditional UNet architecture remains a reliable option but might exhibit slightly less competitive performance in comparison to TransUNet and SE-UNet in specific metrics. These benchmarking results provide valuable insights into the

strengths and weaknesses of the different architectures, aiding in selecting an appropriate architecture for addressing the specific research problem at hand.

Table 3: Performance Comparison of TransUNet, SE-UNet, and UNet Architectures

TF	Model	Loss	AUC	PR_AUC	F1-score	Dice_coef	MCC	Training Time
Ase	trans_unet	0.002645	0.975663	0.689926	0.645413	0.568928	0.647017	936.3531954
Ase	unet	0.003009	0.965169	0.660063	0.611507	0.562571	0.61734	514.6890042
Ase	se_unet	0.002975	0.977289	0.651255	0.60364	0.562019	0.609703	593.4125242
Dpax2	trans_unet	0.001303	0.983446	0.945327	0.910582	0.898781	0.911061	1120.821351
Dpax2	unet	0.001008	0.991062	0.966495	0.928042	0.909853	0.928602	584.3052735
Dpax2	se_unet	0.001022	0.993923	0.967705	0.945337	0.922964	0.945631	596.7592955
Fer1	trans_unet	0.000214	0.810482	0.426238	0.332011	0.524509	0.335585	1319.189374
Fer1	unet	0.00026	0.795095	0.056448	0	0.17644	0	475.9287052
Fer1	se_unet	0.000206	0.855737	0.113755	0	0.195073	0	525.2513454
gtSuH	trans_unet	0.000449	0.980308	0.712724	0.548247	0.547	0.568259	873.5219321
gtSuH	unet	0.00085	0.882279	0.58069	0.545122	0.584594	0.569281	667.5872607
gtSuH	se_unet	0.00048	0.959275	0.702132	0.556813	0.537535	0.566944	559.9713757
Hairless	trans_unet	0.00095	0.909393	0.652263	0.664092	0.60611	0.682183	1159.826265
Hairless	unet	0.000767	0.915369	0.714778	0.6473	0.635554	0.658543	643.5218606
Hairless	se_unet	0.000889	0.906979	0.613967	0.596702	0.529865	0.614598	536.5887361
Insv	trans_unet	0.001594	0.969041	0.717354	0.678717	0.61079	0.685788	1002.319673
Insv	unet	0.001643	0.961843	0.675529	0.571506	0.522292	0.600208	496.7143688
Insv	se_unet	0.001264	0.991107	0.710353	0.620454	0.534788	0.629721	588.2787755
Scute	trans_unet	0.00117	0.992071	0.925538	0.868244	0.84725	0.869274	983.5612061
Scute	unet	0.001153	0.987032	0.925965	0.846969	0.804754	0.851365	459.2233315
Scute	se_unet	0.001118	0.991108	0.920409	0.868296	0.829747	0.869854	621.5647259
Sens	trans_unet	0.00081	0.975266	0.219361	0	0.147607	0	819.563729
Sens	unet	0.00085	0.89243	0.527553	0.45625	0.428856	0.460515	539.7752714
Sens	se_unet	0.000796	0.920659	0.221821	0.072222	0.195597	0.097491	540.3506052
Seq	trans_unet	0.004493	0.938168	0.703903	0.665701	0.640729	0.671742	1281.015062
Seq	unet	0.00459	0.951186	0.627222	0.601689	0.546671	0.607782	563.2387221
Seq	se_unet	0.003856	0.967803	0.673203	0.623815	0.534598	0.626937	513.1339405
Ttk69	trans_unet	0.000381	0.970131	0.794581	0.754281	0.695107	0.758699	1077.272241
Ttk69	unet	0.000343	0.960691	0.795005	0.742679	0.688067	0.758243	582.9376276
Ttk69	se_unet	0.000494	0.966939	0.561094	0.601438	0.534886	0.606009	561.5854514
wtElba1	trans_unet	0.000953	0.982893	0.786207	0.703148	0.60547	0.714505	833.3540401
wtElba1	unet	0.000923	0.993404	0.791191	0.696262	0.635044	0.703574	474.9772944
wtElba1	se_unet	0.000977	0.988419	0.834007	0.752359	0.720486	0.757772	559.9465957
wtElba2	trans_unet	0.000882	0.964489	0.783904	0.737101	0.714682	0.749299	1281.412289
wtElba2	unet	0.000811	0.969028	0.75547	0.660276	0.599807	0.688096	515.3249285
wtElba2	se_unet	0.000682	0.959882	0.768173	0.732198	0.645766	0.742052	748.6498635
wtElba3	trans_unet	0.000538	0.996889	0.937421	0.872415	0.810732	0.875152	960.111058
wtElba3	unet	0.000878	0.985047	0.887316	0.813539	0.782052	0.819123	489.4569919
wtElba3	se_unet	0.000785	0.987702	0.905945	0.833244	0.797024	0.835661	573.6876535
wtInsv	trans_unet	0.001641	0.988993	0.806668	0.736448	0.672214	0.736718	900.9592083
wtInsv	unet	0.002199	0.954767	0.762405	0.688258	0.6473	0.696096	566.0338254
wtInsv	se_unet	0.001634	0.984079	0.83061	0.715666	0.676354	0.731037	564.3196545

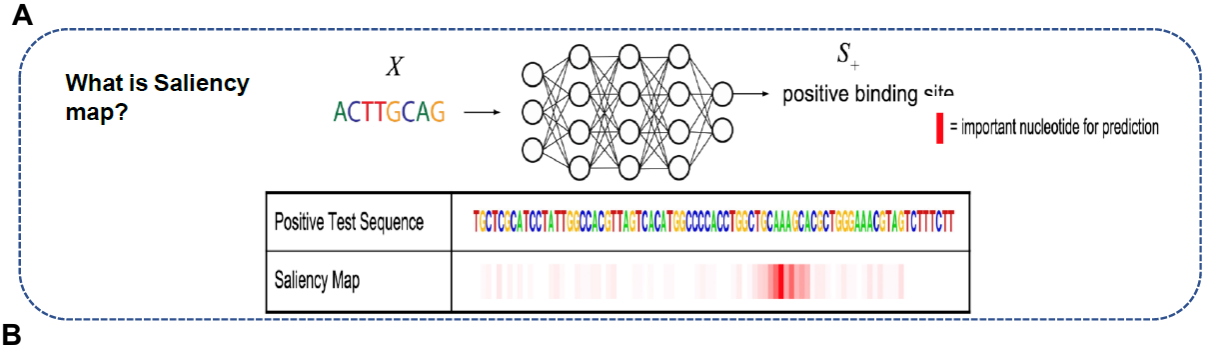
3.2.5 Model Interpretation

To gain biological insights, we employed saliency maps, TF motif logo saliency, to interpret and validate our TF model's predictions. Saliency analysis is a deep learning-based method that highlights important regions in an input sequence that contribute to a model's output. The saliency maps generated from our model's predictions provided an intuitive visualization of the regions in the genomic sequence that were important for predicting TF binding sites. Specifically, motif logo saliency reveals motifs that predominantly impact the prediction of TF binding sites.

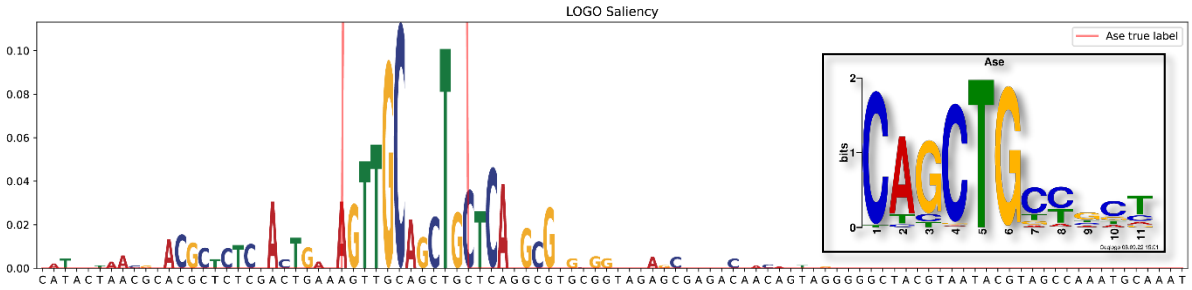
As shown in **Figure 3.2.4**, the motif logo saliency revealed the most important sequence motifs for each of the TF being studied. These motifs were consistent with the known/published motif sequences for respective TFs, suggesting that our model's predictions were biologically relevant and accurate.

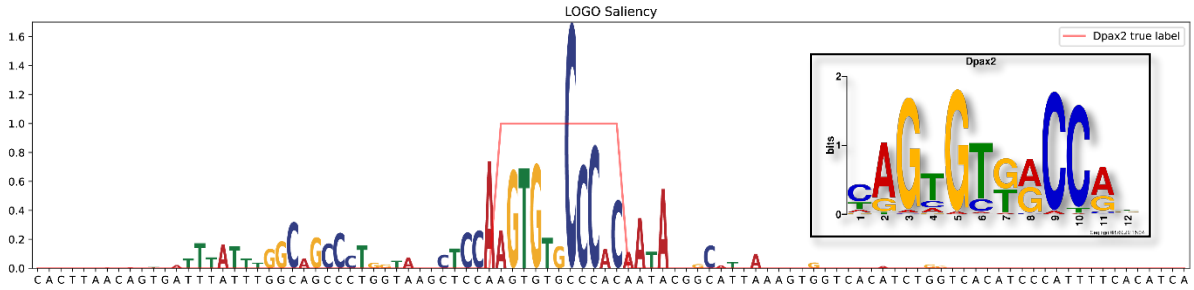
Therefore, our interpretation and validation using saliency and motif logo saliency analysis demonstrate the effectiveness and reliability of our TF model in predicting TF binding sites. These methods provide a valuable tool for understanding the underlying mechanisms of gene regulation and can be used to identify novel regulatory regions and motifs.

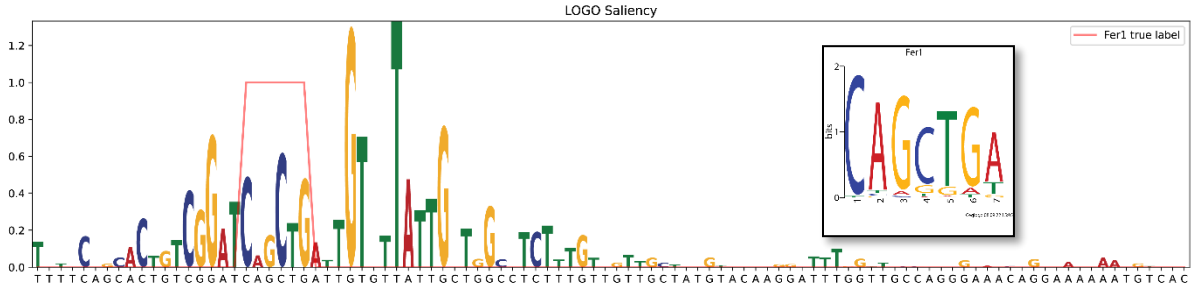
A


What is Saliency map?
X
ACTTGCAG
S_+
positive binding site
= important nucleotide for prediction
Positive Test Sequence
Saliency Map

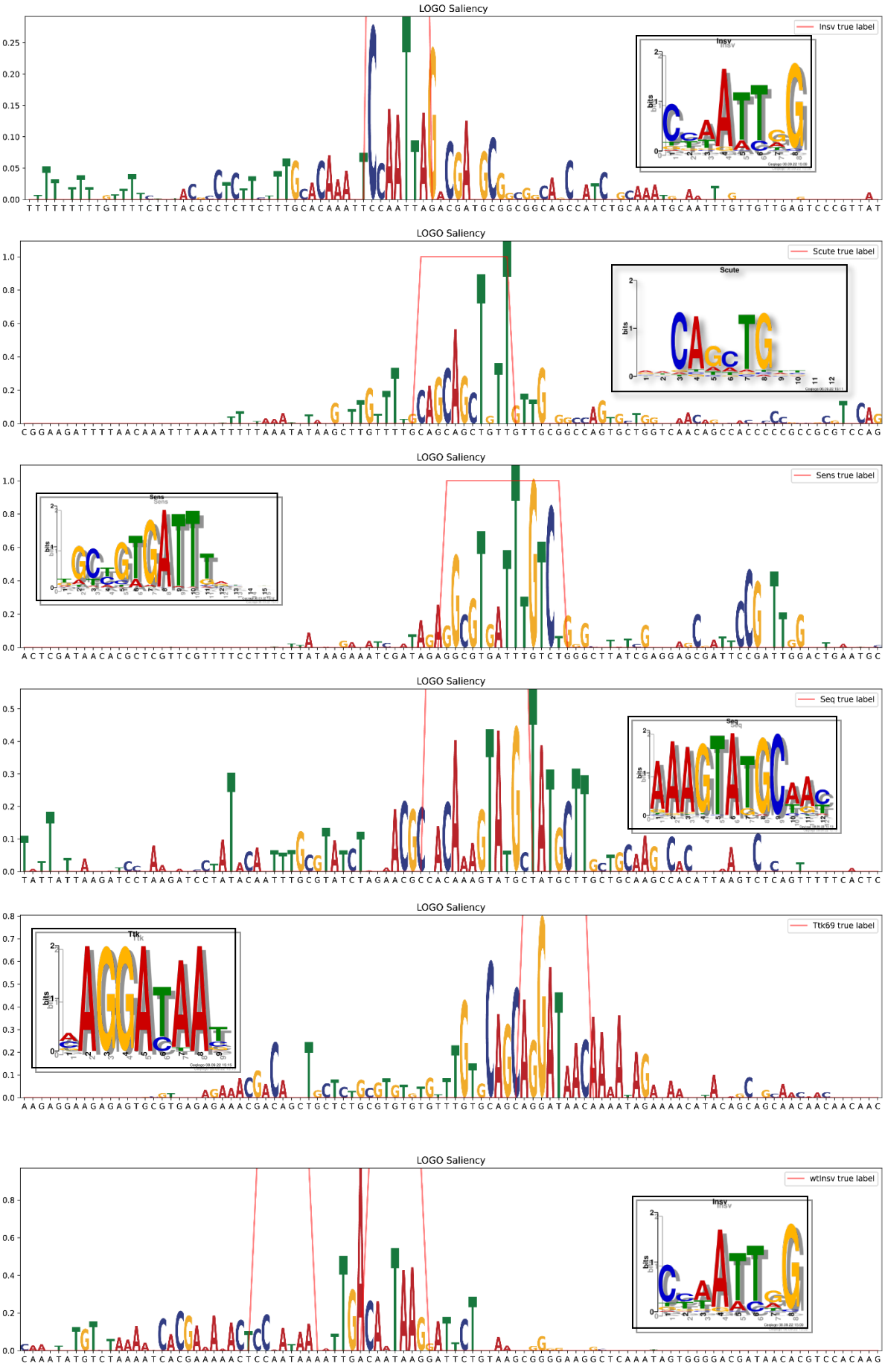
B


LOGO Saliency
Ase true label
Ase
bits


LOGO Saliency
Dpax2 true label
Dpax2
bits


LOGO Saliency
Fer1 true label
Fer1
bits


LOGO Saliency
gtSuH true label
gtSuH
bits


LOGO Saliency
Insv true label
Insv
LOGO Saliency
Scute true label
Scute
LOGO Saliency
Sens true label
Sens
LOGO Saliency
Seq true label
Seq
LOGO Saliency
Ttk69 true label
Ttk
LOGO Saliency
wtInsv true label
Insv

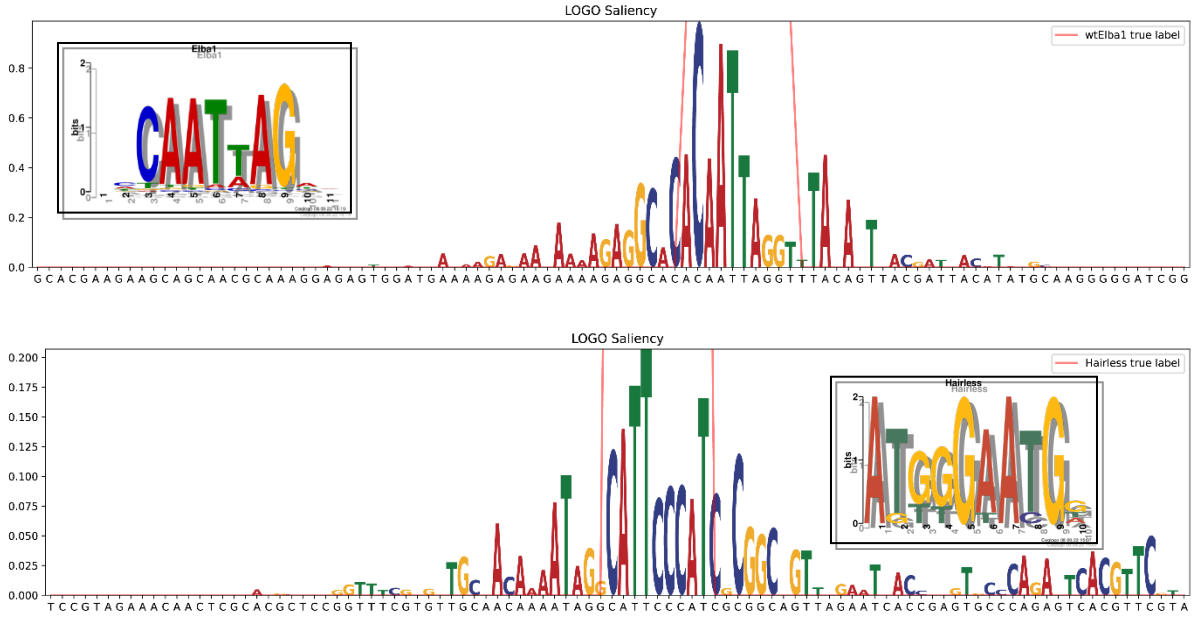


Figure 3.2.4: Interpretation of the classification model (M1) predictions using saliency maps and motif logos to identify the motifs for different TFs. *A. shows what are Saliency maps. It's a method to visualize the influence of each position (i.e. nucleotide) in a DNA sequence on the model's prediction. The saliency map tells us which nucleotides need to be changed the least in order to affect the score the most. B. . The motif logos show the most important sequence motifs for each TF, identified using the saliency maps. These logos depict the nucleotide distribution at each position of the motif, with the height of the letter indicating the frequency of that nucleotide at that position. The motifs in box shows the original motif of these TFs respectively.*

To further investigate the interactions between TFs, we produced both sequence and chromatin accessibility saliency at the predicted TF binding regions. These interactions were subsequently visualized via an interaction plot. This plot delineates the saliency of nucleotide sequences and chromatin accessibility by ATAC, paired by the corresponding true labels and predictions for different TFs. As illustrated in **Figure 3.2.5,** the interaction plot shows the saliency peaks reflecting the saliency scores, indicating the regions in the nucleotide sequence that contribute significantly to the predictions of TF binding. Saliency peaks aligned with high true labels and accurate predictions indicate regions where the nucleotide sequence significantly impacts a TF's binding. Conversely, peaks corresponding to low true labels and imprecise predictions highlight regions potentially negligible for TF binding.

The interaction plot offers a clear visualization of the correlations between TFs, sequence, chromatin accessibility saliency, and their predictions at the TF binding region, helping us understand how specific areas influence the binding of various TFs.

The implications of our findings are broad and significant. First, the effectiveness of TransUNet in handling complex patterns in genomic data suggests that integrating transformer-based models with traditional CNN architectures could be a promising direction for various bioinformatics applications beyond TF binding prediction. Second, the use of advanced interpretation techniques such as saliency maps and potentially Integrated Gradients or LRP in future studies could further illuminate the 'black box' nature of deep learning models, making them more transparent and trustworthy for biological research.

Furthermore, considering the variation in model performance across different TFs, future research could explore more tailored or specialized models that address specific characteristics or behaviours of certain TFs. Additionally, incorporating other types of biological data (e.g., epigenetic markers or 3D chromatin structure data) might enhance the model's ability to predict TF binding sites in a more context-aware manner.

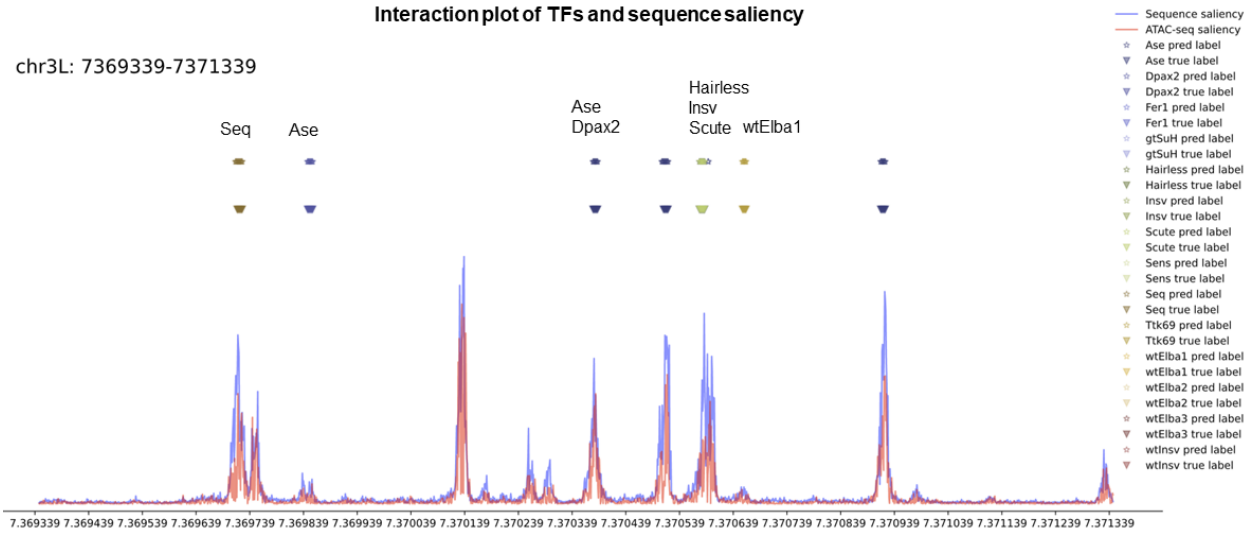


Figure 3.2.5: Interaction plot of TFs and sequence saliency at a predicted TF binding region. *The peaks represent the saliency of nucleotide sequency at Chr3L with the corresponding true labels and accurate predictions of different TFs. The red lines represent DNA sequence saliency and blue lines correspond to ATAC-seq saliency.*

3.3 Prediction of actual downstream gene expression by combining TF binding & chromatin accessibility

3.3.1 Predicting Gene Expression Using PRO-seq Data

We here aimed to develop a regression model to predict downstream gene expression from promoter sequences and chromatin accessibility. We used both wild type PRO-seq "gene body" expression as response variables, and promoter sequences and ATAC-seq coverage as input variables. The "gene body" in PRO-seq spans from the transcription start site (TSS) to the transcription end site (TES) and represents the segment where active transcription and nascent RNA synthesis occur. PRO-seq precisely measures RNA polymerase activity within this region by sequencing nascent RNA transcripts, thereby offering insights into transcriptional dynamics at a single-nucleotide resolution. To accomplish this objective, we employed an architecture similar to the TransUnet model, which has proven effective in classification tasks. We tailored this architecture to the regression task at hand by incorporating an additional linear layer responsible for predicting expression values (For further details, please refer to the methods section).

To evaluate our models, we utilized the coefficient of determination, also known as R-squared (R^2), as the evaluation metric. R^2 quantifies the proportion of the variance in the predicted gene expression values that can be explained by the PRO-seq gene body or RNA-seq expression levels. A high R^2 value indicates a strong correlation between the predicted gene expression values and the true PRO-seq/RNA-seq expression levels, suggesting that the models successfully captured the underlying gene expression dynamics. Conversely, a lower R^2 value suggests that the models may not fully capture the complexity of gene expression and may have room for improvement.

We obtained an R^2 value of 0.599 when evaluating the performance of our predictive models for gene expression (Figure 3.3.1 A). This R^2 value indicates that approximately 60% of the variance in the predicted gene expression values can be explained by the PRO-seq gene body expression levels. An R^2 value of 0.599 suggests

a moderate correlation between the predicted gene expression values and the true PRO-seq expression levels. The Pearson Correlation Coefficient (PCC) value of 0.800 indicates a strong positive correlation between the predicted gene expression values and the actual gene expression levels. This suggests that our model's predictions correlate well with the observed gene expression data. The low p-value of 3.27E-206 further supports the significance of the correlation. The scatter plot shown in **Figure 3.3.1A** compares measured and predicted gene expression values to visually understand the relationship between the actual gene expression levels (measured) and the levels predicted by the regression model (predicted).

Overall, the high correlation, low p-value, and relatively high R-squared value suggest that our model's predictions closely match the actual gene expression levels for the PRO-seq “gene body expression”. This indicates the effectiveness of our regression model in capturing the relationship between promoter sequences, chromatin accessibility, and gene expression levels.

3.3.2 Predicting Gene Expression Using RNA-seq Data

For comparison, we also used wild type RNA-seq expression values as response variables and achieved R-squared value of 0.465, suggesting that 46.5% of the variance in predicted gene expression values can be attributed to actual RNA-seq expression levels **(Fig 3.3.1B**). A PCC value of 0.733 indicates a strong positive correlation between predicted and actual RNA-seq expression levels. The low p-value of 1.44E-183 underscores the statistical significance of this correlation. The prediction performance on RNA-seq expression values is slightly lower than the performance achieved with PRO-seq data, implying that PRO-seq offer greater utility in transcriptional gene expression prediction.

3.3.3 Predicting high/low gene expression

We also sought to assess the capability of our model in distinguishing between high and low gene expression levels. We categorized gene expression into 'high' and 'low' based on whether their values were greater than or less than the median gene

expression, respectively. Using binary classification with our model, we generated a confusion matrix to provide insights into model performance. In this matrix, the rows correspond to the predicted labels (0 and 1), and the columns correspond to the true labels (0 and 1).

For PRO-seq data, the confusion matrix exhibited **(Figure 3.3.1C**):

- True Positive (TP): Predicted label 1 and true label 1, count = 265.
- False Positive (FP): Predicted label 1 but true label 0, count = 100.
- True Negative (TN): Predicted label 0 and true label 0, count = 493.
- False Negative (FN): Predicted label 0 but true label 1, count = 65.
- The prediction accuracy = 82%

For RNA-seq data, the confusion matrix revealed the following (**Figure 3.3.1D**):

- True Positive (TP): Predicted label 1 and true label 1, count = 244.
- False Positive (FP): Predicted label 1 but true label 0, count = 35.
- True Negative (TN): Predicted label 0 and true label 0, count = 547.
- False Negative (FN): Predicted label 0 but true label 1, count = 261.
- The prediction accuracy = 73%

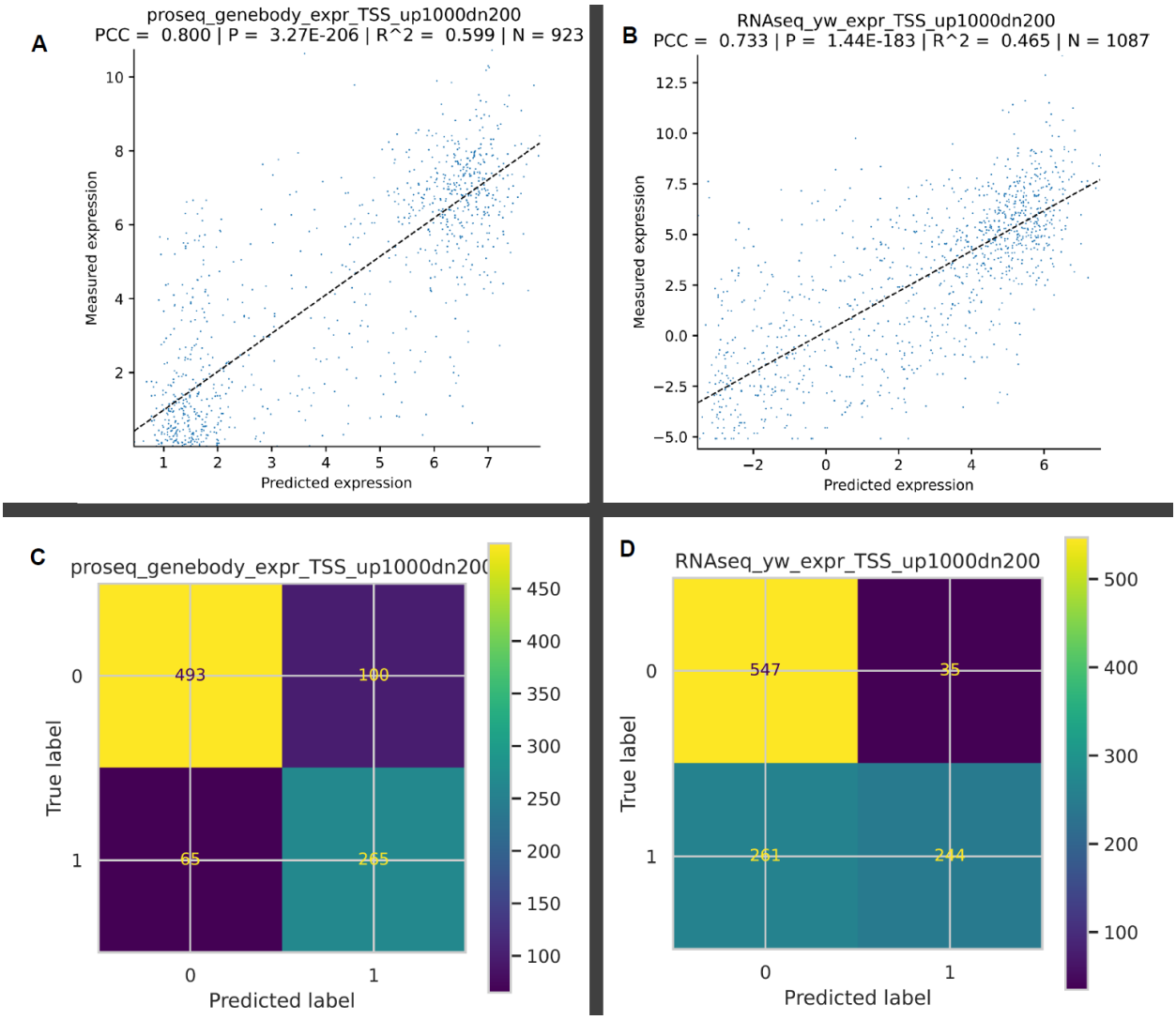


Figure 3.3.1: A) Scatter plot depicting the relationship between measured and predicted gene expression *values generated by the Regression model using PRO-seq gene body expression data. The x-axis represents the predicted values, while the y-axis represents the measured gene expression values obtained from PRO-seq data. The evaluation metrics panel displays Pearson Correlation Coefficient (PCC), P-value, R-squared (R^2), and the number of samples (N). Ideally, the points would align along a straight line with a slope of 1 (45-degree angle), indicating perfect alignment between predicted and measured values.* ***B) Similar scatter plot as in (A), but using RNA-seq wild-type (yw) expression data*** *as the true values. The x-axis presents the predicted gene expression values, and the y-axis corresponds to the actual RNA-seq yw expression values. The evaluation metrics are again displayed in the panel, including PCC, P-value, R^2, and N. The proximity of the plotted points to the 45-degree line indicates the degree of concordance between predicted and actual values.* ***C) Confusion matrix*** *illustrating the relationship between predicted and true labels for PRO-seq gene body expression data. The matrix contains four quadrants representing True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN), indicating model predictions and actual outcomes. D) A parallel confusion matrix, but for RNA-seq wild-type expression data. Similar to (C), the matrix provides insight into model performance, showing TP, FP, TN, and FN instances.*

The results showed that predictions based on PRO-seq gene body expression data were more accurate (82%) compared to those using RNA-seq data (73%). This further suggests that PRO-seq is more adept at capturing transcriptional level regulation.

In conclusion, our work demonstrates the potential of combining deep learning architectures with regulatory genomics data to unravel complex gene regulatory mechanisms. The TransUNet architecture and our regression model hold promise for advancing our understanding of gene expression regulation and its underlying dynamics. These findings contribute to the growing field of computational biology and pave the way for more accurate and insightful predictions of transcription factor binding and gene expression outcomes. Our findings revealed that PRO-seq data, which focuses on active transcription, allowed for more accurate predictions, achieving an R-squared value of 0.599, compared to 0.465 with RNA-seq data. This suggests that real-time transcription activity provides richer information for prediction models. The adaptation of models specifically optimized for different types of expression data (e.g., PRO-seq vs. RNA-seq) could potentially reduce the prediction error and increase the applicability of these models in different biological contexts.

3.4 Cell-type specific prediction of binding-sites at single-cell level

3.4.1 Analysis of scATAC-seq data

Cell-type specific prediction of binding sites at the single-cell level is an important approach that enables the identification and characterization of regulatory DNA regions, known as binding sites, within specific cell types. These binding sites are crucial for understanding the intricate regulatory networks that control gene expression and cellular function. At its core, this analysis aims to unravel the interactions between transcription factors and other DNA-binding proteins with specific DNA sequences in individual cells. By binding to the specific DNA sequences, transcription factors can influence the activity of nearby genes, leading to changes in gene expression patterns. In the context of single-cell genomics, the goal is to decipher these interactions at a single-nucleotide resolution, considering the inherent heterogeneity across different cell types. Each cell type may have distinct regulatory programs and gene expression profiles, and therefore, understanding the binding sites specific to each cell type is crucial for unravelling the complexities of cellular diversity.

In this study, our primary focus was on neuronal cell types, and we utilized a publicly available scATAC-seq data from *Drosophila* early embryos to obtain cell-type specific peaks. First, I present the results from our re-analysis of the scATAC-seq data (Cusanovich et al., 2018) to obtain neuronal specific data. The scATAC-seq analysis outlined in **Figure 3.4.1(A-F)** represents a significant advancement in understanding the complex cellular landscape at the single-cell level. This analysis, carried out using the authors' provided pipeline, offers a comprehensive view of the chromatin accessibility dynamics within the context of cellular development. In the paper, the authors have used Latent Semantic Indexing (LSI) to identify clades of cells with similar chromatin accessibility profiles. The analysis begins by processing data stored in a sparse binary matrix format, with genomic loci as rows and cells as columns. To ensure the data's quality, a filtering step eliminates 2kb windows that intersect with ENCODE-defined blacklist regions. This meticulous curation results in a dataset encompassing 83,290 distinct 2kb windows and data collected from 7,880 cells. Figure **3.4.1A** presents a

histogram depicting the distribution of fragment sizes against their counts. This histogram's three distinct peaks suggest the high quality of the data, indicating well-defined fragment categories that contribute to a reliable dataset. This quality assurance step is essential as it assures the validity of subsequent analyses.

A critical aspect of this analysis involves examining the frequency of insertions within individual windows and understanding the range of windows observed within each cell. The frequency distribution of insertions follows an intriguing pattern, resembling a log-normal distribution when plotted on a logarithmic scale. This distribution indicates that certain sites are observed in a substantial number of cells, with a peak around 6,322 cells, and tapers off as fewer cells exhibit these insertions. The investigation then focuses on retaining the most frequently used sites. Transformation of the data is a crucial step in LSI. Using the TF-IDF approach, which considers both the frequency of insertions and the prevalence of the genomic site across the dataset, the data is transformed into a more informative representation. This representation is then subjected to truncated Singular Value Decomposition (SVD), reducing its dimensionality while preserving significant patterns. The transformed data, or latent semantic index, captures essential characteristics of the genomic data. The resulting latent semantic index representation is analysed through dendrograms, which provide a hierarchical clustering of cells and genomic sites. These dendrograms reveal clusters of cells that exhibit similar insertion patterns and groups of genomic sites that are co-occurring across cells. To visualize these relationships, a bi-clustered heatmap is generated. This heatmap showcases the connections between cells and sites, offering insights into the patterns and associations present in the data. (**Figure 3.4.1C**). It visually captures the organization of cells within specific germ layers, offering a more granular view of cell clustering and accessibility patterns. This understanding is crucial for identifying potential regulatory hotspots and biological processes active in specific regions.

Moving to **Figure 3.4.1B**, the TSNE plot provides an intriguing visualization of cell clustering. The data transformation in TSNE involves applying TF-IDF, similar to LSI. However, this time, the analysis also filters out sex chromosome sites to avoid any gender-related biases. Following the transformation, TSNE is used to generate a lower-

dimensional representation of the data. This lower-dimensional representation retains essential patterns while making the data more manageable for analysis. Density peak clustering, a technique for identifying clusters within a dataset, is then applied to the lower-dimensional TSNE representation. This process aims to group cells with similar patterns together, facilitating the identification of distinct cell populations.

The TSNE plots provide insights into the spatial arrangement of cells based on their patterns. The density peak clusters are visualized, demonstrating how TSNE effectively identifies smaller subclusters within the larger cell populations. The subsequent **Figure 3.4.1D** provides a profile plot and heatmap illustrating read density across peak centres. This visualization offers insights into the distribution of chromatin accessibility, helping to pinpoint regulatory regions with higher activity.

Next, **Figure 3.4.1E** introduces pseudotime analysis, shedding light on the order of cell development. The objective of this analysis is to trace the developmental trajectories of cells using single-cell data. The focus is on the 2-4 hour time point, and a CellDataSet (CDS) is created for cells from this time point. These cells have been classified into 7 distinct cell types based on TSNE analysis: 'Unknown', 'Collisions', 'Blastoderm', 'Neural', 'Ectoderm', 'Mesoderm', and 'Endoderm'. While seven distinct cell types were identified based on the TSNE analysis, the cluster labelled as “Unknown” was excluded from the input for Monocle2. However, the primary interest is in understanding how cells transition from the blastoderm state to the germ layers. The plot reveals how cells evolve through pseudotime, highlighting the shifts in cell types over developmental stages. The trajectory plot indicates that initially, all cells are following a single trajectory, primarily composed of 'Blastoderm' cells. As pseudotime progresses, the trajectory branches out, forming three distinct branches primarily composed of cells from the three major germ layers observed at the 2-4 hour time point. **Figure 3.4.1F** presents a pseudotime analysis revealing the fate of cells within germ layers. This analysis provides a visual narrative of cellular destiny, highlighting how neural cells evolve within the broader context of germ layer development.

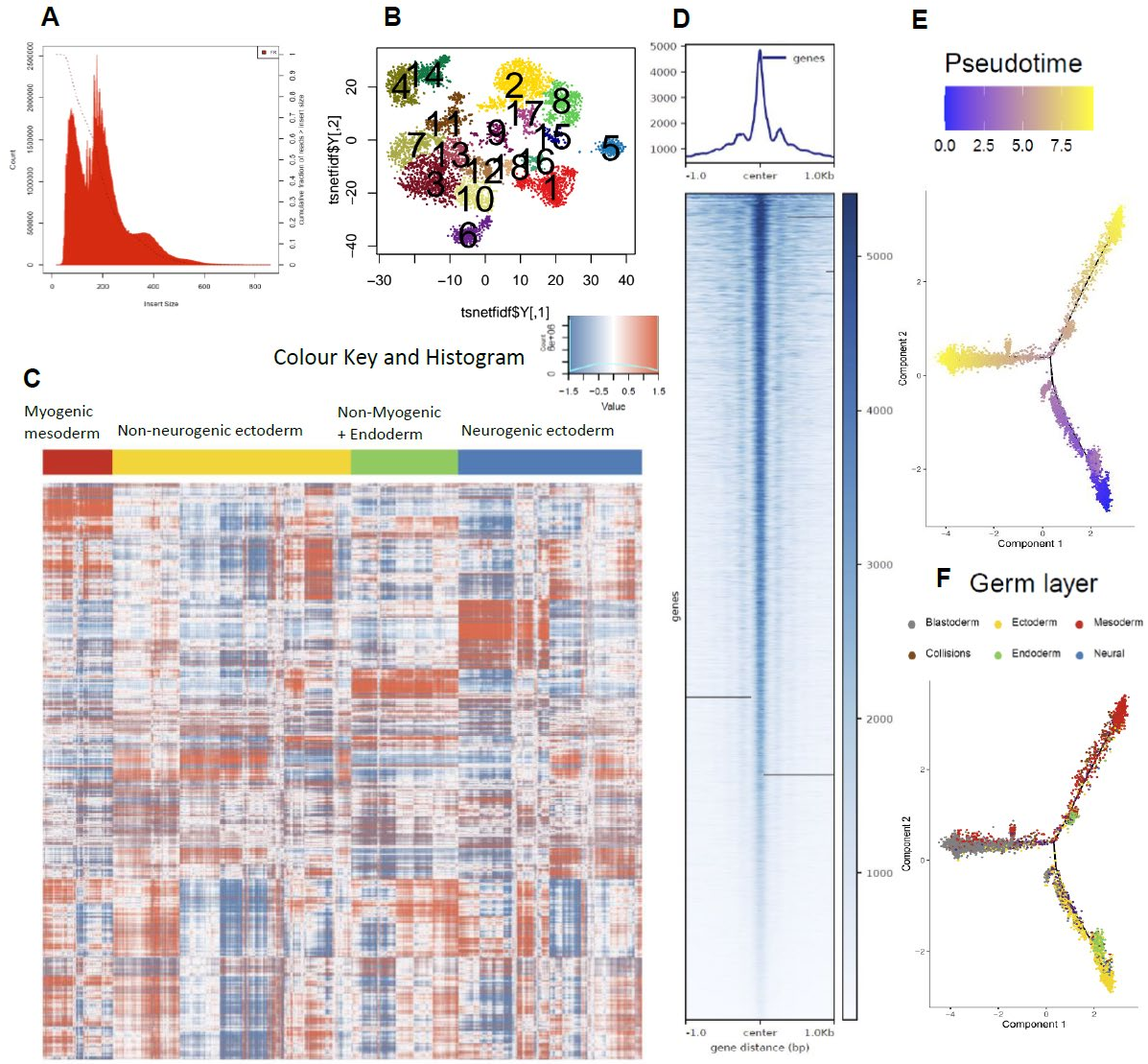


Figure 3.4.1: Reproduced scATAC-seq Analysis*. A) Histogram of Insert-size vs Count. B) TSNE plot showing cell-clustering C) Heatmap of differential accessible sites within the LSI clusters. D) Profile plot and heatmap depicting read density across the peak centre. E) Pseudotime analysis showing order of development of cell. F) Pseudotime showing fate of germ layers in the development.*

Continuing the scATAC-seq data analysis, a comprehensive investigation was conducted using the scATAC-seq explorer tool (https://github.com/JetBrains-Research/sc-atacseq-explorer), designed for in-depth exploration of single-cell ATAC-seq data. This analysis focused on the three distinct time-points provided in the study: 2-4 hours After Egg Laying (AEL), 6-8 hours AEL, and 10-12 hours AEL. To gain a deeper understanding of the chromatin accessibility dynamics, additional analyses were

performed to unveil critical insights. The culmination of these analytical efforts resulted in a series of plots and metrics, one of which is highlighted in **Figure 3.4.2-3.4.3**. This figure provides a condensed visual representation for one of the time-points: 2-4 hours, offering a snapshot of the various parameters investigated.

To begin, an Overlapping Cells (OC) analysis was employed to assess various aspects of the data. This encompassed the determination of the number of cells and peaks across the three time-points. Additionally, the distribution of peak lengths was scrutinized, shedding light on the variability and characteristics of the accessible chromatin regions (**Figure 3.4.2A**). This information was crucial in understanding the scale and complexity of chromatin remodelling events at each time-point.

Furthermore, I invested into the presence of cells exhibiting coverage greater than zero per peak, as well as peaks with coverage greater than zero per cell (**Figure 3.4.2B-C**). This intricate analysis elucidated the extent of chromatin accessibility at a cell and peak level, providing insights into the dynamics of open chromatin regions across developmental time-points.

A pivotal aspect of the analysis involved assessing the coverage of unique molecular identifiers (UMIs) per cell. This metric served as a proxy for the sequencing depth and allowed for normalization and comparison across cells. The summarized distribution of UMI coverage across cells was a valuable metric for understanding the sequencing depth and data quality **(3.4.2D**).

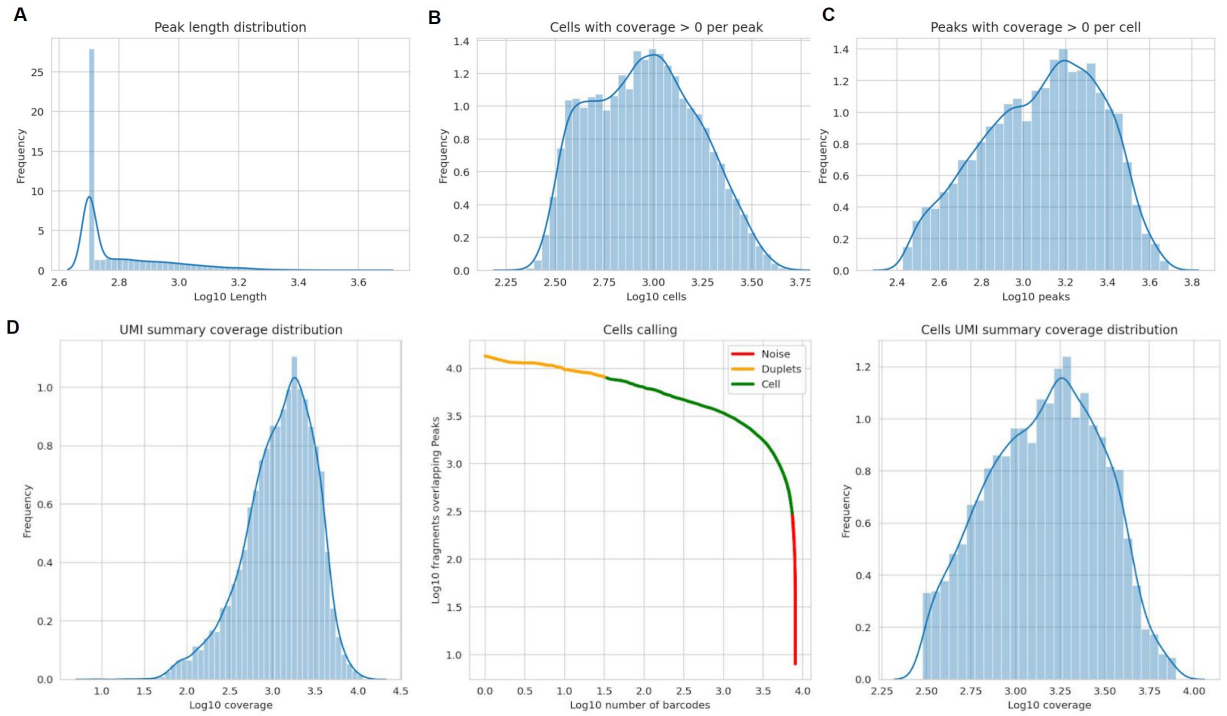


Figure 3.4.2: QC of scATAC-data at individual timepoints. *A) Peak and Cell Enumeration: Peak length distributions at 2-4 hours' time-point. B) Cell-Level Coverage per Peak C) Peak-Level Coverage per Cell D) Analysis of UMI Coverage and Cell Quality Filtering. This section comprises of three sub- plots summarising the quality of scATAC-seq data. First, UMI Summary Coverage Distribution: This plot visualizes the distribution of UMI coverage across all cells on a logarithmic scale, highlighting the general spread and density of UMIs per cell, which provides an overview of sequencing depth across the dataset. Second, Cell Filtering Based on Noise and Duplets: Shows a dual analysis where cells are filtered based on predefined noise and duplet thresholds. The red line indicates cells identified as noise (cells below the noise threshold), while the yellow line marks potential duplets (cells above the duplet threshold). This subplot provides insights into the quality control steps taken to refine cell counts based on UMI counts. Third, Filtered Cells UMI Coverage Distribution: After applying noise and duplet filters, this plot displays the UMI coverage distribution of the remaining cells deemed to be high-quality.*

Another set of plots showcased the characteristics of peak coverage before and after applying a filtering process. The filtering aimed to identify peaks that were informative and potentially biologically relevant (**Figure3.4.3A**). A plot was created to depict the relationship between peak coverage and peak length. This visualization allowed us to explore how the coverage of peaks, which represent regions of open chromatin, correlated with their respective lengths. The x-axis displayed the logarithm (base 10) of the number of cells with coverage for each peak, while the y-axis represented the length of the peaks. Each point on the scatter plot was coloured based on its density,

providing insights into the distribution of peak lengths and their coverage across cells. This plot served as a starting point for understanding the distribution of accessible chromatin regions and their representation in individual cells (**Figure 3.4.3B**). Next, hierarchical clustering was performed on the mean normalized LSA coordinates of cells within each cluster. This analysis aims to capture hierarchical relationships and structural similarities among clusters. The dendrogram resulting from this clustering offers a visual representation of cluster linkage and helps to uncover potential subgroups within the broader clusters (**Figure 3.4.3C**).

Two additional analyses focus on cluster-level characteristics. The first plot displays the distribution of the summary log10 coverage of UMIs per cell within each cluster. This violin plot helps understand the variability in UMI coverage across clusters. Similarly, the second violin plot showcases the distribution of peaks with non-zero coverage per cell, shedding light on the density of accessible chromatin regions across clusters (**Figure 3.4.3D**). Silhouette analysis is employed to assess the quality and coherence of the generated clusters. The silhouette score, calculated for different numbers of clusters, provides an average measure of how well-separated the clusters are. Moreover, individual silhouette coefficients for each cell are computed and visualized, showing the relative distance between the cell and its own cluster compared to other neighbouring clusters. The silhouette plot assists in determining the optimal number of clusters for meaningful separation and interpretation (**Figure 3.4.3E**). The analysis was followed by the hierarchical clustering of coverage in curated markers. This heatmap depicts the coverage patterns of marker regions across different clusters. The rows represent marker regions, and the columns represent clusters. The colour intensity in the heatmap indicates the coverage level of each marker in a specific cluster This examination aids in understanding the chromatin accessibility characteristics associated with markers that are indicative of distinct cellular states or functions (**Figure 3.4.3F**).

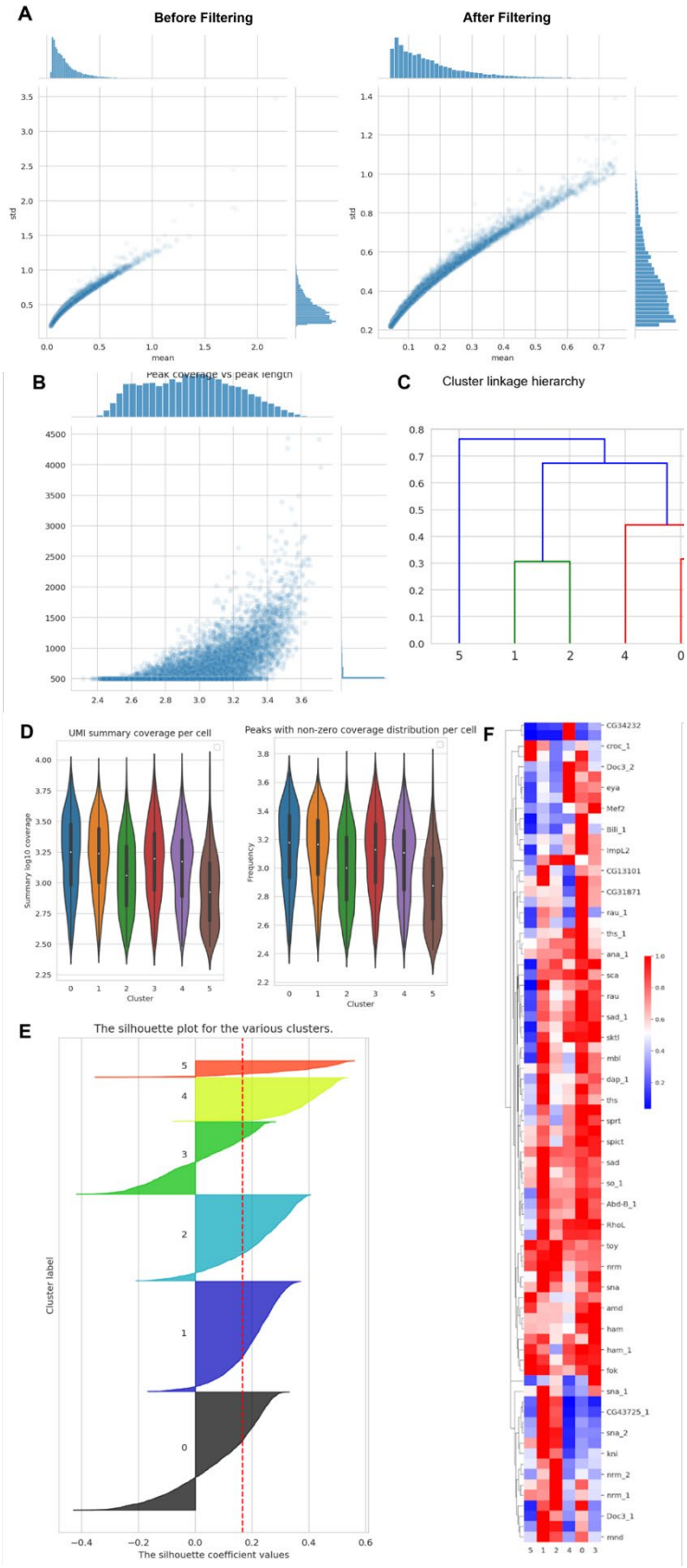

A
Before Filtering
After Filtering
std
mean
B
Peak coverage vs peak length
C
Cluster linkage hierarchy
D
UMI summary coverage per cell
Summary log10 coverage
Cluster
Peaks with non-zero coverage distribution per cell
Frequency
E
The silhouette plot for the various clusters.
Cluster label
The silhouette coefficient values
F
CG34232
croc_1
Doc3_2
eya
Mef2
Bib_1
ImpL2
CG13101
CG31871
rau_1
ths_1
ana_1
sca
rau
sad_1
sktl
mbl
dap_1
ths
sprt
spict
sad
so_1
Abd-B_1
RhoL
toy
nrm
sna
amd
ham
ham_1
fok
sna_1
CG43725_1
sna_2
kni
nrm_2
nrm_1
Doc3_1
mnd

Figure 3.4.3: Exploratory analysis and cluster evaluation of scATAC-seq data at time-point 2-4 hours AEL*. A) Peak Filtering Effects: The impact of peak filtering on informative peak identification is depicted in before and after filtering panels. Peaks were filtered to discern potentially relevant features. B) Peak Coverage vs. Peak Length: A scatter plot illustrating the correlation between peak coverage (log10 of cells with coverage) and peak length. The density-based colouring of points reveals the distribution of accessible chromatin regions across cells. C) Hierarchical Clustering: Hierarchical clustering of mean normalized LSA coordinates for cells within clusters, showcasing hierarchical relationships and structural similarities. The dendrogram depicts cluster linkage and potential subgroup distinctions. D) UMI and Non-Zero Peaks Distribution: Violin plots display the distribution of summary log10 coverage of UMIs and peaks with non-zero coverage per cell in each cluster, elucidating variability and density across clusters. E) Silhouette Analysis: Silhouette score computation and visualization for evaluating cluster quality. Silhouette coefficients for individual cells provide insights into their separation within their clusters. The vertical dashed line represents the average silhouette score of 0.16 for the 6 clusters. The silhouette score, ranging from -1 to 1, measures how similar an object is to its own cluster compared to other clusters. A score closer to 1 indicates better separation and cohesion of clusters, while a score near 0 suggests overlapping clusters. The average silhouette score gives an overall measure of cluster quality. F) Marker Coverage Heatmap: Hierarchical clustering of coverage within curated markers, represented as a heatmap. The heatmap visualizes the coverage patterns of specific genomic markers across the clusters. Each row corresponds to a marker, and each column corresponds to a cluster. The intensity of colour in each cell represents the coverage level, with red indicating higher coverage and blue indicating lower coverage.*

3.4.2 Neuronal cell-types in the developing *Drosophila* embryo

In the early embryonic stages of *Drosophila* development (2-6 hours and 6-12 hours after egg laying), various neuronal cell types begin to differentiate and form within the developing embryo. These stages are critical for establishing the neural architecture and laying the foundation for the nervous system's formation. Our exploration is further elucidated through **Figure 3.4.4A and 3.4.4B**, each offering distinct perspectives on the dynamic distribution and differentiation of various cell types across different time-points.

Figure 3.4.4A portrays the TSNE plots meticulously generated by the study's authors across the three distinct time-points. These plots effectively illustrate the distribution of diverse cell types and their intricate graphical clustering patterns. Moving on to **Figure 3.4.4B**, this figure showcases the TSNE generated through my own iterative re-clustering of the dataset using Seurat's Louvain algorithm with specific parameters. The reanalysis involved running PCA with 30 principal components, finding nearest neighbours using PCA dimensions 1 to 30, and clustering the data with a resolution of 0.4. This reanalysis involved running PCA with 30 principal components, identifying nearest neighbours using PCA dimensions 1 to 30, and clustering the data with a resolution of 0.4. Additionally, the ‘GeneActivity’ function from Signac was utilized to estimate gene activity based on chromatin accessibility.

During the initial time-points (2-4 hours), the cellular landscape is characterized by preliminary developmental stages, often referred to as germ layers. At this juncture, the precise definition of individual cell types is still evolving. The principal clusters identified during this period include Blastoderm, mesoderm, ectoderm, endoderm, and neuroectoderm (neural). As we transition to the 6-8 hour time-frame, a more intricate differentiation and development within the germ layers becomes apparent. Notable transformations encompass the emergence of specific cell-types such as CNS A and CNS B (originating from the neural ectoderm). This phase also reveals subclusters within the ectoderm, distinct regions like foregut and hindgut evolving from the Ectoderm/Blastoderm, and muscle primordia groups A to C originating from the mesoderm. Furthermore, the endoderm gives rise to clusters of haematocytes, midgut,

and fat body. Progressing to the later time-point, 10-12 hours, a clearer resolution of distinct cell types becomes evident. Notably, the neural system evolves into cell types like PNS Sense and Glia, marking a more refined cellular differentiation. This temporal progression towards discernible cell types provides valuable insights into the developmental trajectory and diversification of cellular states.

In the case of **Figure 3.4.4B,** a similar trend of converging clusters over time is observable. The clusters tend to consolidate as development advances, ultimately leading to well-defined and distinct cellular groupings during the later stages of the developmental timeline. This dynamic representation enhances our understanding of how cell types evolve and differentiate over time.

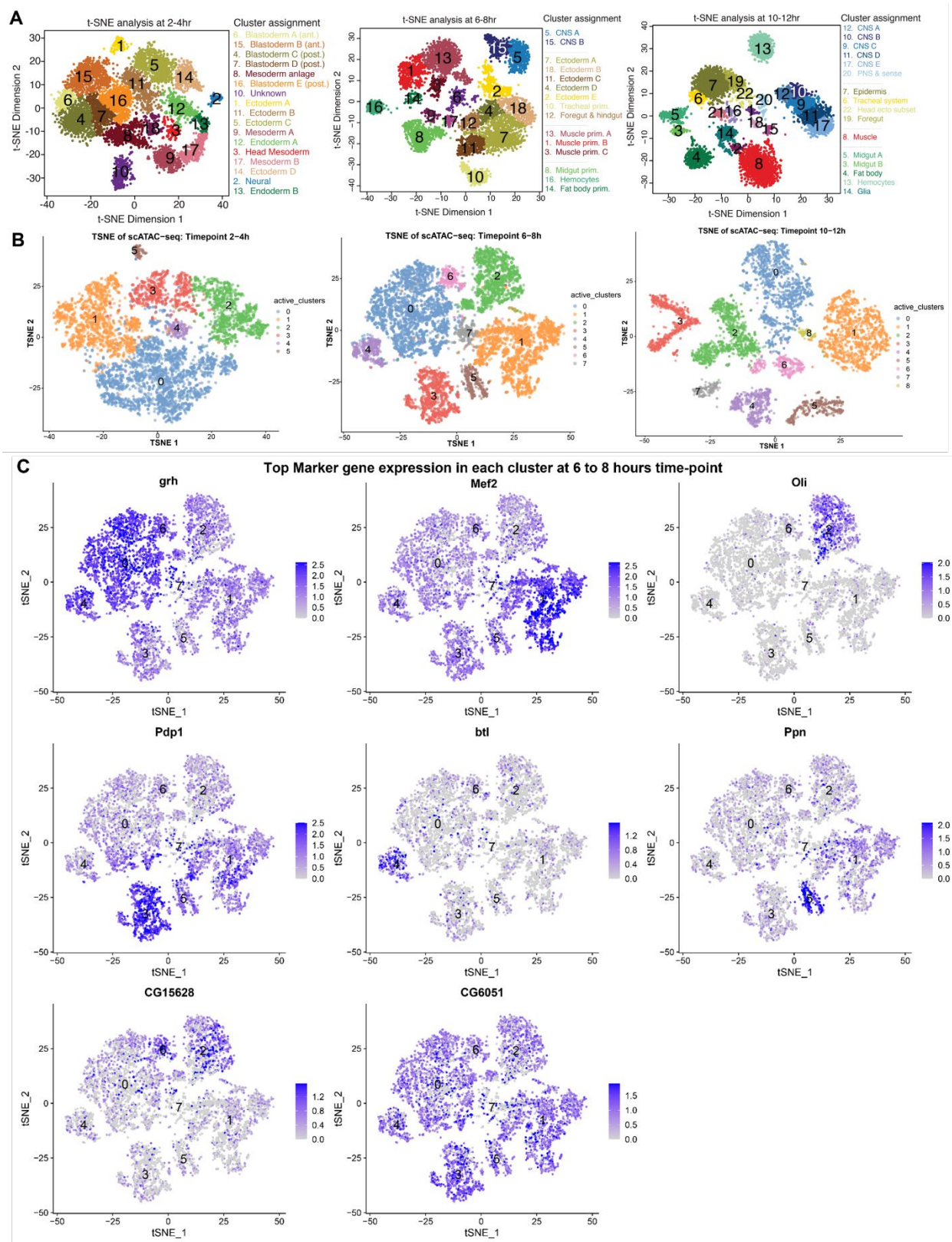

A
t-SNE analysis at 2-4hr
Cluster assignment
6. Blastoderm A (ant.)
15. Blastoderm B (ant.)
4. Blastoderm C (post.)
7. Blastoderm D (post.)
8. Mesoderm anlage
16. Blastoderm E (post.)
10. Unknown
1. Ectoderm A
11. Ectoderm B
5. Ectoderm C
9. Mesoderm A
12. Endoderm A
3. Head Mesoderm
17. Mesoderm B
14. Ectoderm D
2. Neural
13. Endoderm B
t-SNE analysis at 6-8hr
Cluster assignment
5. CNS A
15. CNS B
7. Ectoderm A
18. Ectoderm B
11. Ectoderm C
4. Ectoderm D
2. Ectoderm E
10. Tracheal prim.
12. Foregut & hindgut
13. Muscle prim. A
1. Muscle prim. B
3. Muscle prim. C
8. Midgut prim.
16. Hemocytes
14. Fat body prim.
t-SNE analysis at 10-12hr
Cluster assignment
12. CNS A
10. CNS B
9. CNS C
11. CNS D
17. CNS E
20. PNS & sense
7. Epidermis
5. Tracheal system
22. Head ecto subset
19. Foregut
8. Muscle
5. Midgut A
3. Midgut B
4. Fat body
13. Hemocytes
14. Glia
t-SNE Dimension 1
t-SNE Dimension 2
B
TSNE of scATAC-seq: Timepoint 2-4h
TSNE of scATAC-seq: Timepoint 6-8h
TSNE of scATAC-seq: Timepoint 10-12h
active_clusters
TSNE 1
TSNE 2
C
Top Marker gene expression in each cluster at 6 to 8 hours time-point
grh
Mef2
Oli
Pdp1
btl
Ppn
CG15628
CG6051
tSNE_1
tSNE_2

Figure 3.4.4: Cellular Differentiation and Clustering Dynamics Revealed through dimensional reduction and marker gene expression. *A) TSNE plots published by Cusanovich et al., 2018 depicting scATAC-seq data at three time-points: 2-4 hours AEL, 6-8 hours AEL, and 10-12 hours AEL. Clusters identified at the 2-4 hour time point mainly consist of the germ layers such as Blastoderm, ectoderm, endoderm, mesoderm, and neural cells, denoting early stages of development in fly. At the 6-8 hour time-point, the cell types are further divided into Foregut, hindgut, midgut, muscle prim, CNS A, and CNS B, among others. By the 10-12 hour time point, cell-types like Glia, Fat body, Muscle, epidermis, PNS, and Sense have developed. B) Reproduced t-SNE plots identify the pattern of cell clustering at three different time points. Reproducing the analysis revealed a similar pattern where cells clustered into six, eight and nine distinct clusters respectively for the three time points. C) Feature plots of the top marker gene per cluster for the 6-8 hour time point. Potential cell-type annotation was achieved by examining the expression of the top marker gene for each cluster, highlighted in purple, with the intensity of the colour indicating the level of expression.*

Annotating and interpreting clusters in scATAC-seq data proved more challenging due to the limited knowledge about the functional roles of noncoding genomic regions compared to protein-coding genes. To quantify the activity of each gene in the genome, I assessed the chromatin accessibility associated with gene bodies and promoter regions. A new gene activity assay was derived from the scATAC-seq data by summing fragments intersecting the gene body and a 2 kb upstream region, typically correlated with gene expression. This counting was facilitated by the FeatureMatrix() function, a process automated by the GeneActivity() function in Signac. Although these gene activity profiles were noisier than scRNA-seq measurements due to the sparse nature of chromatin data and the general assumption of correlation between gene body/promoter accessibility and gene expression, they enabled us to begin discerning populations such as mesoderm, endoderm, ectoderm, neural cells, and glia based on these profiles. However, further subdivision of these cell types was challenging based on supervised analysis alone. Consequently, bigwig files of these clusters were downloaded from the published study for further analysis and testing of the model predictions. **Figure 3.4.4C** displays feature plots at the 6-8 hours AEL time-point, showing expression levels characterized by specific marker genes in each cluster indicative of their potential cell types and developmental roles. From these expression levels, I could annotate the clusters into ectoderm (cluster 0), neural cells/CNS A/CNS B (cluster 1), Mesoderm or Muscle primordia (cluster 2), endoderm (cluster 4), Glial cells (cluster 5), Midgut (cluster 6) and cluster 7 remains Unknown due to lack of substantial gene marker expression.

The Ectoderm Cluster (cluster 0) prominently features the marker gene grh (grainy head) along with other genes like ds, uif, ft, dpy, nw, and betaTub60D. Grainy head, a transcription factor crucial for epidermal development, aligns with the ectoderm's function in forming the embryo's outer layer. In cluster 1, the most prominent marker gene is Mef2 (Myocyte enhancer factor 2), accompanied by sns, lmd, Kah, Him, sls, Act87E, Act57B, and Rbfox1. Mef2's known role in muscle and neural development suggests the differentiation of neural cells and formation of the central nervous system (CNS). The Muscle Primordia / Mesoderm Cluster is characterized by the marker gene Myo10A (Myosin 10A), along with ab and Ca-beta. Myo10A's essential role in muscle development aligns with the mesoderm's function in forming muscle tissue. Within the Endoderm Cluster, the most prominent marker gene is dpy (dumpy), accompanied by Myo10A and bowl. Dumpy's involvement in epidermis and muscle attachment supports the endoderm's role in gut development. The Glial Cluster stands out with the marker gene Oli (Oli-neu), along with nerfin-1, msi, and ftz. Oli-neu's marking of oligodendrocytes suggests the presence of glial cells supporting the nervous system. In the Midgut Cluster, the most prominent marker gene is Tollo (Toll receptor), accompanied by arg, tap, chinmo, Kdm4B, CG14459, and elav. Toll receptor's known roles in gut homeostasis and immunity indicate genes related to midgut development and metabolic processes. Lastly, the Unknown Cluster exhibits the marker gene CG6051, alongside Adf1, Gp93, RpS5a, RpII215, CG4266, Dtg, Dpit47, and CG33229. CG6051's function in Drosophila is not well-studied, suggesting this cluster may represent an unknown cell type or developmental process. Transitioning to the adaptation of our bulk-trained model for predicting transcription factor binding sites at the single-cell level, we navigated through the complexity of neuronal cell types. Preprocessing of these neuronal cell types and stages laid the foundation for a deeper and more granular insight. This preprocessing focused on isolating specific cell types and developmental stages that hold critical importance within the neural context.

1. Neural or Neuroectoderm (2-4 hours AEL): This nascent stage represents a crucial foundation for subsequent neural differentiation. Despite the preliminary nature of these cell types, even at this early juncture, we observe glimpses of the complex hierarchical architecture within the neural lineage. This serves as a evidence to the intricate regulatory interactions orchestrating neural development.

2. CNS A and CNS B (6-8 hours AEL): As time progresses to the 6-8 hours AEL time-point, a new vista emerges with the focus on the CNS A and CNS B cell types. This phase marks a significant differentiation event within the neural lineage. These specialized cell types represent distinct branches within the neural hierarchy, each with its unique transcriptional signatures and regulatory dynamics.

A recent study (Calderon et al., 2022), conducted by the same group of authors, generated scRNA-seq analysis for developing *Drosophila* and integrated their results with scATAC seq data for detailed comparison. They curated distinct neuronal clusters at specific developmental time-points. Some key stages included: the Ventral Nerve Cord Primordia (Fly_04-06_12) during the 4-6 hours post-egg laying (AEL) period, followed by the emergence of Peripheral Nervous System (PNS) and Sense cells (Fly_08-10_16) in the 8-10 hours AEL window, and Glia and Glial cells (Fly_10-12_6) within the matured fly at 10-12 hours AEL. I obtained bigwig files for these neuronal clusters as an additional step and applied the model to these individual cell states.

3.4.3 Visualizing predicted TF binding sites at single nucleotide resolution

I applied our bulk TF model to the single-cell context, enabling prediction of TF binding sites within specific neuronal cell types. To achieve this, I utilized scATAC-seq data and DNA sequences as inputs, and then run our pre-trained multi-TF model to predict TF binding sites across various cell types, which were delineated using ATAC-seq in section **3.3.2** (see methods on page 86).

Next, I visualized the genome-wide distribution of predicted transcription factor binding sites for different cell type clusters. Marked with blue vertical lines, these plots depict predicted binding sites for individual TFs for example Sens and Ase across chromosomes in different cell-types, offering an illustrative overview of their genomic dispersion **(Figure 3.4.5)**. I observed different TF binding distribution between different cell types. Figure 3.4.5 features several panels that demonstrate the predicted binding site distribution for these TFs. The first two panels focus on Sens and Ase, showing all predicted binding sites across the TFs' range, with CNS A cluster sites marked in red and CNS B cluster sites in blue.

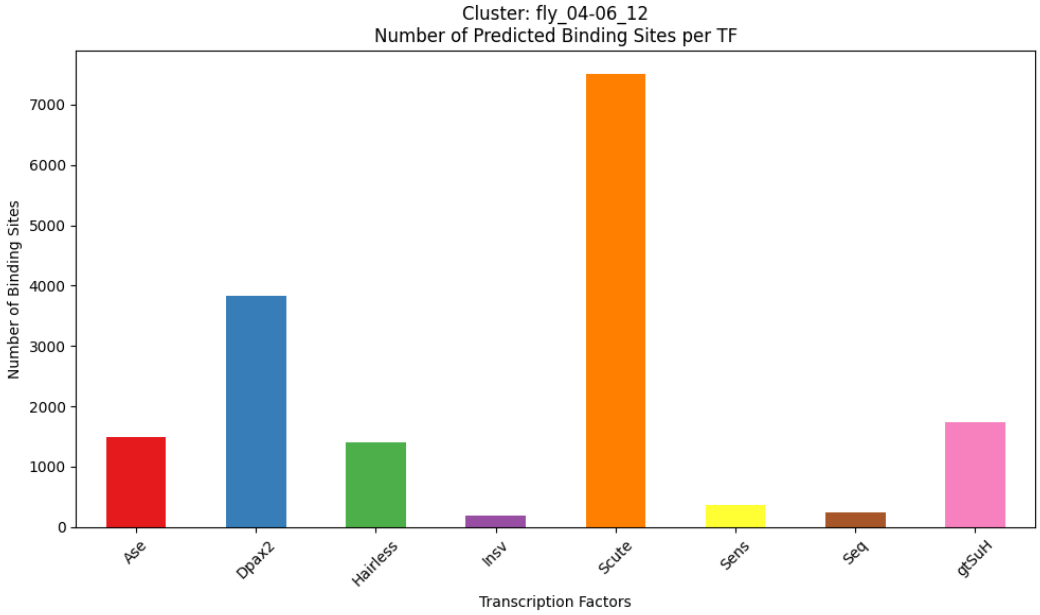
Subsequent panels extend this analysis to other cell clusters namely, VND (Ventral Nerve Cord Primordia), PNS & Sense, and Glial cells each represented by distinct colours (blue, orange, and green, respectively) for both Sens and Ase TFs. These visualizations provide a crucial comparative perspective, revealing variations in TF binding site distribution among different cell types and clusters.

Figure 3.4.5 shows the predicted binding sites across different neuronal cell type clusters, namely VND, PNS & Sense, Glia, CNS A and CNS B. Notably, Scute and Dpax2 exhibit the predominant count of anticipated binding regions within specialized clusters, specifically VND, PNS & Sense, and Glia with Scute exhibiting more than 7,000 predicted sites and Dpax2 around 4,000. Other TFs such as gtSUH, Ase, and Hairless also show substantial activity with over 1,000 predicted binding sites each. Conversely, Insv, Sens, and Seq present the lowest count of binding predictions, each below 500. Further detailing, the plots corresponding to CNS A and CNS B (Figures 3.4.5D and 3.4.5E) reveal that these clusters not only have a higher overall number of predicted binding sites

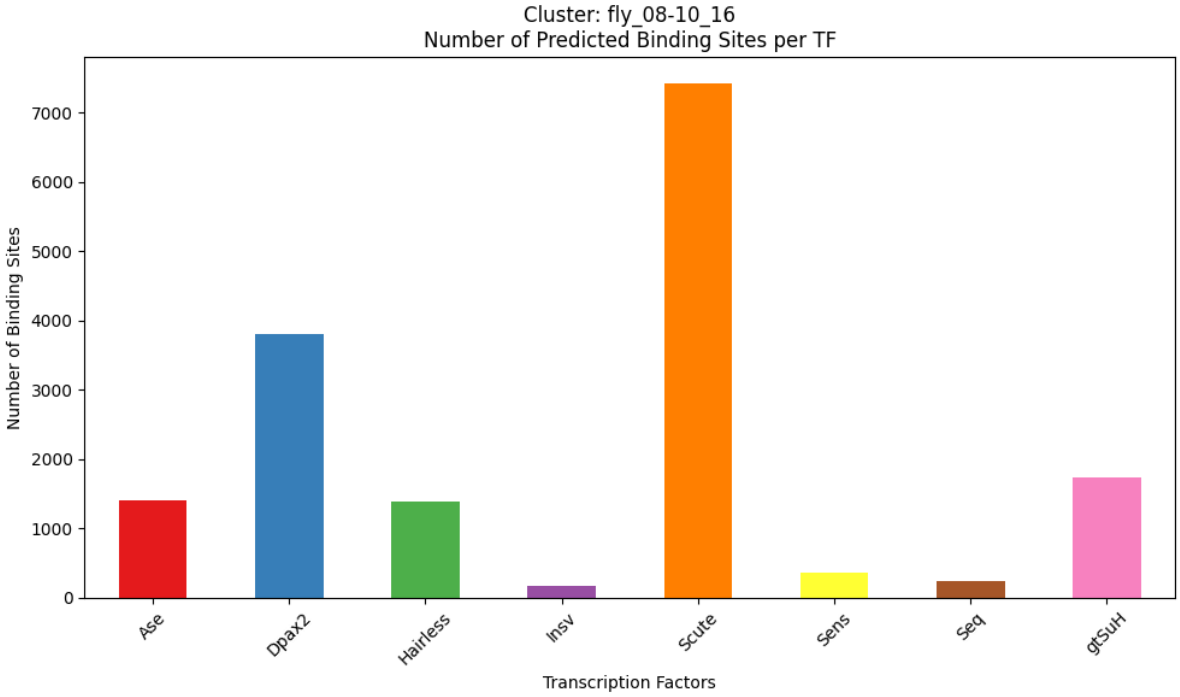
compared to other specialized types but also show significant variation among the TFs. For instance, Dpax2 leads with approximately 34,000 binding predictions in these clusters, followed closely by Ase with around 33,000. This is indicative of a more complex regulatory framework in CNS A and CNS B, with TFs such as Seq, Scute, and wtInsv also showing substantial numbers (25,000, ~20,000, and 18,000, respectively). Meanwhile, TFs like Ttk69, wtElba2, Sens, and Hairless exhibit fewer than 5,000 predicted sites each.

The integration of both qualitative and quantitative visualizations in these figures helps highlight the crucial roles that TFs play in directing gene expression specific to various cell types and developmental stages. It suggests that Scute and Dpax2 may play pivotal roles in the development and differentiation processes of these particular neuronal systems. The analysis not only highlights the distribution of TF binding sites across different neuronal cell clusters but also illuminates the scale of these predictions based on cluster size. CNS A and CNS B, encompassing a broader range of data points, show a relatively higher count of predicted binding sites for all examined TFs. This suggests a more extensive and complex regulatory network within these generalist neuronal regions, likely reflecting their central role in neuronal system functionality and development. Conversely, specialized clusters like Glial, PNS & Sense, and VND correspond to smaller subsets of cells. The fewer number of cells in these clusters results in a lower volume of predicted binding regions. Despite the smaller quantity of binding sites, the impact of TFs such as Scute and Dpax2 in these regions is marked, indicating a strong, targeted regulatory influence in specific developmental contexts.

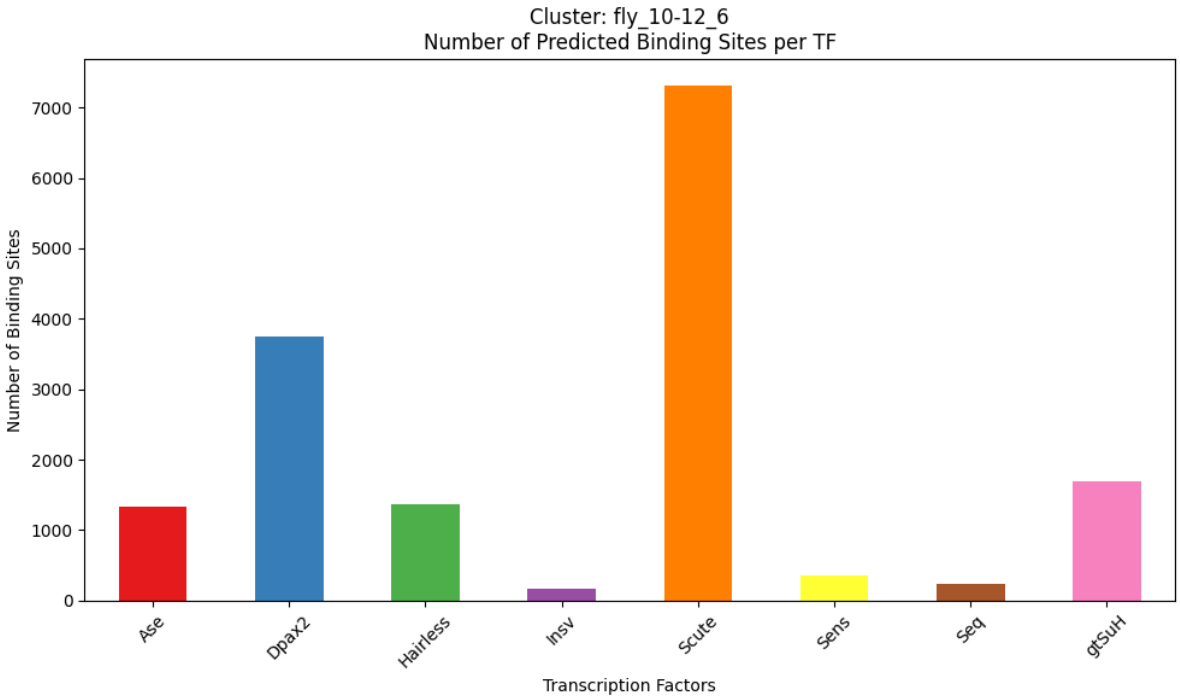
A. Cluster # 12 Ventral Nerve cord Primordia (4-6 hours AEL)



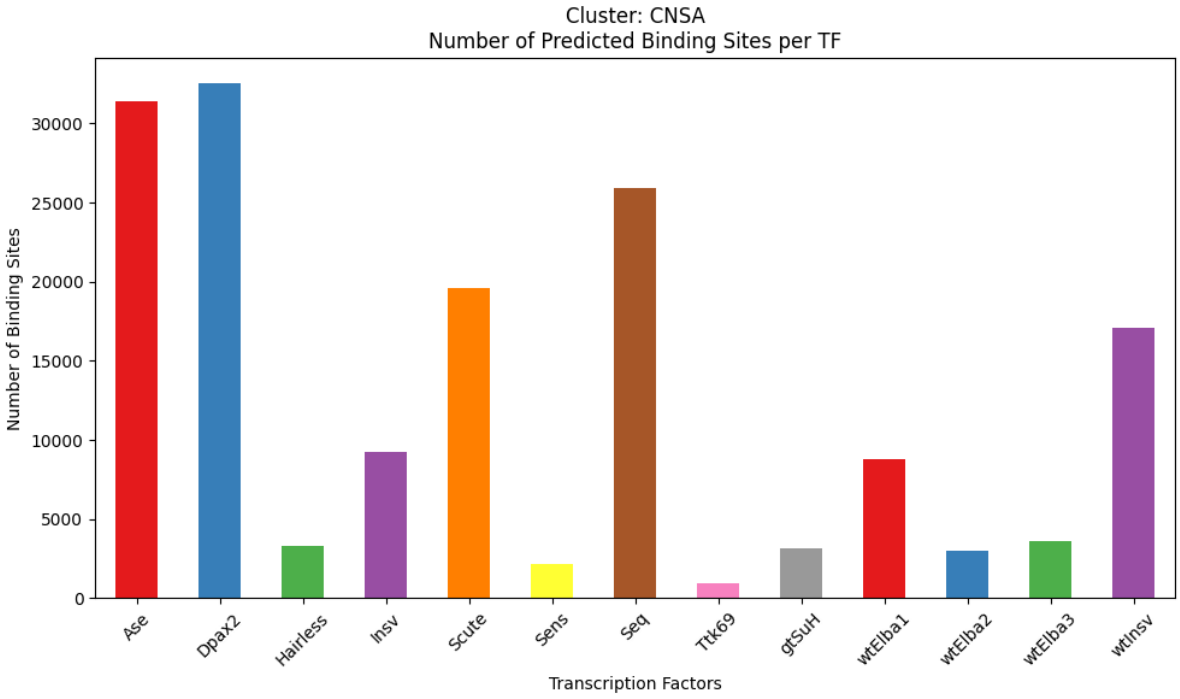
B. Cluster # 16 Glia (8-10 hours AEL)



C. Cluster # 6 PNS and Sense (10-12 hours AEL)



D. Cluster CNS A (6-8 hours AEL)



E. Cluster CNS B (6-8 hours AEL)

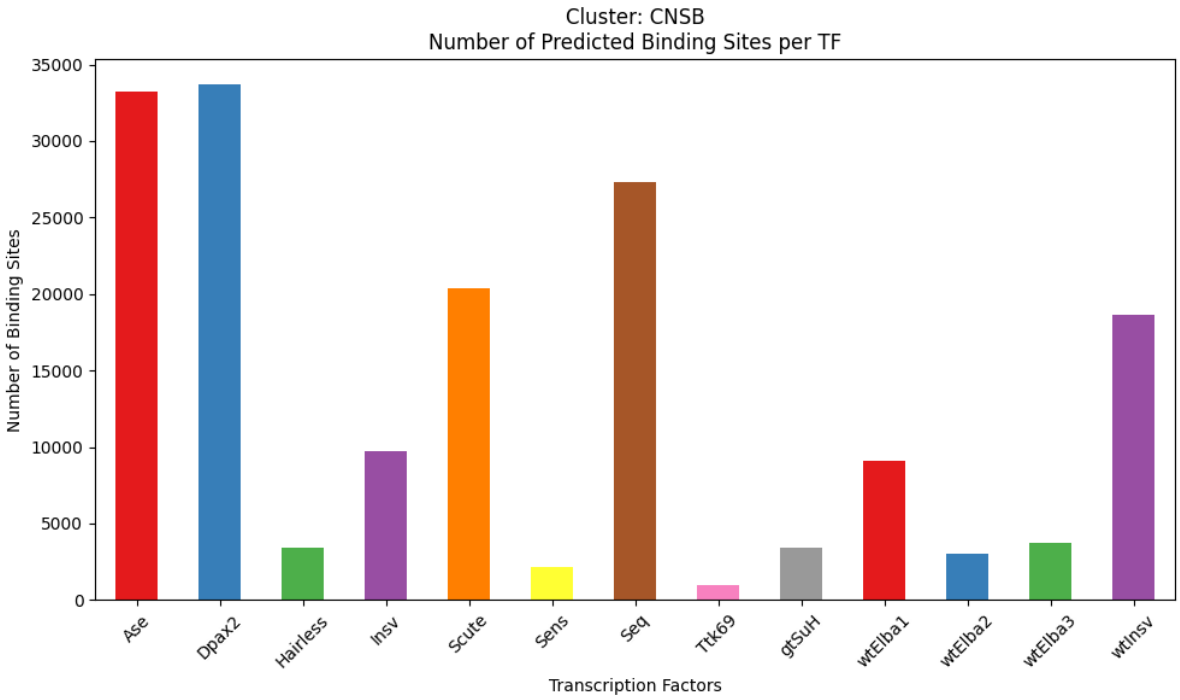


Figure 3.4.5: Bar plots depicting number of predicted binding sites per TF for each neuronal cell-types. *The x-axis represents the different TFs under consideration. Each bar corresponds to a specific TF. The y-axis quantifies the number of predicted binding sites for each TF. Each bar in the graph corresponds to a specific TF, with the height of the bar indicating the frequency of binding sites for the respective TF within the given cell type cluster. The figure includes data from several cell type clusters at various developmental stages, namely Ventral Nerve cord Primordia (4-6 hours After Egg Laying (AEL)), Glia cells (8-10 hours AEL), Peripheral Nervous System (PNS) and Sense organs (10-12 hours AEL), Central Nervous System A (6-8 hours AEL), and Central Nervous System B (6-8 hours AEL). This figure provides a comprehensive overview of the predicted TF binding sites across different cell types and developmental stages, offering valuable insights into the regulatory landscape of these cells.*

3.4.4 Model validation and downstream analysis

To validate the accuracy and reliability of our study's predictions of TF binding sites in single cells, we employed several strategies for downstream analysis. These strategies provided indirect evidence of the accuracy of our predictions and allowed us to assess the robustness and consistency of the predicted TF binding patterns.

A pivotal step in validating our model's interpretations was the comparison of enriched motifs within the predicted binding sites regions with the known TF motifs to correlate with TF binding. Using the HOMER tool, we conducted motif enrichment analysis to identify enriched sequence motifs in the predicted TF regions. The motif enrichment analysis revealed a significant enrichment of nearly all motifs known to be associated with these neuronal TFs except for Fer1 and Seq. **Figure 3.4.6A** displays the enriched sequence motifs for Ase, Dpax2, Sens, H, and GtSu(H), as identified in the predicted TF binding regions. Each motif is represented by a logo that depicts the nucleotide distribution at each position of the motif. The significant enrichment of known motifs associated with neuronal TFs, reinforces the validity of our predictions and suggests that our model effectively captured the underlying regulatory elements governed by these TFs in single cells.

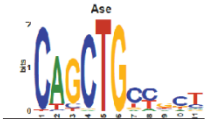
To further explore the functional implications of our predictions, I performed hierarchical clustering based on the predicted TF binding regions and their corresponding gene activity score. The gene activity score was calculated using Cicero (Pliner et al., 2018). It is defined as the overall accessibility of a promoter and its associated distal sites. This score of regional accessibility has a better concordance with gene expression. For each TF, I utilized the predicted binding regions obtained from the model and applied hierarchical clustering using the gene activity score. The clustering algorithm grouped the cells based on the similarity of their predicted binding regions for the TF. Once the clustering analysis was performed for each TF, I visualized the clusters and their relationships using Uniform Manifold Approximation and Projection (UMAP). Figure 3.4.6B demonstrates the UMAP visualization of the resulting clusters for three TFs Ase, Insv and gtSu(H).

Through motif analysis and hierarchical clustering, we validated the accuracy and functional relevance of our predicted TF binding regions. The observed correspondence between our predictions and known motifs, as well as the clustering patterns and enriched TFs within specific clusters, provided robust evidence for the validity of the model's predictions.

A. Known / *de novo* Motif Enrichment (HOMER) in predicted TF binding sites

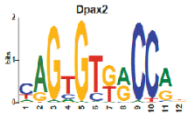
Ase

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	STD(Bg STD)	Best Match/Details
1		1e-220	-5.070e+02	20.87%	2.74%	41.2bp (71.5bp)	E-box/Drosophila-Promoters/Homer(0.957) More Information | Similar Motifs Found
2		1e-198	-4.577e+02	23.98%	4.34%	44.2bp (64.3bp)	Unknown1(NR/Ini-like)/Drosophila-Promoters/Homer(0.900) More Information | Similar Motifs Found
3		1e-132	-3.047e+02	24.24%	6.63%	45.7bp (66.5bp)	Unknown2/Drosophila-Promoters/Homer(0.983) More Information | Similar Motifs Found



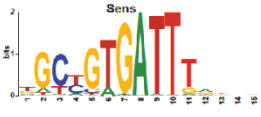
Dpax2

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	STD(Bg STD)	Best Match/Details
1		1e-1579	-3.638e+03	46.57%	4.10%	38.7bp (62.2bp)	Unknown1(NR/Ini-like)/Drosophila-Promoters/Homer(0.934) More Information | Similar Motifs Found
2		1e-274	-6.329e+02	15.06%	2.73%	44.6bp (58.9bp)	DREF/Drosophila-Promoters/Homer(0.985) More Information | Similar Motifs Found
3		1e-274	-6.310e+02	28.22%	9.48%	45.5bp (63.7bp)	Unknown2/Drosophila-Promoters/Homer(0.957) More Information | Similar Motifs Found



Sens

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	STD(Bg STD)	Best Match/Details
1		1e-37	-8.656e+01	23.75%	3.42%	39.0bp (61.6bp)	Unknown1(NR/Ini-like)/Drosophila-Promoters/Homer(0.955) More Information | Similar Motifs Found
2		1e-28	-6.638e+01	10.37%	0.54%	53.9bp (68.3bp)	onecut/MA0235.1/Jaspar(0.700) More Information | Similar Motifs Found
3		1e-24	-5.744e+01	20.40%	4.01%	50.1bp (65.2bp)	pnr/MA0536.1/Jaspar(0.974) More Information | Similar Motifs Found

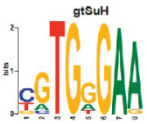


H

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	STD(Bg STD)	Best Match/Details
1		1e-58	-1.346e+02	16.60%	1.59%	31.7bp (54.8bp)	Unknown1(NR/Ini-like)/Drosophila-Promoters/Homer(0.967) More Information | Similar Motifs Found
2		1e-41	-9.649e+01	19.27%	3.64%	54.7bp (70.0bp)	CG11617/MA0173.1/Jaspar(0.619) More Information | Similar Motifs Found
3		1e-32	-7.449e+01	20.61%	5.41%	52.4bp (71.0bp)	Unknown4/Drosophila-Promoters/Homer(0.933) More Information | Similar Motifs Found

gtSuH

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	STD(Bg STD)	Best Match/Details
1		1e-63	-1.467e+02	16.15%	2.17%	40.1bp (60.3bp)	M1BP(Z)/S2R+-M1BP-ChIP-Seq(GSE49842)/Homer(0.958) More Information | Similar Motifs Found
2		1e-58	-1.346e+02	15.60%	2.25%	58.5bp (68.1bp)	Su(H)/dmmpmm(Papatsenko)/fly(0.972) More Information | Similar Motifs Found
3		1e-49	-1.148e+02	14.65%	2.38%	46.2bp (58.9bp)	DREF/Drosophila-Promoters/Homer(0.982) More Information | Similar Motifs Found



B. Clustering of predicted binding sites

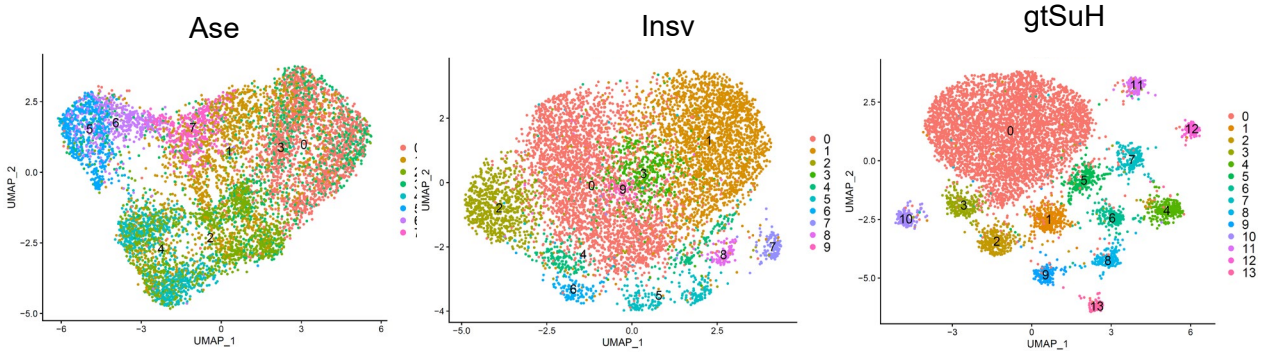


Figure 3.4.6: Validation of model interpretation by downstream analysis. *A. HOMER motif enrichment on predicted regions and comparison with original motifs. B. UMAP clustering of predicted regions with corresponding TF-gene activity.*

To explore the co-occurrence patterns of TF binding motifs, I applied motif co-occurrence analysis (see methods in section 2.2.5 “Strategies for downstream analysis of predicted values” on page 85) to the predicted binding regions. Co-occurrence in this context refers to the simultaneous presence or occurrence of two motifs or TFs at the same binding sites. This involved quantifying the correlation between pairs of TF motifs across the predicted binding regions. The resulting correlation matrix was visualized as a heatmap shown in **Figure 3.4.7,** where each cell represented the strength and direction of correlation between two TFs. Positive values indicated co-occurrence, while negative values indicated exclusion or competitive binding. Notably, the heatmap revealed distinct co-occurrence patterns that could signify the concerted actions of TFs in the context of gene regulation. For instance, the observed higher correlation between Ase, Sens, scute, and gtsuH in the motif co-occurrence heatmap resonates with the findings from the ChIP-seq analysis, as discussed in Section 3.1. This alignment between computational predictions and experimental validation further underscores the reliability and consistency of our results.

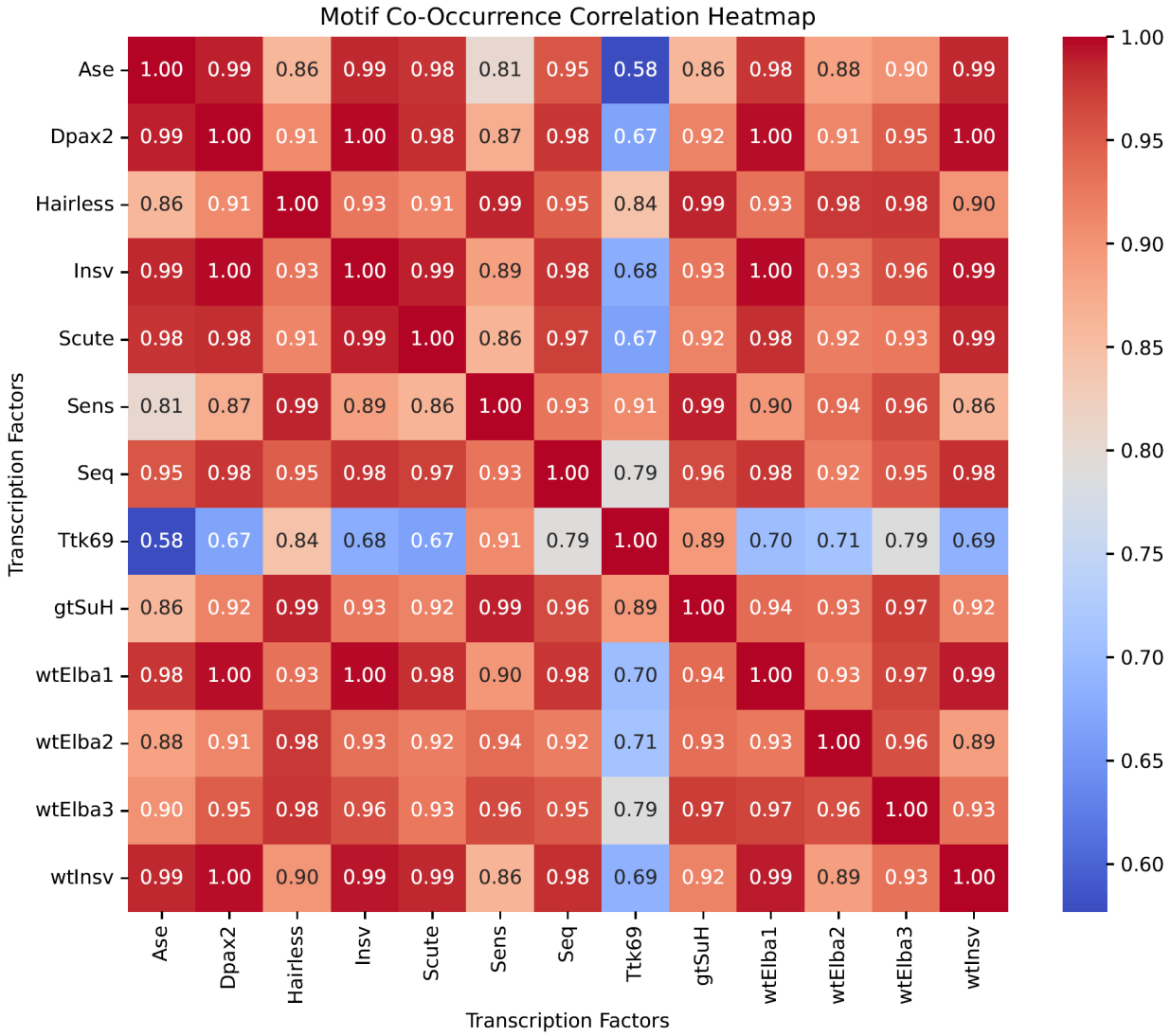


Figure 3.4.7: TF motif co-occurrence from predicted binding sites. It visualizes the correlation between different motifs/ transcription factors. *The x and y axes list various motifs/TFs included in the study. Each axis represents the same set of TFs, allowing for comparison between them. Each cell in the heatmap grid represents the correlation between the motifs/factors on the x and y axes. The color of each cell indicates the strength of the correlation. A colour scale is provided, where blue represents a low correlation (value of -1.00) and red represents a high correlation (value of 1.00). The cells along the diagonal from top left to bottom right are all coloured in red, indicating a perfect correlation of 1.00 with themselves. The highest motif co-occurrence correlation was observed between Dpax2 and Insv, and wtElba1, with a correlation coefficient of 1 (highest). Other notable correlations include Ase with Dpax2 and Insv (0.99), Hairless with gtSuH (0.99), Insv with Scute (0.99), and Sens with Hairless and gtSuH (0.99). The lowest motif co-occurrence was observed between Ttk69 and Ase, with a correlation coefficient of 0.58.*

With the primary aim of developing a sophisticated deep learning model to predict the binding sites of diverse neuronal TFs, this study centred on the construction and

validation of comprehensive bulk and single-cell prediction models. However, the subsequent steps involving experimental validation and the interpretation of the intricate gene regulation variations are intricate tasks that require specialized expertise and extended investigations beyond the current scope. While unravelling the downstream implications of these regulatory differences remains a challenging endeavour, the present work effectively showcases the accurate prediction of TF binding sites at a single-cell resolution. By harnessing advanced computational methods, we substantiate our capability to decipher the intricate regulatory patterns of neuronal TFs with an unprecedented level of precision.

Chapter 4: CONCLUSION

4.1 Summary

In conclusion, this study presented a comprehensive and innovative approach for predicting cell-type specific combinatorial binding of neuronal transcription factors using deep learning techniques. Through the development of a Transformer encoded U-Net architecture (Trans-UNet), the model successfully captured complex regulatory patterns governing gene expression in specific neuronal cell types in early embryonic development of *Drosophila*. We have achieved a remarkable level of prediction accuracy by harnessing the power of deep learning techniques. Our developed approach, featuring the Trans-UNet architecture, has demonstrated an exceptional ability to predict cell-type specific combinatorial binding patterns of neuronal transcription factors (TFs). We obtained high AUC score ranging between 0.95-0.99 for majority of TFs. A noteworthy aspect of our approach is its capability to capture the complexity of combinatorial TF binding. By employing a multi-TF model, we have not only identified individual TF binding sites but also deciphered the cooperative interactions among multiple TFs. This distinction between predicting single TF binding versus capturing combinatorial binding provides a more holistic understanding of transcriptional regulation. We observed that the performance of the model enhances with muti-TF model when predicting binding sites for all TFs simultaneously.

The benchmarking process further underscored the effectiveness of our Trans-UNet architecture. In comparison to other deep learning architectures such as CNN, UNet, our model demonstrated superior performance and significant predictive capabilities.

The Trans-UNet architecture had advantages over the standard Unet with CNN architecture. The high accuracy achieved in predicting TF binding sites underscores the power of deep learning in deciphering complex transcriptional regulatory mechanisms.

The incorporation of classification and regression models within the framework allowed for simultaneous prediction of TF binding sites and gene expression levels from

DNA sequence and ATAC-seq coverage at the bulk level. Moreover, benchmarking our model against similar published approaches and DL architectures has showcased the advancements and competitive advantages that our architecture holds. These benchmarks reinforced the reliability and functional relevance of the model's predictions.

Model interpretation techniques, such as saliency scores and motif logo saliency maps, provided a window into the specific nucleotides and genomic regions that wield substantial influence over TF binding. Notably, the motif logo saliency maps have also highlighted the overlapping binding sites for multiple transcription factors.

Importantly, our approach extends the prediction of TF binding beyond the bulk level, enabling us to predict TF binding sites at single-cell resolution. Knowing TF binding sites at the single-cell level offers several key advantages:

- Cellular Heterogeneity Understanding: Single-cell predictions enable us to explore how TF binding varies across different cell types within a tissue. This helps us comprehend the diverse roles of TFs in governing cell-specific processes and responses.
- Cell State Dynamics: Predicting TF binding sites in individual cells provides insights into the dynamic changes that occur during cellular differentiation, activation, or response to stimuli. This temporal information enhances our understanding of gene regulatory networks.
- Unveiling Regulatory Networks: Identifying TF binding sites in single cells allows us to reconstruct intricate gene regulatory networks at a cellular resolution. This is crucial for deciphering how different TFs collaborate to control gene expression.

The implications of this study extend beyond the accurate prediction of TF binding sites. By shedding light on the cell-type specific combinatorial binding patterns of neuronal TF networks, the research enhances our understanding of the complex regulatory landscape governing neuronal development. Furthermore, the success of this deep learning approach demonstrates its potential for applications in deciphering transcriptional regulatory networks in other biological contexts. By embracing its findings

and navigating the course of future investigations, researchers are well-positioned to solve the mystery of cell-type specific gene expression that shapes the dynamic landscape and production of specific proteins and molecular components that define the unique characteristics and functions of that cell type.

4.2 Significance

By leveraging deep learning models, we adeptly capture the complex interplay of TFs, yielding more precise insights into their cooperative and synergistic effects on gene regulation. The utilization of advanced deep learning architectures, including U-NET and TransUNet, stands as a hallmark of this study. It elevates the precision and interpretability of our predictions. The strategic focus on TF binding in the early embryo of *Drosophila* holds multifaceted significance. The paramount significance is amplified by our privileged access to *Drosophila* early embryo TF chip-seq and ATAC-seq, PRO-seq and RNA-seq data from our collaborators at Stockholm University. This alignment of research interests and data access expedites the validation of our predictions, enriching the study's credibility and impact.

The study emphasizes the accuracy and reliability of the developed prediction model through various validation and interpretation methods. It also highlights the importance of using both bulk and single-cell data for understanding gene regulation dynamics at different levels of resolution.

Continuing along this path, applying these methodologies to human data holds promise for uncovering the mysteries of human neurodevelopment and disease. Comparative studies and extensive benchmarking against existing methods could validate the model's strengths further and uncover novel insights. The exploration of long-range interactions and non-coding RNAs could unravel further layers of regulatory complexity.

4.3 Limitations and Future directions

Despite its significant contributions, this study has certain limitations that warrant consideration. By utilizing sequences from the promoter region for model development, our approach has an inherent limitation: it might not capture the complete spectrum of regulatory interactions. Notably, regulatory elements like enhancers located outside the promoter regions might be excluded, potentially oversimplifying the system's complexity. Additionally, the model's performance might be influenced by the quality and quantity of available training data, potentially limiting its generalizability to other cell types or organisms. The study also focused on a specific set of neuronal-specific TFs and may not encompass the full spectrum of regulatory factors involved in neuronal development. The study's focus on a specific set of 13 neuronal TFs introduced variability in the distribution of true label predictors for each TF. This variability stems from the varying numbers of binding sites associated with different TFs, which could potentially lead to a disparity in the amount of available training data for each TF. While this diversity enriches the dataset by capturing TFs with more prominent roles, it also presents an inherent challenge in terms of data imbalance. This consideration raises thoughtful questions about how this data distribution might influence the model's performance without diminishing the study's overall contribution. In light of these considerations, several avenues for future research emerge:

1. Incorporating Longer Range Interactions: To achieve a more complete regulatory network, we would need to factor in long-range interactions between distant genomic elements like enhancers and their influence on TF binding and gene expression. This, however, would necessitate significantly larger datasets and increased computational resources.
2. Integration of Multi-Omics Data: Future research could involve incorporating multi-omics data, such as epigenomic profiles, to enhance the accuracy of TF binding predictions and provide a more comprehensive view of gene regulation.
3. Interpretable Deep Learning: Exploring methods for more interpretable deep learning models could provide insights into the decision-making process of the model, aiding in understanding the regulatory logic.

4. Translational Applications: Applying similar methodologies to human data could contribute to understanding the regulatory mechanisms underlying neurodevelopmental disorders or disease-related gene expression patterns.
5. Exploration of Non-Coding RNAs: Investigating the interplay between TFs and non-coding RNAs, such as microRNAs or long non-coding RNAs, could unveil additional layers of gene regulation.

To further facilitate progress in this field, the scripts for running the model will be made available on GitHub, enabling researchers to test their specific sequences and data. This collaborative effort can refine the model's applicability and enrich its insights. Therefore, this study provides a foundational platform for delving into cell-type specific combinatorial binding patterns of neuronal transcription factor networks. Addressing its limitations and embarking on these proposed future directions will undoubtedly unveil deeper insights into the intricate regulatory mechanisms steering neuronal development and function.

PROJECT 2

Using Single-cell RNA seq to explore bone marrow immune landscape by IL-2 therapy

Chapter 5:

5.1 Introduction

5.1.1 Overview of bone marrow and Immune System

Bone marrow (BM) is a highly specialized tissue located within the cavities of bones (National Cancer Institute, 2023). It is responsible for producing and maintaining the blood cells that are essential for life. The bone marrow contains hematopoietic stem cells (HSCs), which have the unique ability to self-renew and differentiate into all blood cell lineages, including white blood cells, red blood cells, and platelets. These cells are critical for maintaining normal physiological function and for fighting against infections and diseases. White blood cells, also known as leukocytes, are an integral part of the immune system and play a crucial role in defending the body against infections. There are several types of white blood cells, including neutrophils, eosinophils, basophils, lymphocytes, and monocytes, each with its own specific function. Red blood cells, or erythrocytes, are responsible for transporting oxygen from the lungs to the body's tissues and organs. Platelets, or thrombocytes, are essential for blood clotting and preventing excessive bleeding (Rosenberg et al., 2011) (**Figure 5.1**).

The process of blood cell production is known as haematopoiesis, which occurs in several stages. HSCs divide and differentiate into progenitor cells, which give rise to the various types of blood cells (Rothenberg, 2014).This process is tightly regulated by a complex network of signalling pathways, cytokines, and transcription factors. In addition to producing blood cells, the bone marrow also serves as a reservoir for memory B and T cells, which can provide long-term protection against infections. The bone marrow also contains stromal cells, which support the growth and differentiation of HSCs and immune cells, as well as regulate the immune response.

BM contains multiple immune cell subsets with critical functions and is considered an immune regulatory organ. It contains osteoclasts and immune cells fundamentally involved in physiological and pathological bone remodelling (Raggatt &

Partridge, 2010). In autoimmune diseases, inflammation can impair the BM niche, disturb hematopoietic and immune development, and induce osteoporosis (Mbalaviele et al., 2017).

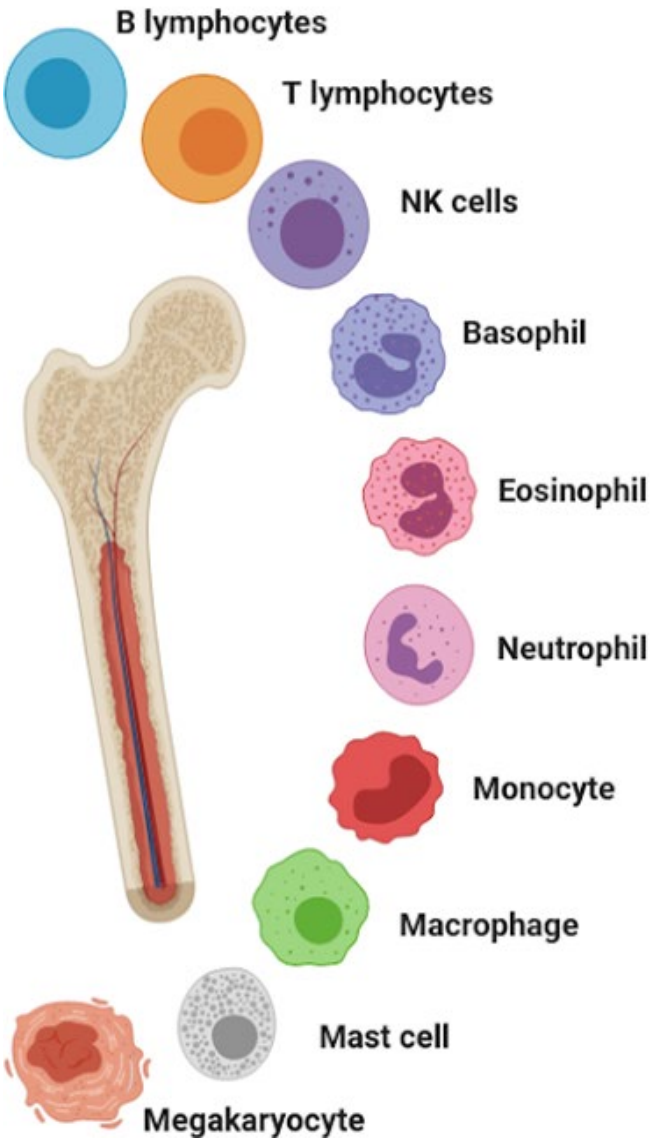


Figure 5.1: Bone marrow Immune landscape. *BM contains multiple immune cell subsets with critical functions and is considered an immune regulatory organ. It contains osteoclasts and immune cells fundamentally involved in physiological and pathological bone remodelling.*

5.1.2 Bone marrow immune landscape in health and disease

The bone marrow is a complex and dynamic system that plays a critical role in maintaining normal immune function. Altered immune cell production and function in the bone marrow can contribute to various diseases. In healthy individuals, the bone marrow maintains a complex and dynamic immune landscape, with distinct niches and cell populations that support the development, maturation, and maintenance of immune cells. However, in various diseases, including cancer and autoimmune disorders, the bone marrow immune landscape can be altered, leading to abnormal immune cell production and function. For example, in leukemia, the abnormal expansion of immature white blood cells can lead to the replacement of normal HSCs and impairment of normal immune cell production (Cheng et al., 2000). Similarly, in autoimmune disorders, such osteoporosis, there is a decrease in bone density due to the loss of bone tissue, which can lead to an increased risk of fractures. The bone marrow microenvironment also undergoes changes in osteoporosis, including alterations in the numbers and function of bone marrow immune cells, such as T cells, B cells, and osteoclasts (Tsukasaki & Takayanagi, 2019). Research has shown that the interaction between bone marrow immune cells and bone cells plays a crucial role in bone remodelling and maintenance. In osteoporosis, dysregulation of this interaction can lead to an imbalance in bone resorption and formation, contributing to bone loss (Tsukasaki & Takayanagi, 2019).

Understanding the role of the bone marrow immune landscape in osteoporosis may lead to the development of new therapies to prevent or treat bone loss. For example, targeting specific immune cells or cytokines involved in bone remodelling may help to restore the balance between bone resorption and formation and prevent or reverse bone loss(Tsukasaki & Takayanagi, 2019).

5.1.3 Interleukin-2(IL-2) therapy

Interleukin-2 (IL-2) is one of the crucial cytokines that exhibit pleiotropic effects on the immune system which is chiefly produced by CD4+ T cells, and to a lesser extent, by CD8+ T and Natural Killer cells (Boyman & Sprent, 2012). It is a 15 kDa protein containing 4-bundles of α-helices (Liao et al., 2013) and was initially discovered as a growth factor for bone-marrow-derived T lymphocytes in 1976 (Morgan et al., 1976). IL-2 plays a significant role in the maintenance of CD4+ regulatory T cells, along with the differentiation of CD4+ T cells into different subsets (Paul & Zhu, 2010). The strength of the IL-2 signal affects the differentiation process with strong signals leading to short-lived effector T cells (Kalia et al., 2010) and low-level signals leading to follicular helper (Tfh) or central memory T-cells. Discovery of IL-2 in the regulation of survival, differentiation and propagation of activated T cells paved the path for its direct clinical implications in immunotherapy.

5.1.4 Introduction to single-cell RNA sequencing

Single-cell RNA-seq (scRNA-seq) is a powerful technique that isolates single cells and captures their transcript to generate sequencing libraries wherein the mapping of transcripts is done to individual cells. It measures the distribution of expression levels for each gene across a population of cells. This accounts for unprecedented resolution and provides valuable insights into a broad spectrum of research topics like principal cell-specific changes in transcriptome, e.g. cell type identification, heterogeneity of cell responses, stochasticity of gene expression, the inference of gene regulatory networks across the cells, and comparison of diseased and healthy tissue at cellular as well as the molecular level to detect changes leading to the diseased state (Massaia et al., 2018).

scRNA-seq has several advantages over traditional RNA sequencing. First, scRNA-seq allows researchers to identify rare cell populations that may be missed in bulk sequencing data. Second, scRNA-seq can provide insights into the heterogeneity of cell populations and identify subpopulations of cells with distinct gene expression

profiles. Third, scRNA-seq can reveal cell-cell interactions and signaling pathways that are difficult to detect using traditional sequencing methods.

scRNA-seq has been particularly useful in studying immune cell heterogeneity and function, allowing for the identification of new cell types and signalling pathways involved in immune responses (Papalexi & Satija, 2018). This project is designed to address the fundamental question of how low-dose IL-2 therapy modulates BM immune landscape by comprehensively mapping the therapy-induced changes using single-cell technologies and then dissecting the underlying mechanisms by mouse models. IL-2 therapy has been utilized for treating various types of cancer, including advanced melanoma and kidney cancer, with some patients achieving long-lasting remissions (Rosenberg et al., 2011). Nonetheless, severe side effects, such as fever, fatigue, and low blood pressure, often curtail its effectiveness and tolerability (National Cancer Institute, 2021). To mitigate the adverse effects of IL-2 therapy and boost its efficacy, researchers are exploring novel strategies. For example, combining IL-2 with other immune-modulating agents, like checkpoint inhibitors, or utilizing low-dose IL-2 to activate regulatory T cells, which can inhibit the immune response and avert autoimmune reactions, may improve the safety and effectiveness of IL-2 therapy and extend its use to other cancer types (Sim & Radvanyi, 2014). Since IL-2 is capable of activating both Treg and cytotoxic lymphocytes, using low-dose IL-2 administration increases the number of only Treg cells which is beneficial for conditions like autoimmunity and chronic inflammatory diseases, whereas high-dose IL-2 treatment

expands the population of cytotoxic lymphocytes, useful for metastatic cancer (Klapper et al., 2008) **(Figure 5.2)**.

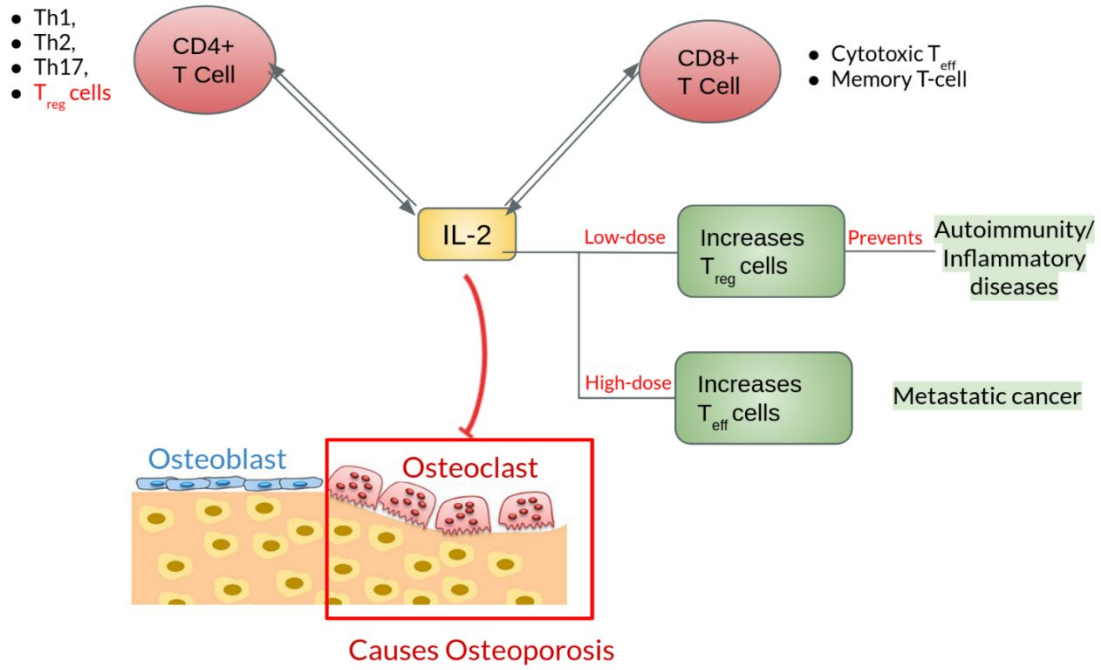


Figure 3.2: Role of Interleukin-2 in tissue-specific regulation. *Since IL-2 is capable of activating both Treg and cytotoxic lymphocytes, using low-dose IL-2 administration increases the number of only Treg cells which is beneficial for conditions like autoimmunity and chronic inflammatory diseases, whereas high-dose IL-2 treatment expands the population of cytotoxic lymphocytes, useful for metastatic cancer.*

5.2 Project design

Diseased vs treated conditions

The data for this project was provided by the collaborating lab of Dr. Di Yu. Bone marrow cells were isolated from female mice. The four different samples included- a control (sham), control treated with IL-2 (sham+IL-2), ovariectomy-induced osteoporosis (OVX) and OVX treated IL-2 (OVX+IL-2). $CD45^+$ cells were sorted from BM of control or OVX mice treated with recombinant human IL-2 (rhIL-2) for a duration of four weeks. scRNA-seq was conducted for the $CD45^+$ cells by a drop-based encapsulation approach using the 10X Genomics Chromium system.

5.3 Rationale and Objectives of the study

The aim of the proposed study is to investigate how low-dose IL-2 therapy modulates the bone marrow landscape.

Aim 1- To perform sc-RNA-seq data analysis to find different populations of immune cell types present in the bone marrow of mice.

Aim 2- To compare sub-types of cells across different conditions using single-cell data.

Aim 3- To conduct pseudotime trajectory analysis to study the cellular dynamics and for identification of branching points and the fate of different cell types.

5.4 Materials and Methods

5.4.1 Quality control of scRNA-seq data and alignment of reads

Since scRNA-seq is affected by technical noise arising due to library preparation methods, cell capture or sequencing procedure, it is essential to discard poor quality libraries before heading to the downstream analysis. I evaluated the quality metrics to diagnose the library size, the number of expressed features and the proportion of Mitochondrial genes to ensure low-quality cells are eliminated.

5.4.2 Normalization and batch correction

Batch correction is a crucial step in the analysis of scRNA-seq data. Batch effects refer to the technical variability caused by differences in the experimental conditions during sample preparation, sequencing, and data processing. Batch effects can lead to spurious correlations between cells and distort the biological interpretation of the results. To address batch effects, we used mutual nearest neighbours (MNN) approach. MNN is a data-driven method that identifies cells that are similar across batches and uses them to correct for batch effects. The MNN method detects these cells by computing the distances between cells in different batches and identifying the cells with the closest distances. The fastMNN function of the scran package was used to implement this method in our analysis.

The purpose of normalization is to remove the technical variation present in the data that is caused by factors such as differences in capture efficiency and sequencing depth between cells. Next, we performed normalization using the "scTransform" function in Seurat. The normalized expression counts were then used for downstream analyses such as dimensionality reduction and clustering.

5.4.3 Data Integration

Integrating data from multiple experiments or batches is a crucial step in the analysis of single-cell RNA sequencing data as it can increase statistical power and capture a broader range of biological heterogeneity. To integrate the data, we used the "IntegrateData" function in Seurat, which aligns the datasets based on their mutual

nearest neighbours (MNNs) and accounts for differences in the cellular composition of each dataset. We chose the top 10 principal components (PCs) for integration, and the "anchors" were identified using the "FindIntegrationAnchors" function.

5.4.4 Dimensionality reduction

The scRNA-seq data generated for this study contained thousands of genes and a large number of cells, making it a high-dimensional dataset. To handle this complexity, dimensionality reduction techniques were employed to project the data onto a lower-dimensional space while retaining the essential features. In this study, we used popular techniques such as principal component analysis (PCA), UMAP (uniform manifold approximation and projection), and t-distributed stochastic neighbour embedding (t-SNE) to visualize the data in reduced dimensions. The TSNE method captured the non-linear relationship and produced a two-dimensional visualization of distinctly isolated clusters (van der Maaten and Hinton, 2008), while UMAP overcame this and constructed a topological representation of the data.

As the integrated data contained thousands of genes and a large number of cells, we performed dimensionality reduction using the principal component analysis (PCA) algorithm. We scaled the integrated data with the "ScaleData" function and then ran PCA with the "RunPCA" function in Seurat. Next, we used the Uniform Manifold Approximation and Projection (UMAP) algorithm to visualize the data in two dimensions with the "RunUMAP" function. UMAP reduces the dimensionality of the data while preserving the local structure and distances between cells, allowing us to visualize the data in a more interpretable format. This allowed us to identify cell clusters and examine the relationships between different cell types.

5.4.5 Clustering

After dimensionality reduction and data integration, cells were clustered using Seurat's graph-based clustering method. Graph-based clustering recognizes different cell populations based on shape, size, and density (Jaitin et al., 2014). This method constructs a shared nearest neighbour graph based on Euclidean distances and uses the

Louvain algorithm to detect clusters. The clustering parameters included a resolution of 0.5 and 30 principal components for analysis.

5.4.6 Marker gene analysis and annotation of cell-types

To identify marker genes for each cluster, the FindAllMarkers function in Seurat was employed. This function performs a statistical test (Wilcoxon rank-sum test) to identify genes that are differentially expressed between cells in the cluster and cells in all other clusters. The significance level for this test was set to 0.05.. A list of previously known markers as shown in Table 5.1 was used to identify different clusters of immune cells, and unsupervised clustering methods were employed for the identification of de novo markers specific to subpopulations (Stuart et al., 2019). The closely related clusters were grouped into different immune cell types, and their annotation was validated using the SingleR package, which allows automatic annotation for scRNAseq data (Aran et al., 2019). This package compares new cells in a test dataset to a reference dataset of samples (single-cell or bulk) with known labels, identifying similarities to assign annotations.

For this project, we used the murine ImmGen data (Immunological Genome Project) as reference dataset (Heng et al., 2008), which contains labels for 20 main cell types and 253 further subtypes derived from 830 microarray samples.

5.4.7 Differential gene expression

Differential gene expression analysis is a critical step in scRNA-seq data analysis to identify genes that are significantly upregulated or downregulated in specific cell populations. This analysis helps to unravel the functional heterogeneity of different cell types and the underlying biological processes that distinguish them. In this study, we performed differential gene expression analysis using Seurat's FindMarkers function, which uses a Wilcoxon rank sum test to identify genes that are differentially expressed between two groups of cells.

First, we defined the two groups of cells to compare, such as a particular cluster of interest and the remaining cells. The FindMarkers function was then applied to identify differentially expressed genes between the two groups. The function calculates a log-fold

change and p-value for each gene and adjusts the p-value for multiple comparisons using the Bonferroni correction method.

To further refine the differential gene expression analysis, we performed a functional enrichment analysis on the identified differentially expressed genes using the gene ontology (GO) database. The GO terms provide information about the biological processes, cellular components, and molecular functions associated with the differentially expressed genes. This analysis allows us to gain insights into the underlying biological processes and functions that distinguish the different cell populations.

5.4.8 Trajectory analysis

Pseudotime trajectory analysis is a powerful tool for understanding the cellular dynamics of biological systems. In our study, we aimed to computationally reconstruct the cell trajectory and pseudotime based on scRNA-seq data to study cellular dynamics. Pseudotime is an abstract unit of progress that measures how much progress an individual cell makes while undergoing differentiation. The length of the trajectory depends on the number of significant transcriptional changes occurring during the process.

To conduct pseudotime trajectory analysis, we utilized the Monocle method, implemented using the Monocle (version 2) package in R.

The scRNA-seq data were loaded into the Monocle framework, and preprocessing steps such as size factor estimation, dispersion estimation, and gene detection were performed. Genes with a minimum expression of 0.1 were considered for downstream analysis. The differential expression analysis was conducted to identify genes associated with cell clusters. Specifically, the top 1000 ordering genes were selected based on their differential expression across clusters. Subsequently, dimensionality reduction was performed using the DDRTree method, and cells were ordered along the trajectory using the orderCells function. The resulting trajectory was visualized using the plot_cell_trajectory function, color-coded by state, pseudotime, cluster, and sample.

Monocle method proved to be a useful tool for studying the cellular dynamics of immune cell subtypes identified using graph-based clustering.

5.5 Results and Discussion

5..5.1 Identification of immune cell types in mouse bone marrow using sc-RNA-seq analysis

The first aim of the study includes the identification of subpopulations of immune cells from scRNA-seq data analysis and highlighting the diversity of various cell populations in the bone marrow niche. The analysis was divided into the following steps-

Integration of scRNA-seq data reveals novel insights into bone marrow immune cell types and their interactions.

In this project, data integration was performed to compare the compositional differences among the four single-cell datasets. Seurat's data integration and label transfer method was used, which identified shared cell states known as 'anchors' that were common among all experimental conditions. These pairwise correspondences were assumed to originate from the same biological state. The number of cells expressed in the four samples were 6162, 5036, 6267, and 5885 in Ovx, Ovx+IL2, Sham, and Sham+IL2, respectively. A total of 3000 anchors or shared features were used to combine the different experimental conditions.

Overall, data integration and unsupervised clustering allowed us to identify and characterize the different populations of immune cells present in the bone marrow of mice. These findings provide insight into the bone marrow immune landscape and can inform future studies on immune cell development and function.

Visualizing compositional differences among experimental conditions using dimensionality reduction techniques

In this study, we used popular techniques such as principal component analysis (PCA), UMAP (uniform manifold approximation and projection), and t-distributed stochastic neighbour embedding (t-SNE) to visualize the data in reduced dimensions.

The resulting plots depict each cell as a point in the reduced-dimensional space, and the distance between the points represents the degree of variation between the cells. **Figure 5.3** shows UMAP plots of all four conditions before and after data integration. As shown in the figure, the data integration process successfully combined the four datasets and enabled clear visualization of the different cell populations in the bone marrow. The plots revealed distinct clusters of immune cell types, such as T cells, B cells, and myeloid cells, in all four conditions. These results highlight the power of dimensionality reduction and visualization techniques in scRNA-seq data analysis, facilitating the identification of distinct immune cell types in the bone marrow.

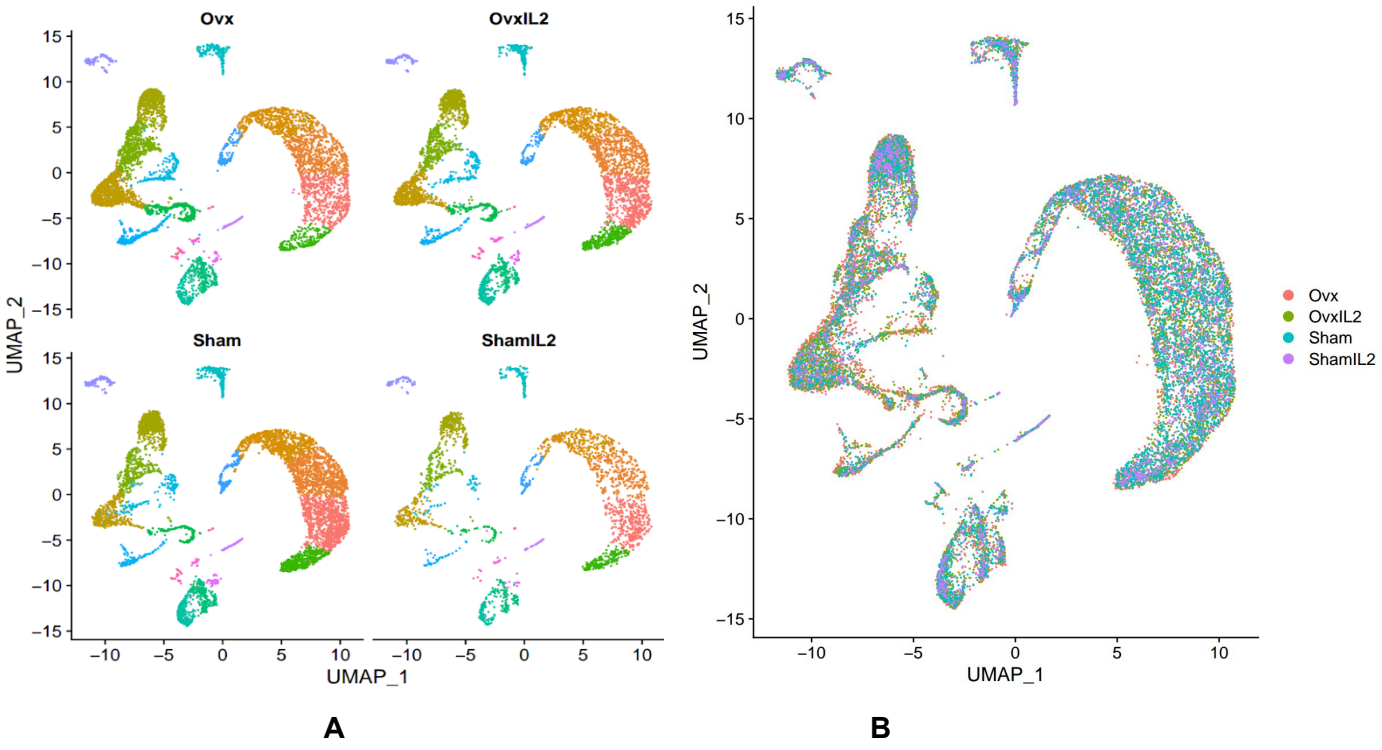


Figure 5.3: Data Integration using Seurat package. *A. UMAP projection of four experimental conditions- Ovx, Ovx +IL2, Sham and Sham + IL2 respectively coloured according to clusters. B. UMAP projection of integrated object combining all the four experimental conditions using 3000 features (anchors) where colour represents sample type.*

Clustering and Identification of subpopulations

To investigate cellular heterogeneity at single-cell resolution, I performed graph-based clustering for the identification of cells exhibiting similar gene-expression patterns. Once putative clusters were identified, selected marker genes were used for determining specific subpopulations. A total of 24 clusters were obtained **(Figure 5.4A)**, indicating the presence of diverse cell populations in the bone marrow. The major populations identified in our study were B cells, Monocyte-macrophages, T cells, Neutrophils, Basophils, Natural-killer cells, and ILC. To complement the automatic annotation, I also used established markers mentioned in **Table 5.1** for manual identification of clusters. The most prominent marker genes for each cell type were Thy1 for T cells, Cd19 for B cells, Cxcr4 for ILC2, Ncr1 for NK cells, Itgam for Monocytes, Fcer1a for Dendritic cells, Cd177 for Neutrophils, Il3ra for Eosinophils, Cd192 for Basophils, and Fcer1g for Mast cells. The combination of manual and automatic annotation allowed us to merge 24 clusters into 14 main immune cell types, enabling a more detailed and comprehensive analysis of the immune landscape in the bone marrow.

The cluster numbers 11, 16, and 19 were merged into the B cell population, while clusters 7, 10, 12, 14, and 17 were merged into Neutrophils. These two cell-types correspond to two large clusters, as visualized in **Figure 5.4B**. Cell-type compositions were compared across the four conditions as shown in Figure 5.4C, the cell type count for Neutrophils and B-cell populations significantly decreased in IL-2 treated samples. Gene expression patterns for selected genes were compared across all the samples to investigate changes in expression levels. **Figure 5.5** shows a higher expression of CD19, a B-cell marker, in IL2-treated samples compared to the diseased, indicating a higher immune response.

The second approach involved the identification of de novo markers specific to subpopulations by measuring the differential expression between clusters. Genes with higher DE are likely to fall in separate clusters, and their expression patterns can help identify subpopulations with specific functions or characteristics. A heatmap exhibiting

gene expression patterns of the top 10 highly variable genes from each cluster was plotted in **Figure 5.6**, providing a comprehensive overview of the different subpopulations present in the bone marrow. The analysis of these subpopulations can help shed light on the mechanisms underlying immune responses and inform the development of novel immunotherapeutic approaches.

Table 5.1: List of known markers for immune cell types used for annotation of clusters.

T-cells	B-cells	ILC2	NK	Monocyte	Dendritic cells	Neutrophils	Eosinophils	Basophils	Mast cells
Thy1	**Cd19**	**Cxcr4**	**Ncr1**	**Itgam**	**Fcer1a**	**Cd177**	**Il3ra**	**Cd192**	**Fcer1g**
Btla	Cd180	Cxcr6	Itga2	Cd68	Cst3	Cd93	Il5ra	Cd33	Cd9
Cd27	Ms4a2	Il17rb	Klrd1	Ly6g		Cxcr1	Siglecf	Cd38	Fcer1a
Cd4	Cd22	Il12l1	Klrk1	Tlr2		Nectin2	Cd193	Itga2b	
Cd44	Cd36	Il2ra	Klrb1c	Tnfrsf11a		Spn	Cd9	Spn	
Cd69	Cd40	Il7r	Il2rb	Tnfrsf18		Tlr2		Tlr2	
Cd7	Cd80	Ly6a	Cd226			Tlr6		Tlr4	
Cd28	Cd81	Thy1							
Cd8a	Cd86	Ccr9							
Icos	Crc2								
Il7r	Fcer1a								
Tnfrsf9	Il2ra								
Cd5									

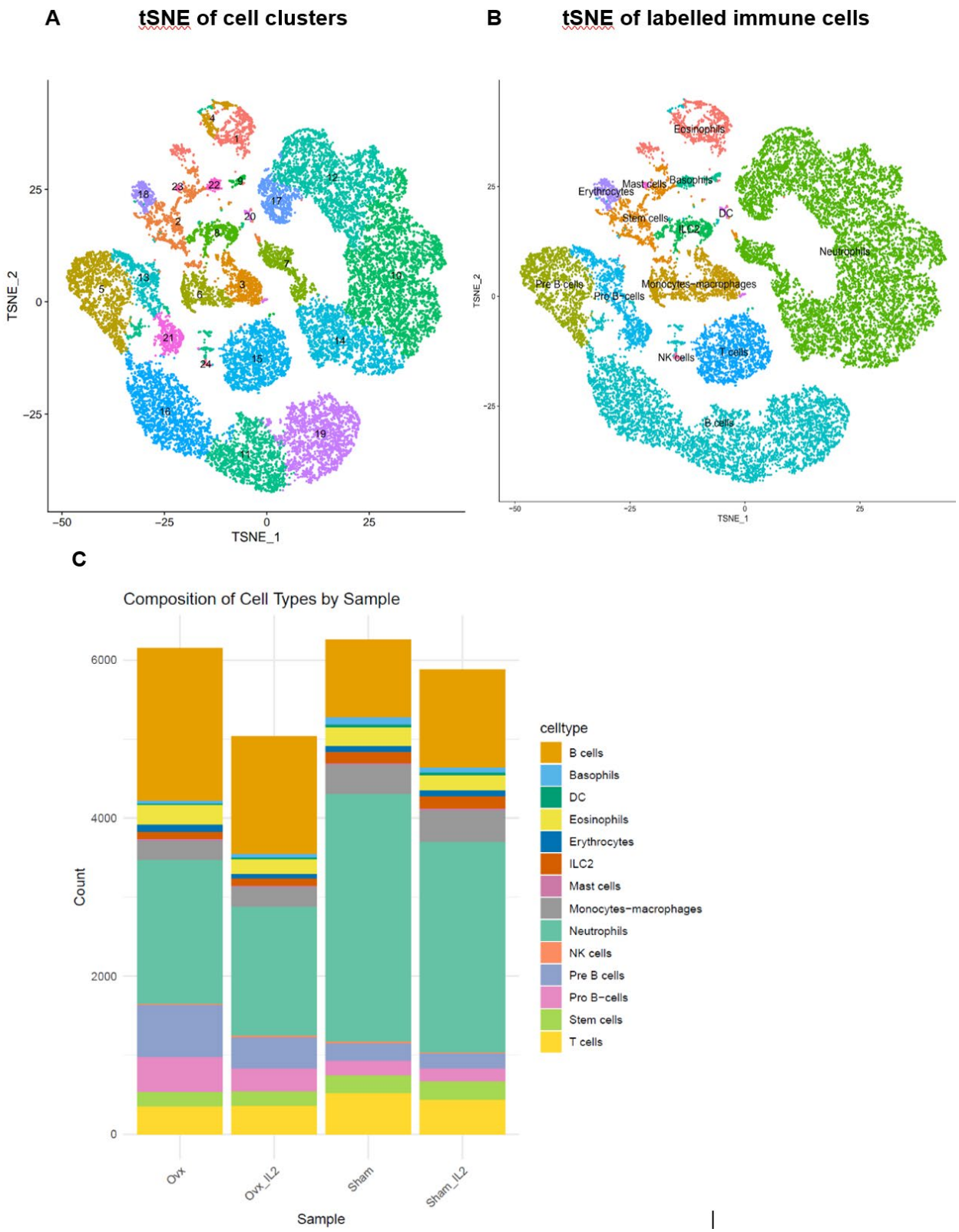
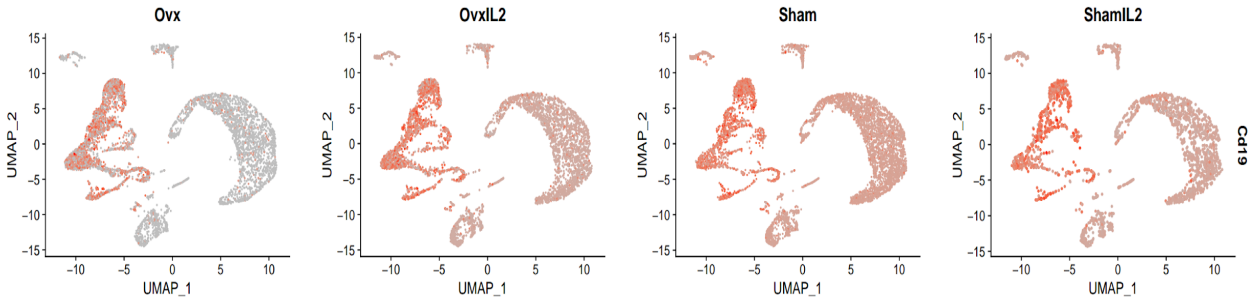

A
tSNE of cell clusters
B
tSNE of labelled immune cells
TSNE_1
TSNE_2
Eosinophils
Mast cells
Basophils
Erythrocytes
Stem cells
DC
ILC2
Neutrophils
Pre B cells
Pro B−cells
Monocytes−macrophages
NK cells
T cells
B cells
C
Composition of Cell Types by Sample
Count
Sample
Ovx
Ovx_IL2
Sham
Sham_IL2
celltype
B cells
Basophils
DC
Eosinophils
Erythrocytes
ILC2
Mast cells
Monocytes−macrophages
Neutrophils
NK cells
Pre B cells
Pro B−cells
Stem cells
T cells

Figure 5.4: Graph-based clustering. *t-SNE visualization shows single-cell transcriptomes of mice BM cells. Cells with similar gene expression are grouped together and their colour depicts the cluster to which it belongs. A. t-SNE with all 24 clusters, B. t-SNE with 24 clusters merged into 14 cell types. C. Compositions of cell-types on different conditions. This stacked barplot shows proportions of different immune cell types in the 4 conditions, depicting significant decrease in Neutrophils and B-cell populations in IL-2 treated cells.*

A. CD19 (B-cell marker) expression in different datasets



B. T cell markers expression in different datasets

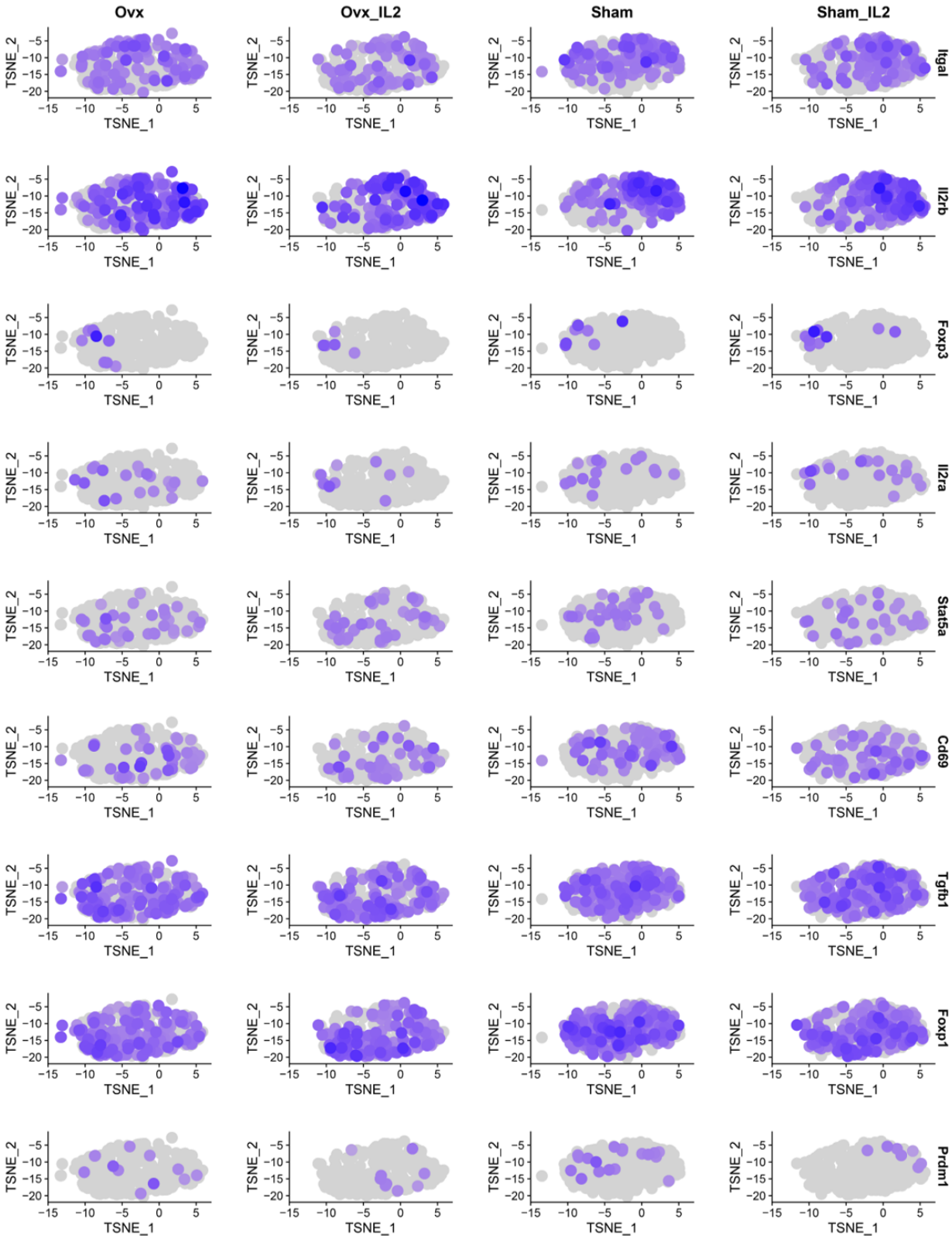


Figure 5.5: Gene Expression across different experimental conditions. *A. Feature Plot showing the expression of Cd19 (B cell marker) in all four samples. B. Feature plot of T cell subcluster showing the expression of various T cell markers such as Itgal, Ilr2b, Foxp3, Il2ra etc across 4 conditions.*

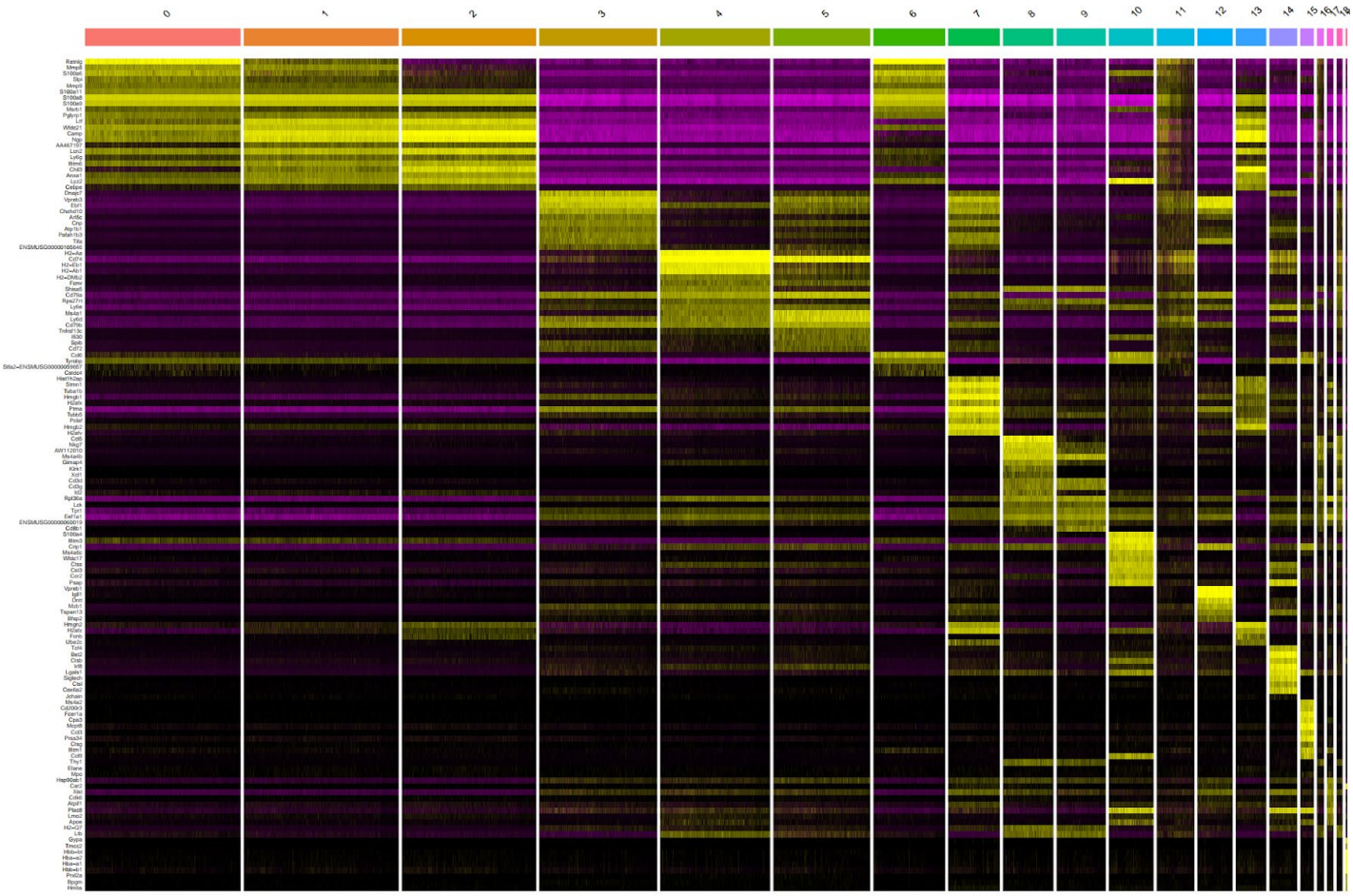


Figure 5.6: Marker Gene Identification. *Heatmap of the top 10 differentially expressed marker genes from each cluster. The column corresponds to the cells and the rows correspond to the genes. Cells are grouped by clusters. The colour scale is based on a Z-score distribution from -2 (purple) to 2 (yellow).*

5.5.2 Comparison of cell subtypes across different conditions uncovered differentially expressed genes and pathways.

To compare the different cell subtypes obtained in the previous step across different conditions, a differential expression (DE) analysis was conducted. This analysis is essential for identifying quantitative differences between different groups or conditions and can help elucidate the molecular basis of phenotypic variation. To conduct the DE analysis, pairwise cluster analysis was used to compare the gene expression patterns between closely associated clusters. By comparing the expression levels of genes in different conditions, we were able to identify important genes and pathways that may be responsible for the observed differences in cell subtypes. DE analysis provided valuable insights into the molecular mechanisms underlying the phenotypic differences observed across different conditions.

Differential Expression between Clusters

Statistical methods such as the Wilcoxon test was employed to compare each cluster against all other clusters of cell subpopulations. To ensure the robustness of the DE signature, each gene was tested for differential expression between the chosen cluster and every other cluster in the dataset, and their expression profiles were examined. The resulting DE gene signatures were then used to perform pathway enrichment analysis using methods such as gene ontology and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis to identify biological processes and pathways that were significantly enriched in each cluster. These analyses provided insights into the functional roles of different cell types and the molecular mechanisms that drive their phenotypic diversity across different conditions. The DE results and enriched pathways for each cluster are summarized in **Figure 5.7**.

Differential Expression across Conditions

To further elaborate, differential expression analysis allowed us to identify genes that are differentially expressed between two conditions or groups.. I performed comparisons between different experimental conditions, namely Ovx vs Sham and Ovx vs Ovx + IL2, for four main cell types. This enabled us to identify genes that are differentially expressed between the control and diseased groups, as well as between the diseased and treatment groups.

To carry out these comparisons, first I utilized a comparative approach to identify differentially expressed genes across various conditions in B cells, specifically focusing on the contrasts between Ovx vs. Ovx_IL2 and Ovx vs. Sham (control). Initially, I integrated single-cell RNA sequencing data from these conditions. By subsetting the data to focus on B cells, I set the cell identities to their respective sample labels (Sample here refers to the condition- OVX, OVX_IL2, Sham and Sham_IL2), enabling pseudo-bulk analysis. I calculated the average expression levels for each gene across samples and applied a log transformation to normalize the data. Next, I created a new metadata column combining cell type and condition information and set this as the current identity class. Using the FindMarkers function, I identified genes that were differentially expressed between the conditions of interest. Seurat's standard method for DE analysis of the same cell type in two conditions considers each cell as a sample. As an alternative method, I used edgeR to find differentially expressed genes between the conditions of interest using pseudobulk approach. In the design matrix, the columns represented the 4 conditions and rows corresponded to cells (instead of samples). The two contrast groups were- Control (Sham) vs Ovx (Disease) and Ovx (Disease) vs Ovx_IL2 (treated with IL2).The results of differential expression analysis can be visualized in Figure 5.7A, which shows that the number of significant DE genes in the diseased vs treatment group is less compared to diseased vs Control. These findings suggest that the low dose of IL2 treatment has a regulatory effect on gene expression, leading to a reduction in the number of significantly differentially expressed genes.

The list of differentially expressed genes obtained from pairwise comparisons was further analysed using pathway/gene set analysis to identify the biological pathways that are differentially regulated between the different conditions. For instance, we found that the osteoclast differentiation pathway was downregulated in the Sham group but upregulated in the Ovx group, indicating that osteoporosis promotes osteoclast differentiation and bone resorption. However, when treated with a low dose of IL2, the osteoclast differentiation process was downregulated, thereby preventing bone resorption. These findings provide insights into the molecular mechanisms underlying osteoporosis and its treatment. The list of differentially expressed genes obtained from the comparison was then used for pathway and gene set analysis. This analysis helped to highlight whether the marker genes were responsible for upregulation or downregulation of significant pathways. Interestingly, it was observed that the osteoclast differentiation pathway was downregulated in the control group (Sham) whereas upregulated in the diseased group (Ovx). This finding is significant as it explains how osteoporosis promotes osteoclast differentiation, leading to bone resorption. However, when treated with a low dose of IL2, the osteoclast differentiation process is downregulated, thereby preventing bone resorption (as depicted in **Figure 5.7B).**

The findings from our differential expression and pathway analysis contribute to a more nuanced understanding of how immune modulation can influence bone health. Comparing our results with existing literature, it appears that IL2's role may extend beyond traditional immunological functions, also implicating it in bone metabolism and the pathophysiology of diseases like osteoporosis. These results underscore IL2's potential as a therapeutic agent for preventing bone loss in osteoporosis patients. Future studies should explore the dose-response relationship of IL2 in osteoporosis and other bone-related disorders to optimize therapeutic strategies. Furthermore, our use of pathway and gene set analysis has broadened our understanding of the biological mechanisms underlying these differential expression patterns, enhancing our ability to target these pathways in disease treatment.

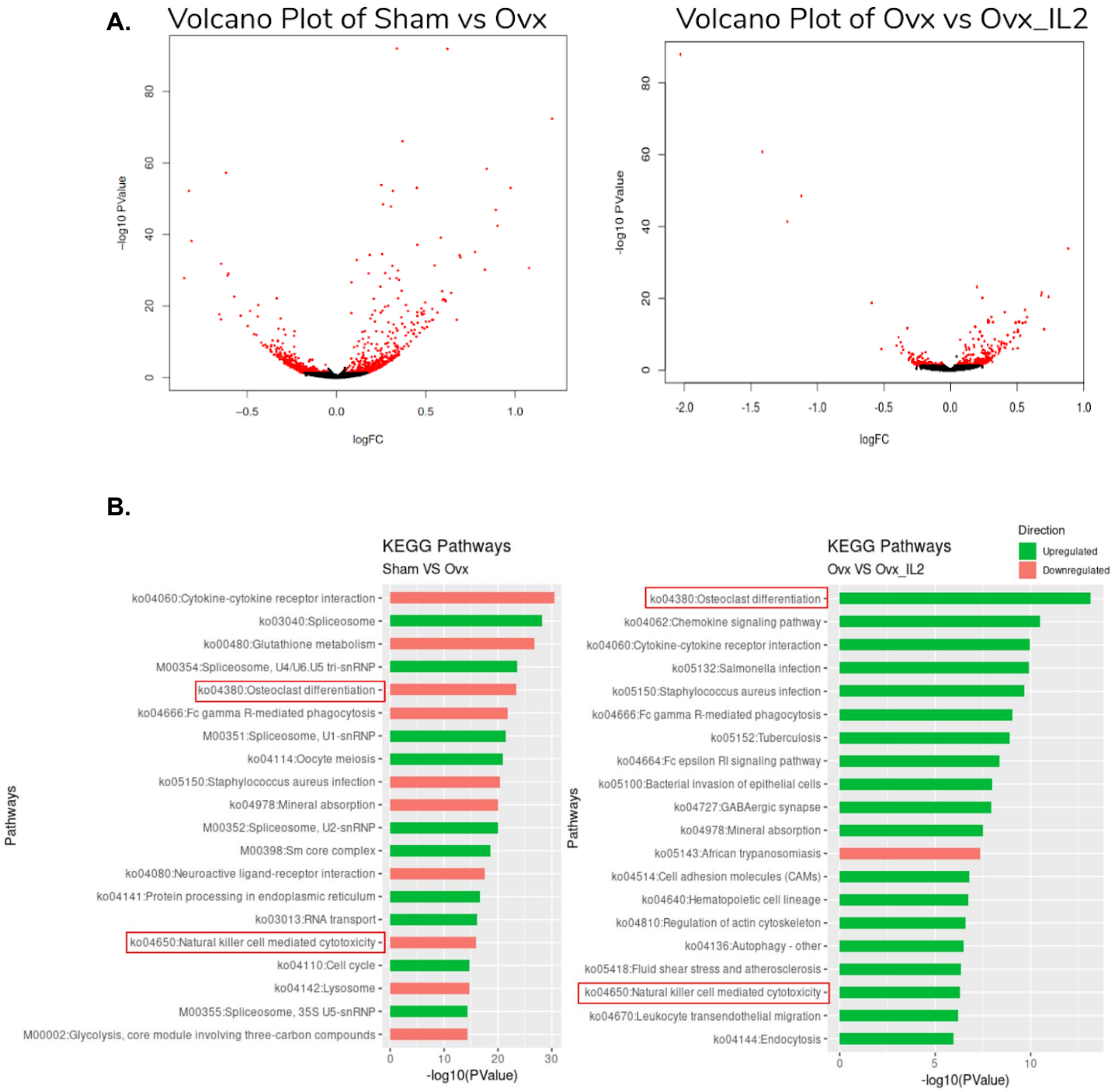


Figure 5.7: Differential Expression results across conditions*. A. Volcano Plots showing significant DE genes in B cells in red for both comparisons- Sham vs Ovx and Ovx vs Ovx +IL2. In these plots, the significant differentially expressed genes are highlighted in red. The x-axis of the volcano plot represents the log2 fold change between the two conditions, indicating the magnitude of difference in gene expression. The y-axis represents the -log10 of the p-value from the differential expression test, indicating the statistical significance of the difference. Thus, the red points in the upper left and right of the plot represent genes that are significantly upregulated or downregulated, respectively. B. Kegg Pathways for both comparisons showing upregulated pathways in green while down-regulated pathways in red arranged in decreasing order of their log P Values, which suggests that the pathways at the top of the list have the most significant changes in gene expression. This part of the figure provides an overview of the biological processes that are most affected by the condition.*

5.5.3 Pseudotime trajectory analysis and fate mapping of cell populations.

Pseudotime trajectory analysis revealed cellular dynamics and lineage in our data The lineage of the B-cell population, which includes immature B cells, Pro B-cells, and Mature B-cells, was observed by deriving the trajectory for B-cell populations. Clusters 5, 11, 13, 16, 19, and 21, corresponding to B-cells, were subsetted, and trajectories were conducted using Monocle v2. The trajectory analysis revealed that clusters 5 and 13 are the starting point of the trajectories and are referred to as immature or Pre/Pro B-cells. The first branching point revealed that cluster 21 and 16 are next in the developmental phase, followed by the second branching with clusters 11 and 19 deemed as Mature B-cells. This result correlates with our previous understanding and identification of clusters **(Figure 5.8)**.

Similar analyses were conducted for four other cell populations - Neutrophils, Monocytes-macrophages, T-cells, and ILC2 - to gain insights into their direction of development. The pseudotime trajectory analysis provided valuable information about the cellular dynamics of these cell types, including the identification of branching points and the fate of different cell types. These results contribute to a better understanding of the molecular mechanisms that govern cellular differentiation and development.

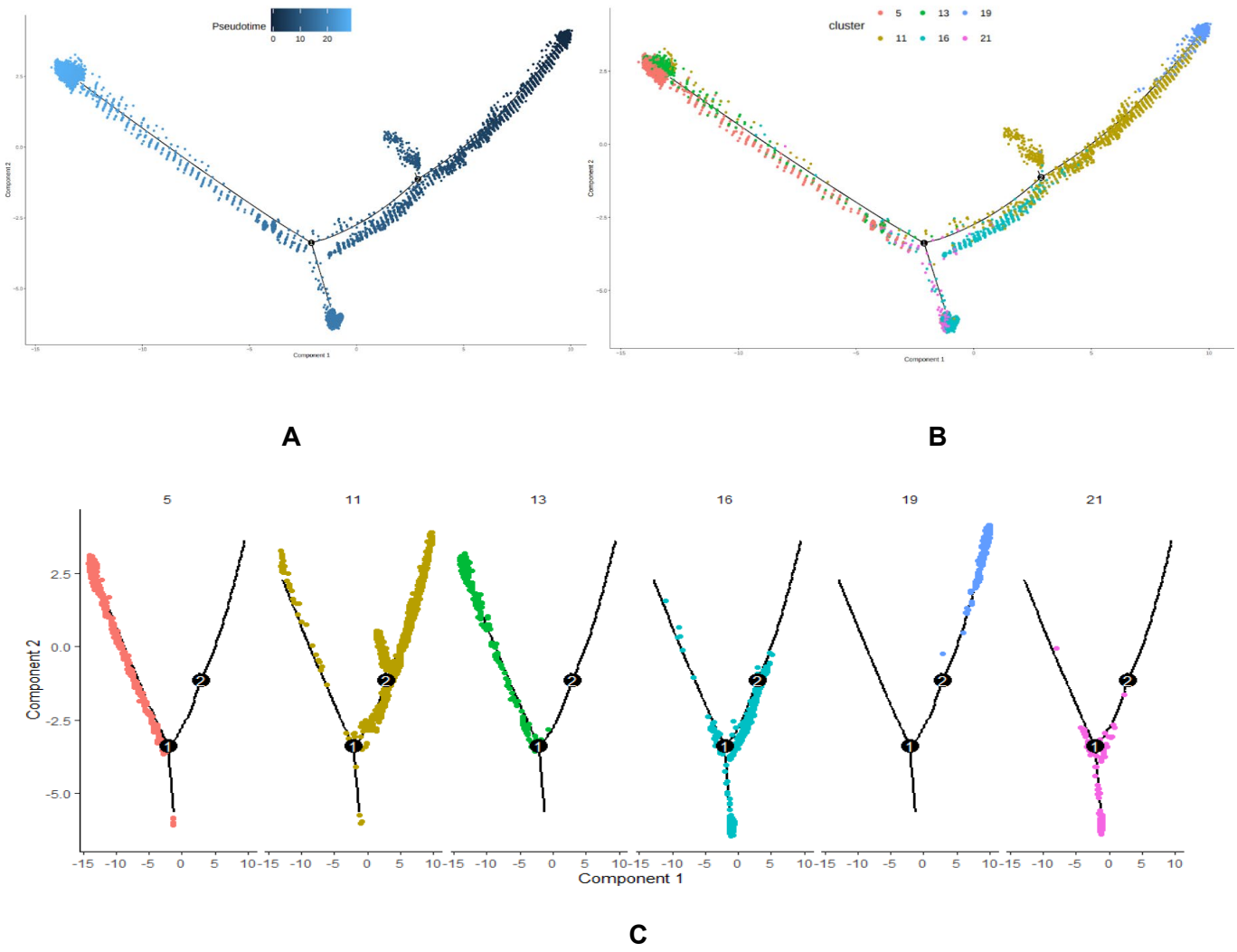


Figure 5.8: Pseudotime Trajectory Analysis of B-cell Population. *Clusters identified as B-cells (5, 11, 13, 16, 19 and 21) were extracted from the data object to determine their trajectory. Each point represents a single cell. A. B cells ordered by their pseudotime with 0 or dark blue representing the starting point of differentiation and the colour lightens as the cells mature. B. Development trajectory of B cells with colours representing the cluster to which it belongs, C. Individual clusters and their branching point.*

5.6 Conclusion

This project aimed to investigate how low-dose IL-2 therapy modulates the immune landscape of the bone marrow (BM) using single-cell RNA sequencing (scRNA-seq). The scRNA-seq data of CD45+ cells from BM samples of mice were analyzed to identify different immune cell types and compare their expression across four experimental conditions: control (sham), control treated with IL-2 (Sham + Treatment), ovariectomy-induced osteoporosis (OVX), and OVX treated with IL-2 (OVX + Treatment). The analysis pipeline included quality control, normalization, data integration, dimensionality reduction, clustering, and downstream processing. Pairwise differential expression analysis confirmed the identified cell types and their assignment into clusters. Trajectory analysis using pseudotime analysisrevealed the cellular dynamics and lineage of subpopulations in B cells.

The immune landscape of the bone marrow plays a critical role in bone remodeling and immune regulation. In autoimmune diseases and inflammation, the BM niche can be disrupted, affecting hematopoietic and immune development and leading to osteoporosis. Interleukin-2 (IL-2) is a cytokine with diverse effects on the immune system.

In this project, scRNA-seq data from BM CD45+ cells were analysed to understand how low-dose IL-2 therapy modulates the immune landscape. The analysis pipeline involved several steps, including quality control to ensure data reliability, normalization to account for technical variations, data integration to compare different experimental conditions, dimensionality reduction techniques to visualize the data, clustering to identify distinct cell types, and downstream processing for further analysis.

One of the key findings from our computational analysis was the decrease in Neutrophils and B cell populations in IL2 treated samples. This suggests that low-dose IL-2 therapy may modulate the immune response by affecting these cell populations. This finding is crucial because it indicates that IL-2, traditionally known for its role in T cell biology, particularly in expanding regulatory T cells (Tregs), may also influence other aspects of the immune system. Neutrophils are primary innate immune cells known for

their roles in acute inflammation and defence against infections. Their reduction upon IL-2 treatment could suggest a potential dampening of inflammatory responses, which is beneficial in diseases where chronic inflammation leads to tissue damage. For instance, excessive neutrophil activity is linked to various inflammatory and autoimmune diseases (such as rheumatoid arthritis), and their decrease might mitigate such inflammatory responses, as suggested by studies like Mantovani *et al.* (2011).

Similarly, B cells are central to adaptive immunity, responsible for antibody production, antigen presentation, and cytokine secretion. They also play critical roles in both humoral immunity and in the pathogenesis of several autoimmune diseases through the production of autoantibodies. The reduction in B cell populations could suggest that IL-2 therapy may help reduce the autoantibody-mediated components of autoimmune diseases. This is aligned with research indicating that modulation of B cell activity can be beneficial in conditions like systemic lupus erythematosus (SLE) as reviewed by Mackay & Browning (2002).

Furthermore, the specific decrease in these cell types underlines the possibility of IL-2's role in rebalancing the immune system towards a more regulatory and less inflammatory state. This is particularly relevant in the context of osteoporosis and other bone marrow niche disorders, where inflammatory processes can disrupt bone remodeling and homeostasis. By reducing the populations of cells that contribute to inflammation (neutrophils) and autoimmunity (B cells), IL-2 therapy might foster an environment more conducive to bone health and less prone to the detrimental effects of immune dysregulation.

Therefore, these results not only expand our understanding of IL-2's effects beyond T cell populations but also open up potential therapeutic avenues for using IL-2 to treat a broader range of immune-related conditions by targeting multiple cell types within the immune system. The pairwise differential expression analysis confirmed the presence of different immune cell types and their assignment into clusters. This analysis provided insights into the changes in gene expression between different experimental conditions, highlighting the impact of low-dose IL-2 therapy on immune cell populations

in the bone marrow.. By analysing the trajectory of specific cell populations, such as T cells or B cells, it is possible to gain insights into their development and differentiation processes.

Overall, this project contributes to our understanding of how low-dose IL-2 therapy modulates the immune landscape of the bone marrow. The analysis of scRNA-seq data provides valuable insights into the changes in gene expression and cellular dynamics, shedding light on the underlying mechanisms of IL-2 therapy and its potential implications for autoimmune diseases and chronic inflammation.

. IL-2 exhibits pleiotropic effects on the immune system, and its role in immune cell function and response has been extensively investigated (Boyman et al., 2006). Our findings of differential gene expression and enriched pathways in response to IL-2 treatment further support its regulatory role in immune cell function. The identification of these gene expression changes provides valuable insights into the molecular mechanisms underlying the immunomodulatory effects of IL-2 in the bone marrow.

The comparison of different experimental conditions, including control, IL-2-treated control, ovariectomy-induced osteoporosis, and IL-2-treated ovariectomy, has allowed us to assess the effects of low-dose IL-2 therapy on the bone marrow immune landscape. Ovariectomy-induced osteoporosis serves as a model for understanding the pathological changes in the bone marrow niche. By examining the immune cell populations and gene expression changes in response to IL-2 treatment, we have gained insights into how IL-2 modulates immune cell populations in the context of bone remodelling and diseases like osteoporosis. These findings contribute to our understanding of the potential therapeutic applications of IL-2 in managing bone-related disorders.

Our study adds to the growing body of literature on single-cell transcriptomics and its applications in investigating immune cell types and their responses in various conditions (Poulin et al., 2016). By integrating scRNA-seq data and employing advanced analysis techniques, we have provided valuable insights into the immune landscape of the bone marrow and the effects of IL-2 therapy. The application of single-cell technologies in immunological research holds great promise for unravelling the

complexities of the immune system and informing the development of targeted therapies for immune-related disorders.

Chapter 6: Identification of Alternative Splicing and Polyadenylation in RNA-seq Data

6.1 Preface

This chapter predominantly features a publication in ***Journal of Visualized Experiments***, for which I had the privilege of being the first author. My contribution to this research was multi-faceted, including identifying the appropriate RNA-seq dataset for alternative splicing analysis, structuring and drafting the initial manuscript, conducting a thorough search for relevant methods in alternative splicing analysis and benchmarking them, performing the alternative splicing analysis, creating and assembling the figures, and editing subsequent versions of the manuscript. I was also responsible for responding to feedback from external peer reviewers, with valuable guidance from my supervisor, Associate Professor Jean Wen.

ABSTRACT:

As well as the typical analysis of RNA-Seq to measure differential gene expression (DGE) across experimental/biological conditions, RNA-seq data can also be utilized to explore other complex regulatory mechanisms at the exon level. Alternative splicing and polyadenylation play a crucial role in the functional diversity of a gene by generating different isoforms to regulate gene expression at the post-transcriptional level and limiting analyses to the whole gene level can miss this important regulatory layer. Here, we demonstrate detailed step by step analyses for identification and visualization of differential exon and polyadenylation site usage across conditions, using Bioconductor and other packages and functions, including DEXSeq, diffSplice from the Limma package, and rMATS.

Dixit, G., Zheng, Y., Parker, B. and Wen, J., 2021. Identification of Alternative Splicing and Polyadenylation in RNA-seq Data. JoVE (Journal of Visualized Experiments), (172), p.e62636.

6.2 Introduction

RNA sequencing (RNA-seq) has been a widely used technology in transcriptome analysis over the years, primarily for estimating differential gene expression (DGE) and gene discovery purposes (Wang et al., 2009). However, in addition to these conventional applications, RNA-seq can also be leveraged to gain insight into the regulatory mechanisms that govern gene expression at the post-transcriptional level by measuring exon-level expression patterns (Merkin et al., 2012). Alternative splicing (AS) is one such mechanism that plays a crucial role in the functional diversity of eukaryotic genes by generating different mRNA isoforms using different splice sites (Wang et al., 2008). AS events can be divided into different patterns: skipping of complete exons (SE) where a ("cassette") exon is completely removed out of the transcript along with its flanking introns; alternative (donor) 5' splice site selection (A5SS) and alternative 3' (acceptor) splice site selection (A3SS) when two or more splice sites are present on either end of an exon; retention of introns (RI) when an intron is retained within the mature mRNA transcript and mutual exclusion of exon usage (MXE) where only one of the two available exons can be retained at a time (Merkin et al., 2012; Wang et al., 2008). Alternative polyadenylation (APA) also plays an important role in regulating gene expression using alternative poly (A) sites to generate multiple mRNA isoforms from a single transcript (Elkon et al., 2013). Most polyadenylation sites (pAs) are located in the 3' untranslated region (3' UTRs), generating mRNA isoforms with diverse 3' UTR lengths. As the 3' UTR is the central hub for recognizing regulatory elements, different 3' UTR lengths can affect mRNA localization, stability and translation (Mayr & Bartel, 2009). There are a class of 3' end sequencing assays optimized to detect APA that differ in the details of the protocol (Lianoglou et al., 2013).

In this study, we presented a pipeline of differential exon analysis methods, which can be divided into two broad categories: exon-based (DEXSeq) (Anders & Huber, 2010; Love et al., 2014) and event-based (replicate Multivariate Analysis of Transcript Splicing [rMATS]; (Shen et al., 2014). The exon-based methods compare the fold change across conditions of individual exons, against a measure of overall gene fold change to call differentially expressed exon usage, and from that compute a gene-level measure of AS

activity. Event-based methods use exon-intron-spanning junction reads to detect and classify specific splicing events such as exon skipping or retention of introns, and distinguish these AS types in the output (Wang et al., 2008). Thus, these methods provide complementary views for a complete analysis of AS. We selected DEXSeq (based on the DESeq2 DGE package) and diffSplice (based on the Limma DGE package) for the study as they are amongst the most widely used packages for differential splicing analysis (Anders & Huber, 2010; Ritchie et al., 2015). rMATS was chosen as a popular method for event-based analysis (Shen et al., 2014).

The RNA-seq data used in this study was acquired from the Gene Expression Omnibus (GEO) (GSE138691) (Sznajder et al., 2020). The data consisted of mouse RNA-seq data with two condition groups: wild-type (WT) and Muscleblind-like type 1 knockout (Mbnl1 KO) with three replicates each. To demonstrate differential polyadenylation site usage analysis, we obtained mouse embryo fibroblasts (MEF) PolyA-seq data (GEO Accession GSE60487) (Batra et al., 2014). The data had four condition groups: Wild-type (WT), Muscleblind-like type1/type 2 double knockout (Mbnl1/2 DKO), Mbnl 1/2 DKO with Mbnl3 knockdown (KD), and Mbnl1/2 DKO with Mbnl3 control (Ctrl). Each condition group consisted of two replicates.

6.3 Results and Discussion

The AS analysis conducted in this study identified exons with differential usage across the conditions, and genes that exhibited significant overall splicing activity of one or more of its constituent exons, ranked by statistical significance. The authors of the original study (Sznajder et al., 2020) experimentally validated the differential splicing patterns in the top-ranking genes from our analysis, including Mbnl1, Tcf7, and Lef1 as depicted in Figures 6.1 and 6.2. Additionally, the authors (Sznajder et al., 2020) confirmed the differential splicing patterns in five genes, namely Mbnl1, Mbnl2, Lef1, Tcf7, and Ncor2, through experimental validation. Our approach successfully detected differential splicing patterns in all the above-mentioned genes.

We present the FDR levels obtained for each gene using DEXSeq, diffSplice, and rMATS, respectively, as shown in Supplementary Tables 1-3: Mbnl1 (0, 6.6E-61, 0), Mbnl2 (0, 0.18, 0), Lef1 (1.4E-10, 1.3E-04, 0), Tcf7 (0, 1.1E-6, 0), and Ncor2 (9.2E-11, 1.6E-06, 0). These results demonstrate the effectiveness of the protocol in identifying differential splicing events in RNA-seq data.

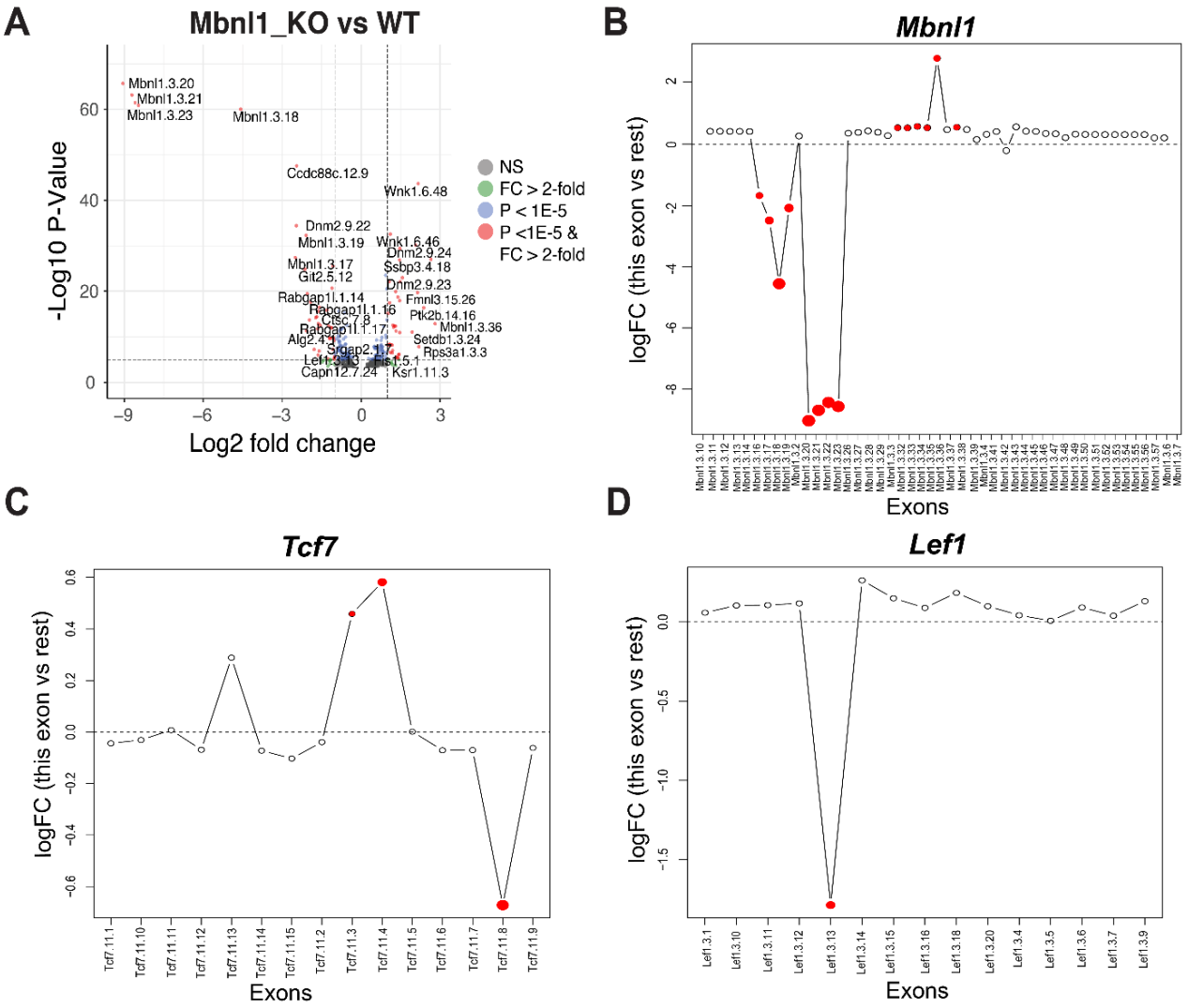


Figure 6.1: RNA-Seq analysis of differential splicing with Limma diffSplice. A. Volcano plot *of RNA-Seq from Limma diffSplice analysis: The x-axis shows log exon fold change; the y-axis shows -log$_{10}$ p-value. Each point corresponds to an exon. Horizontal dashed line at p-value=1E-5; vertical dashed lines at two-fold change (FC). Red exons show substantial FC and statistical significance.* ***B-D. Differential splicing plots****: Splicing patterns are exhibited for example genes* ***Mbnl1, Tcf7,*** *and* ***Lef1****, respectively, where the x-axis shows exon id per transcript; the y-axis shows exon relative log fold change (the difference between the exon's logFC and the overall logFC for all other exons). Exons highlighted in red show statistically significant differential expression (FDR < 0.1).*

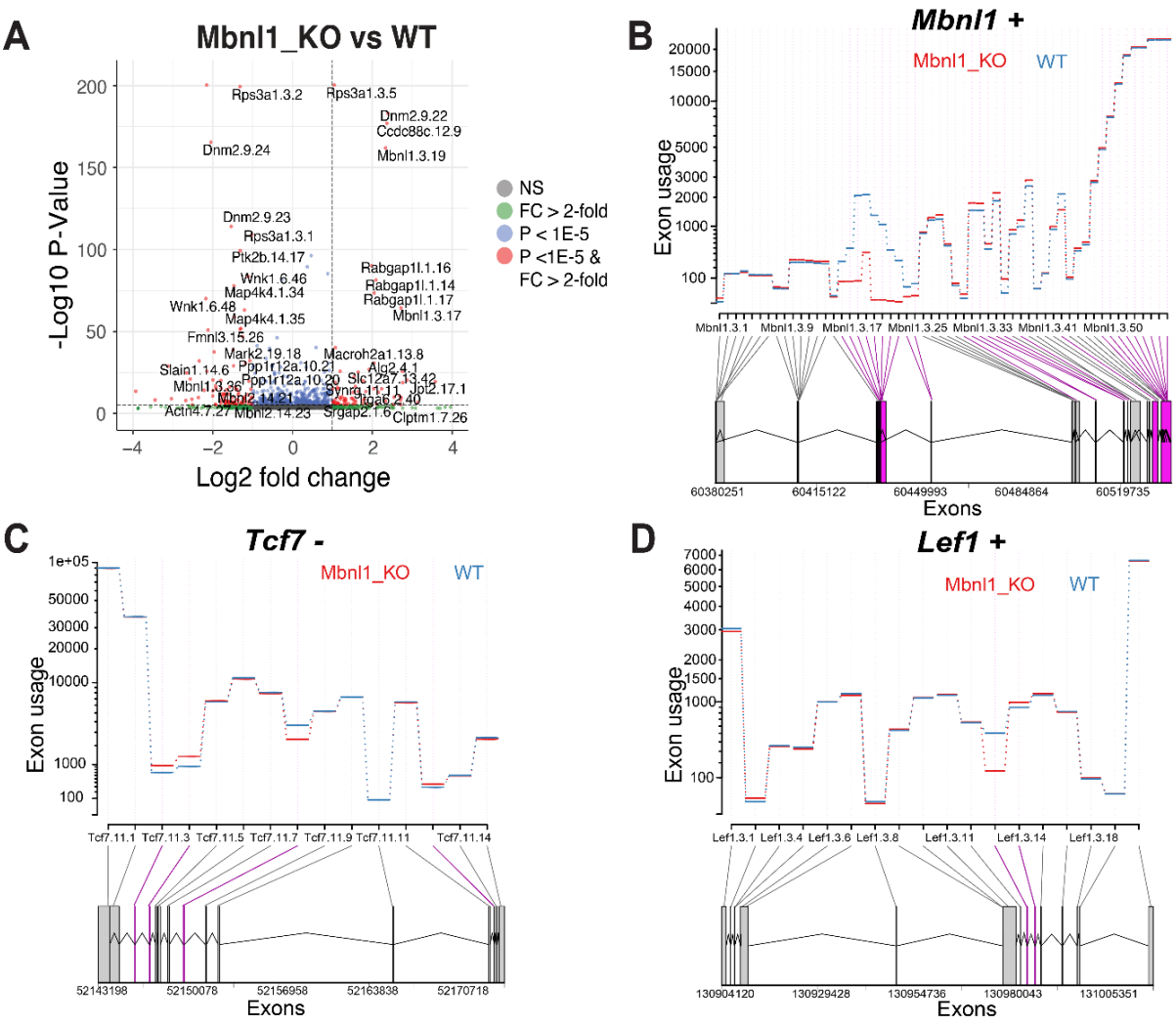


Figure 6.2: RNA-Seq analysis of differential exon usage with DEXSeq. *A. Volcano plot. B-D. Differential exon usage of RNA-Seq obtained from DEXSeq analysis. Genes Mbnl1, Tcf7, and Lef1, respectively, show significant differential usage of exons between WT and Mbnl1 knock-out highlighted in pink, which correspond to the same differential exons in figure 5.1.*

Figure 6.3 displays a comparison between outputs obtained from three different tools and illustrates alternative splicing patterns in *Wnk1* gene. In the volcano plots, the y-axis shows statistical significance (-log10 P-values) and the x-axis shows effect size (fold change). Genes located at the top left or right quadrants indicate substantial FC and strong statistical evidence of genuine differences.

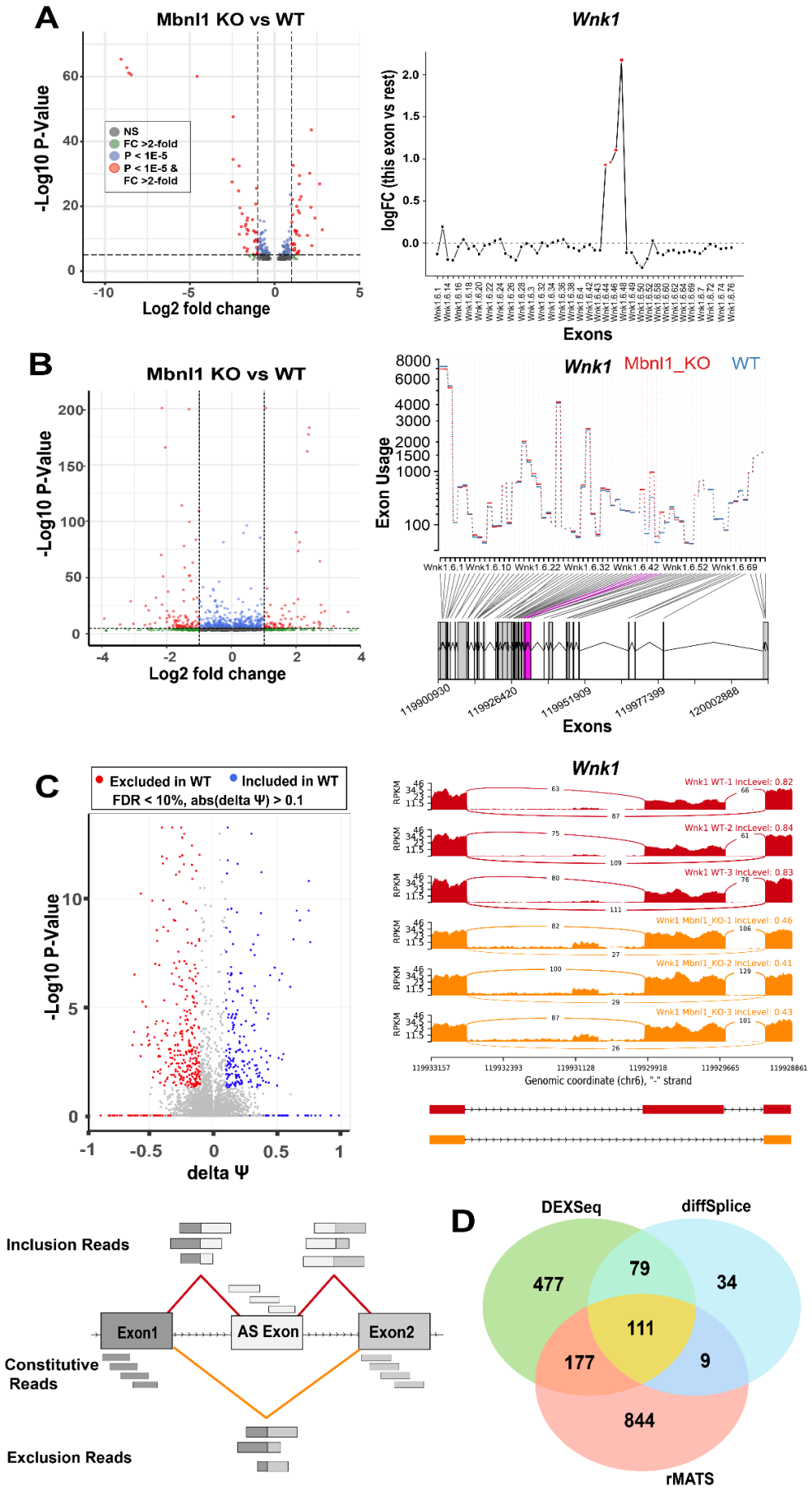

A
Mbnl1 KO vs WT
-Log10 P-Value
Log2 fold change
NS
FC >2-fold
P < 1E-5
P < 1E-5 & FC >2-fold
Wnk1
logFC (this exon vs rest)
Exons
B
Mbnl1 KO vs WT
-Log10 P-Value
Log2 fold change
Wnk1
Mbnl1_KO
WT
Exon Usage
Exons
C
Excluded in WT
Included in WT
FDR < 10%, abs(delta Ψ) > 0.1
-Log10 P-Value
delta Ψ
Wnk1
Wnk1 WT-1 IncLevel: 0.82
Wnk1 WT-2 IncLevel: 0.84
Wnk1 WT-3 IncLevel: 0.83
Wnk1 Mbnl1_KO-1 IncLevel: 0.46
Wnk1 Mbnl1_KO-2 IncLevel: 0.41
Wnk1 Mbnl1_KO-3 IncLevel: 0.43
RPKM
Genomic coordinate (chr6), "-" strand
Inclusion Reads
Exon1
AS Exon
Exon2
Constitutive Reads
Exclusion Reads
D
DEXSeq
diffSplice
rMATS
477
79
34
111
177
9
844

Figure 6.3: Comparison of alternative splicing results obtained from diffSplice, DEXSeq and rMATS.

A. Volcano plot (left) of RNA-Seq from Limma diffSplice analysis: The x-axis shows log exon fold change; the y-axis shows -log10 p-value. Each point corresponds to an exon. Horizontal dashed line at p-value=1E-5; vertical dashed lines at two-fold change (FC). Red exons show substantial FC and statistical significance. Differential splicing plot (right): Splicing patterns are exhibited for an example gene Wnk1 where the x-axis shows exon id per transcript; the y-axis shows exon relative log fold change (the difference between the exon's logFC and the overall logFC for all other exons). Exons highlighted in red show statistically significant differential expression (FDR < 0.1).

B. Volcano plot (left) and Differential exon usage (right) of RNA-Seq obtained from DEXSeq analysis. Wnk1 gene shows significant differential usage of exons between WT and Mbnl1 knock-out highlighted in pink, which correspond to the same differential exons in (A).

C. Volcano plot (left) and Sashimi plot (right) for Wnk1 obtained from rMATS analysis. Volcano plot depicting significant skipped (cassette) exon (SE) event in wild type compared to knockout at 10% FDR with change in percent spliced in (PSI or ΔΨ) values > 0.1. The x-axis shows change in PSI values across conditions, and the y-axis shows log P-Value. The Sashimi plot shows a skipped exon event in the Wnk1 gene, corresponding to a significant differential exon in (A) and (B). Each row represents an RNA-Seq sample: three replicates of wild-type and Mbnl1 knock-out. The height shows read coverage in RPKM and the connecting arcs depict junction reads across exons. Annotated gene model alternative isoforms are shown at the bottom of the plot. The bottom panel of C illustrates junction reads used to compute the PSI statistic.

D. Venn diagram showing the number of significant differentially spliced exons.

Figure 6.3A (right-panel) shows a diagrammatic display of exon differences of one of the top ranked genes, showing logFC on the y-axis and exon number on the x-axis. This example shows a cassette exon varying between conditions for gene *Wnk1*. The differential exon usage plot from DEXSeq shows evidence of differential splicing at five exon sites near Wnk1.6.45. The highlighted exons in pink are likely to be spliced out in Mbnl1 KO samples compared to WT. These exons are complementary to the results obtained by diffSplice which shows a similar pattern at the specific genomic position. **Figure 6.3C** shows a volcano plot of genes that were alternative spliced, with the effect size measured by the differential “percent spliced in” (PSI or ΔΨ) statistic.

PSI refers to the percentage of reads consistent with the inclusion of a cassette exon (i.e., reads mapping to the cassette exon itself or junction reads overlapping the exon) compared with reads consistent with exon exclusion i.e. junction reads across adjacent upstream and downstream exons (The bottom panel of **Figure 6.3C**). The right

panel of **Figure 6.3C** shows sashimi plot of *Wnk1* gene with differential splicing event overlaid on coverage tracks for the gene, with a skipped exon in *Mbnl1* KO. Arcs joining exons show the number of junction reads. **Figure 6.3D** compares the significant differentially spliced exons detected by the three tools.

Figure 6.4 illustrates types of splicing events SE, A5SS, A3SS, MXE and RI with the help of Sashimi plots of the top significant genes of those events. On comparing the three replicates of both WT and Mbnl1_KO, a total of 1272 SE events, 130 A5SS, 116 A3SS, 215 MXE and 313 RI events were detected at FDR 10%. Sashimi plot illustrates the type of event using top genes as an example.

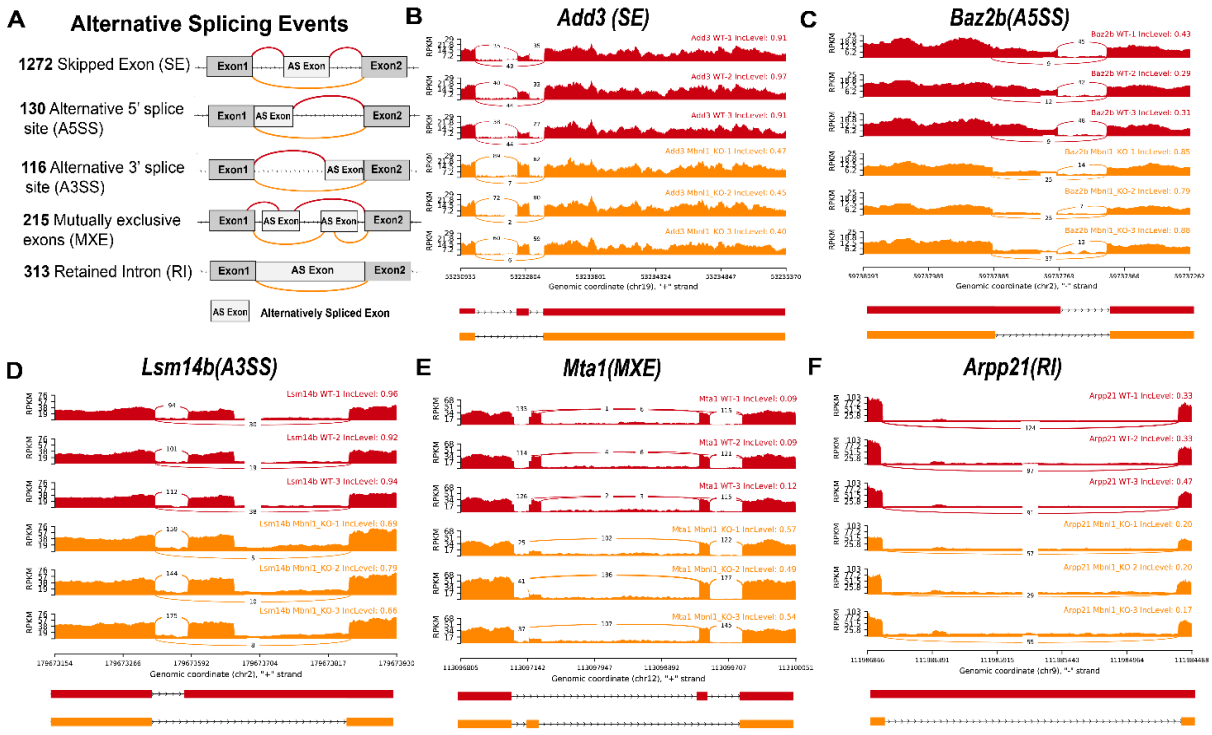


Figure 6.4: Alternative splicing events acquired by rMATS analysis. *A. Types of AS events. This figure is adapted from rMATS documentation11 explaining the splicing event with constitutive and alternatively spliced exons. Labelled with the number of each event at FDR 10%. B. Sashimi plot of Add3 gene exhibiting skipped exon (SE). C. Sashimi plot of Baz2b gene exhibiting alternative 5' splice site (A5SS). D. Sashimi plot of Lsm14b gene exhibiting alternative 3' splice site (A3SS). E. Sashimi plot of Mta1 gene exhibiting mutually exclusive exons (MXE). F. Sashimi plot of Arpp21 gene exhibiting retained intron (RI). Red rows represent three replicates of wild-type and orange rows represent Mbnl1 knock-out replicates. The x-axis corresponds to genomic coordinates and strand information, y-axis shows coverage in RPKM.*

In this study, exon-based and event-based approaches were evaluated to detect AS and APA in bulk RNA-Seq and 3' end sequencing data. The exon-based AS approaches produced a list of differentially expressed exons and a gene-level ranking ordered by the statistical significance of overall gene-level differential splicing activity. For the diffSplice package, differential usage was determined by fitting weighted linear models at an exon-level to estimate the differential log fold-change of an exon against the average log fold-change of the other exons within the same gene (called per exon FC). The gene-level statistical significance was computed by combining individual exon-level significance tests into a gene-wise test by the Simes method. The ranking by gene-level differential splicing activity could subsequently be used to perform a gene set enrichment analysis (GSEA) of key pathways involved (Ritchie et al., 2014). A similar strategy was used by DEXSeq, which fitted a generalized linear model to measure differential exon usage, differing in certain steps such as filtering, normalization, and dispersion estimation.

The preprocessing and mapping of reads to the genome assembly involved several key steps. Initially, the quality of the raw reads was assessed using FASTQC ensuring the acceptability of read quality. Following QC, read alignment to the reference genome was conducted using STAR Aligner. This process included building the index for the reference genome and aligning raw reads to it, resulting in the generation and sorting of BAM files for each sample. Subsequently, exon annotations were prepared by running a supplementary code file with the downloaded annotation in GTF format. The GTF file contained multiple exon entries for different isoforms, and this file was used to "collapse" the multiple transcript IDs for each exon, which was an important step to define exon counting bins. Finally, the number of reads mapped to different transcripts/exons was counted using RSubread, resulting in the exon-level read count matrix. To prepare the RNA-seq library for an AS analysis, it was important to use a strand-specific bulk RNA-seq protocol (Conesa et al., 2016), as a large fraction of genes in vertebrate genomes overlapped on different strands, and a non-strand-specific protocol was unable to distinguish these overlapping regions, confounding final exon detection. Another consideration was read depth, with splicing analyses requiring deeper sequencing than DGE, e.g., 30–60 million reads per sample, versus 5–25 million reads per sample for DGE

(https://sapac.support.illumina.com/bulletins/2017/04/considerations-for-rna-seq-read-length-and-coverage-.html).

All the tools demonstrated in the protocol supported both single-end and paired-end sequencing data. The importance of using appropriate methods for detecting AS and APA in RNA-seq data was underscored, as these events play important roles in regulating gene expression and function. The evaluation of the different approaches provided useful insights into the strengths and limitations of each method and highlighted the need for careful validation and interpretation of the results.

Chapter 7: CRMnet: a deep learning model for predicting gene expression from large regulatory sequence datasets

7.1 Preface

This chapter features a publication in ***Frontiers in Big Data***, of which I was a co-author along with **Ke Ding**. My contribution to this research was identifying the DREAM challenge and collecting the associated data, pre-processing of the data and performing motif analysis. I also drafted the method section for Motif analysis in the manuscript. I would like to take this opportunity to thank my co-authors and supervisor Jean for this opportunity.

ABSTRACT:

Recent large datasets measuring the gene expression of millions of possible gene promoter sequences provide a resource to design and train optimised deep neural network architectures to predict expression from sequences. High predictive performance due to the modelling of dependencies within and between regulatory sequences is an enabler for biological discoveries in gene regulation through model interpretation techniques.

To understand the regulatory code that delineates gene expression, we have designed a novel deep-learning model (CRMnet) to predict gene expression in Saccharomyces cerevisiae. Our model outperforms the current benchmark models and achieves a Pearson correlation coefficient of 0.971. Interpretation of informative genomic regions determined from model saliency maps, and overlapping the saliency maps with known yeast motifs, support that our model can successfully locate the binding sites of transcription factors that actively modulate gene expression. We compare our model's training times on a large compute cluster with GPUs and Google TPUs to indicate practical training times on similar datasets.

7.2 Introduction

Cis-regulatory sequences, also known as cis-regulatory modules consist of promoters, enhancers, silencers, and insulators (Davidson & Erwin, 2006). These sequences are recognized and bound by transcription factors (TFs), which are regulatory proteins that control gene expression (Ni & Su, 2021). Changes to the cis-regulatory sequences can alter their interaction with TFs, leading to changes in cell phenotype and cell-state transitions (de Boer et al., 2020). There is growing evidence of the importance of cis-regulatory element modification in various diseases such as cancer and diabetes (Mathelier et al., 2015). As a result, understanding how cis-regulatory elements regulate gene expression is crucial for comprehending transcriptional gene regulation. However, predicting the expression of DNA sequences directly has been challenging due to the lack of high-quality data. Recently, the expression levels of over 100 million random promoter sequences were measured in a high-throughput manner using the Gigantic Parallel Reporter Assay (GPRA) to regulate yeast gene constructs, providing large training sets for developing models to decode cis-regulatory logic. Thus, sequence-to-expression models have been proposed to predict how changes in cis-regulatory sequences impact gene expression (Vaishnav et al., 2022).

As deep neural network (DNN) architectures become increasingly complex, parallel acceleration hardware such as GPUs and TPUs, along with high performance clusters/cloud computing, are being developed to reduce the overall training time (Wang et al., 2020). As a result, the issue of model training time on these different high performance computing architectures has become important, especially as neural networks become more sophisticated and the volume of scientific data grows.

In this research, a new DNN model called CRMnet was introduced, which was a Transformer encoded U-Net. It predicted the expression levels of yeast promoter DNA sequences and outperformed previous benchmark models proposed in Vaishnav et al. (2022), achieving a Pearson correlation coefficient of 0.971 in the test dataset. Accurately predicting expression from promoter sequences allowed such models to be used predictively to design new regulatory sequences in synthetic biology, study the predicted

effects of mutations, and help in understanding gene regulation by interpreting the model. The model was interpreted by visualizing saliency maps, which identified key regions in the promoter sequences that affected the corresponding expression the most. It was shown that the model could learn biologically meaningful information by quantifying the saliency information over known yeast sequence motifs. Finally, the performance of the model was compared on large datasets using parallel hardware such as GPUs and TPUs on an HPC cluster.

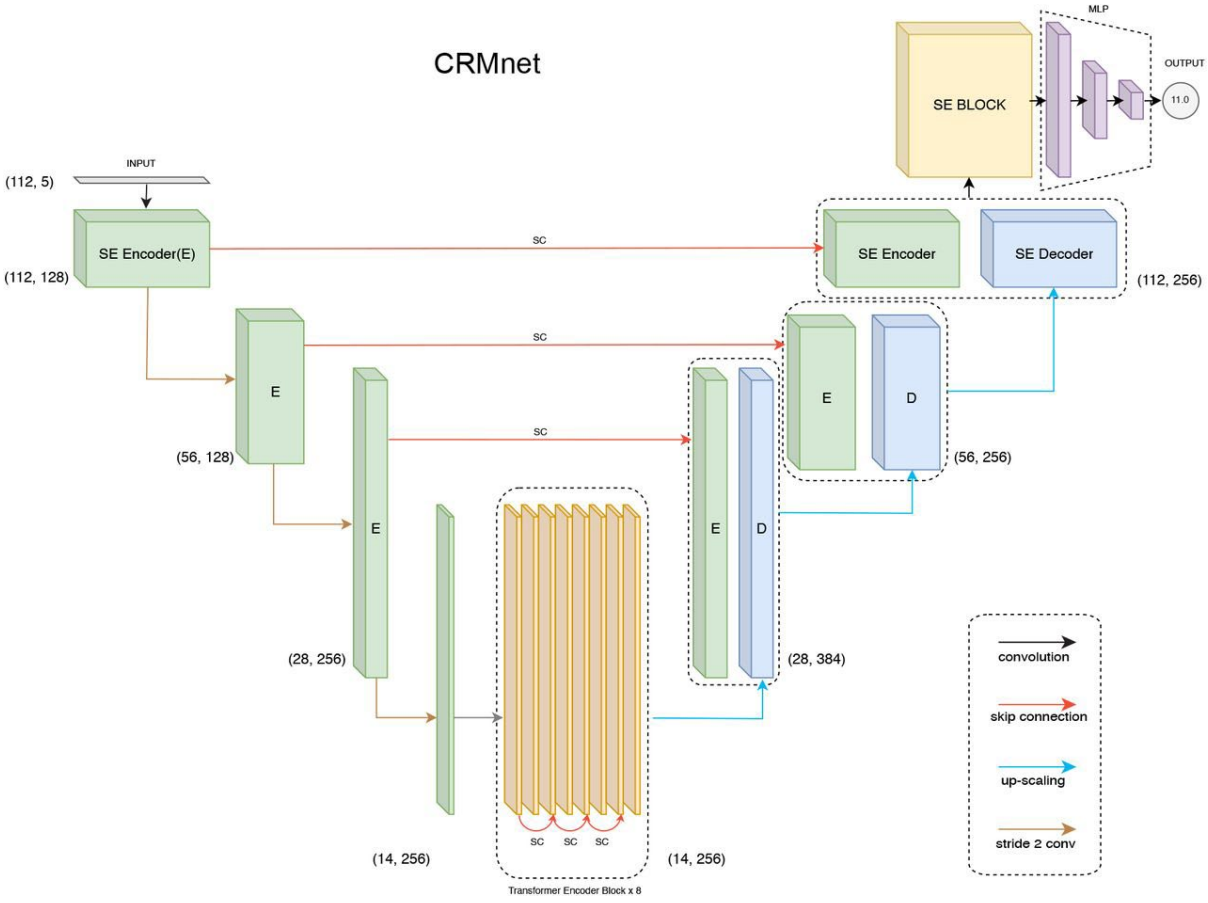


Figure 7.1: CRMnet model's architecture. *CRMnet consists of Squeeze and Excitation (SE) Encoder Blocks, Transformer Encoder Blocks, SE Decoder Blocks, SE Block and Multi-Layer Perceptron (MLP). Similar to the UNet architecture, the encoder and corresponding decoder at the same level have a skip connection (SC) so the decoder utilizes the concatenation of the upsampled feature map with the corresponding higher resolution encoder feature map at that level.*

7.3 Results and Discussion

We began by assessing the performance of the model and comparing it to existing deep learning models. Through ablation studies, we investigated the roles of the different subparts of model. To highlight the biological significance of the model, we utilized saliency maps for model interpretation and compared them with enriched transcription factor binding site motifs discovered through probabilistic motif discovery. Finally, we compared the training time between TPUs and GPUs.

In the first phase of our evaluation, we presented the predictive performance of our fine-tuned deep learning model on independent experimental test sets of both random and native promoter sequences found in yeast, in both complex and defined mediums. To measure the performance of our deep learning models, we calculated the Pearson Correlation Coefficient (r) and Coefficient of Determination (R2) on the test datasets. Our results demonstrated that our fine-tuned CRMnet model achieved outstanding prediction performance on both native and random sequences, with r values of 0.971 and 0.987, respectively, in complex medium **(Figure 7.2)**. In defined medium, our model achieved r values of 0.955 and 0.973, respectively.

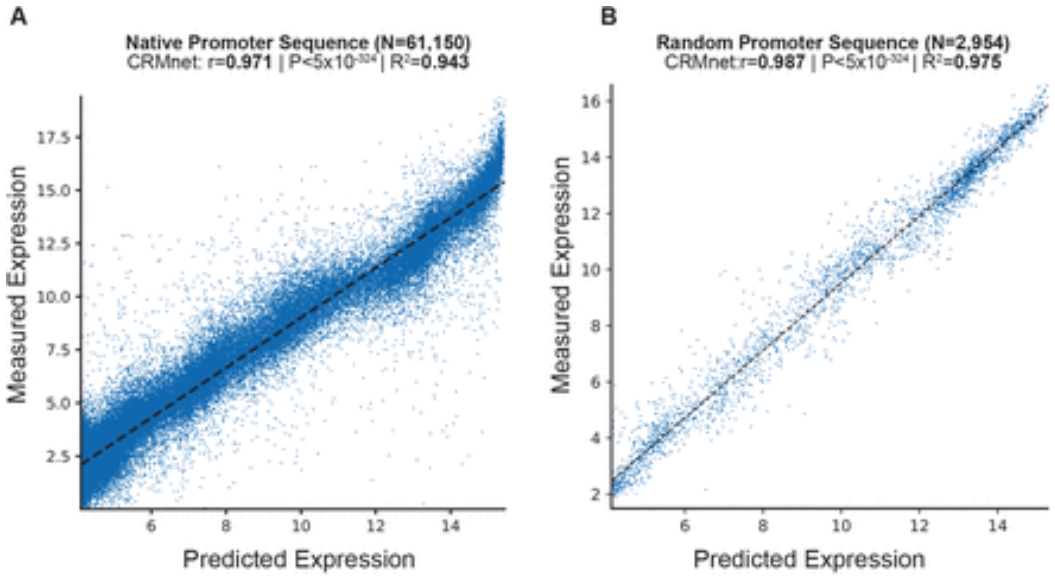


Figure 7.2: Prediction of expression from yeast native sequences from CRMnet*. CRMnet tested on A. native promoter sequences; and B. random promoter sequences. The y-axes represent measured expression levels, while the x-axes represent predicted expression levels. As a benchmark, the model*

performance metrics of the Pearson r value, associated two-tailed p-values, and R-square for the transformer model from (Vaishnav et al., 2022) showed: A: r=0.963, P < 5 × 10−324, R2=0.927; B: r=0.978, P < 5 × 10−324, R2=0.95.

Next, we utilized saliency maps to interpret the model's predictions by highlighting predictive motifs. Gradient backpropagation-based saliency maps have been commonly used to identify model-derived features in input data and interpret the relationship between the input and prediction of the trained model. In this method, a higher saliency value of a particular segment of the input data indicates its significance in the model's prediction. By combining gradient values with the input sequences, input-masked gradients can be used to visualize segments that significantly impact the model's prediction. To compare the results, we used probabilistic motif discovery based on expression levels to identify significant TF motifs.

We discovered that the top five motifs associated with higher expression levels were NHP10 (High-mobility group (HMG) domain factors), REB1 (Myb/SANT domain factors), ABF1 (Basic helix-loop-helix factors (bHLH), AZF1 (C2H2 zinc finger factors), and RAP1 (Myb/SANT domain factors).

Consequently, we generated saliency map logos from our fine-tuned model over yeast native sequences and compared them to the significant motifs discovered through probabilistic motif discovery. The saliency map matched known yeast motif logos, indicating that the model learned to predict expression levels based on biologically meaningful features. We further quantified this by calculating the mean saliency map gradients over the positions in the sequences that matched the top five motifs and found that these motifs were associated with significantly higher saliency gradients than the mean over all sequences as a control **(Figure 7.3)**.

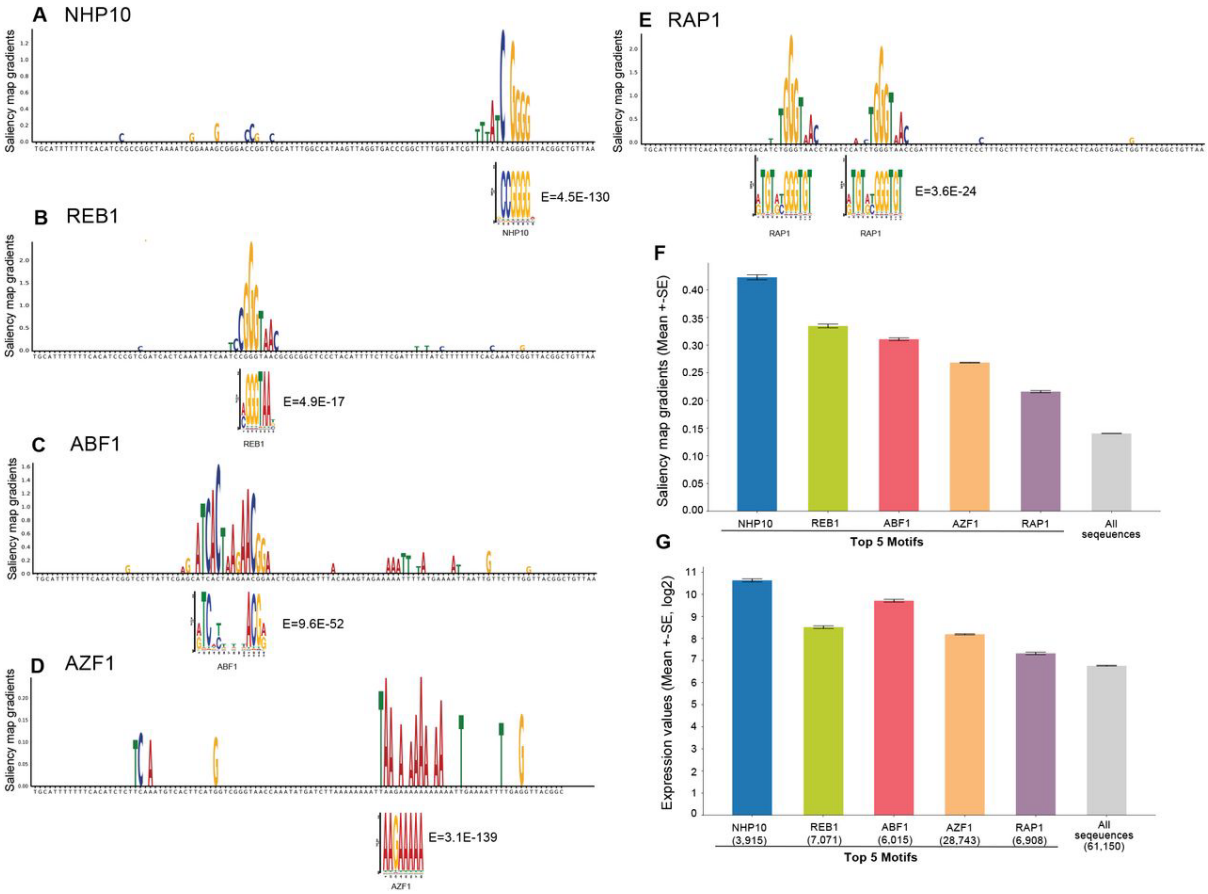


Figure 7.3: Model interpretation by saliency maps. *A-E. The top 5 yeast TF motifs detected by motif discovery: NHP10, REB1, ABF1, AZF1, and RBP1. Shown is an example sequence with its saliency map gradients over 80-nt for each motif, aligned with the known TF motif logo and E-values. F. Mean saliency map gradients over these top 5 motif matches in yeast native sequences, and mean saliency map gradients over all sequences as the controls. G. Mean expression levels of yeast native sequences containing these top 5 TF motifs, and all native sequences as the control.*

Chapter 8: A unique histone 3 lysine 14 chromatin signature underlies tissue-specific gene regulation

8.1 Preface

This chapter features a publication in ***Molecular cell***, of which I was a co-author. My contribution to this research included Differential expression analysis of RNA-seq reads from four replicates, Differential exon usage analysis and gene set enrichment analysis. I would like to thank my co-authors and my supervisor **Jean** for making me a part of this highly influential work.

ABSTRACT:

Organismal development and cell differentiation critically depend on chromatin state transitions. However, certain developmentally regulated genes lack histone 3 lysine 9 and 27 acetylation (H3K9ac and H3K27ac, respectively) and histone 3 lysine 4 (H3K4) methylation, histone modifications common to most active genes. Here we describe a chromatin state featuring unique histone 3 lysine 14 acetylation (H3K14ac) peaks in key tissue-specific genes in *Drosophila* and human cells. Replacing H3K14 in *Drosophila* demonstrates that H3K14 is essential for expression of genes devoid of canonical histone modifications in the embryonic gut and larval wing imaginal disc, causing lethality and defective wing patterning. We find that the SWI/SNF protein Brahma (Brm) recognizes H3K14ac, that brm acts in the same genetic pathway as H3K14R, and that chromatin accessibility at H3K14ac-unique genes is decreased in H3K14R mutants. Our results show that acetylation of a single lysine is essential at genes devoid of canonical histone marks and uncover an important requirement for H3K14 in tissue-specific gene regulation.

8.2 Introduction

Cellular differentiation and response to environmental stimuli involve changes in gene expression and chromatin state, as indicated by various histone post-translational modifications (Allis & Jenuwein, 2016; Atlasi & Stunnenberg, 2017). For instance, H3K27me3 is found in repressed genes while H3K4me3 is present in active gene promoters (Black et al., 2012). Histone acetylation, specifically H3K27ac and H3K9ac, is linked with active transcription (Verdin & Ott, 2015), but some developmental genes lack these marks (Pérez-Lluch et al., 2015). In this study, the role of H3K14ac in developmental gene expression was investigated and it was discovered that a group of tissue-specific genes lacking canonical histone marks were uniquely decorated with this modification.

Studying the functions of individual histone modifications in multicellular organisms is challenging due to the dispersed nature of the encoding genes. However, in *Drosophila melanogaster,* the histone complex (HisC) contains clusters of histone genes that can be manipulated to investigate the direct biological roles of specific histone modifications (Bongartz & Schloissnig, 2019; Lifton et al., 1978; McKay et al., 2015). This approach was used to examine the function of H3K14ac and whether it was functionally redundant or had unique functions compared to other acetylation marks. The findings demonstrate that H3K14 is essential for the expression of genes uniquely marked with H3K14ac, which is necessary for wing patterning and animal survival.

8.3 Results and Discussion

Differential expression analysis

Following adapter sequence trimming using Trimmomatic, we mapped RNA-seq reads from four replicates of mutant Df(2L)HisCFRT40A;6xHisGUH3K14R (H3K14R) embryos, Df(2L)HisCFRT40A; 6xHisGU (H3wt), and heterozygous Df(2L)HisCFRT40A/CyO Dfd-GFP; 6xHisGUH3K14R (ΔHisC/+;H3K14R) samples to the *Drosophila melanogaster* (dm6) genome assembly using HISAT2. After detecting and removing an outlier using PCA analysis, we used the DEseq2 Bioconductor package for RNA-seq differential expression analysis (DE) of H3K14R versus H3wt and H3K14R versus ΔHisC/+;H3K14R. Differentially expressed genes were identified and presented in Table S1, with a false discovery rate (FDR, Benjamini-Hochberg) estimation. A pathway enrichment analysis was performed using HOMER2 for a set of highly differentially expressed genes between H3K14R and both controls with a FDR < 10% and fold change ≥ 2-fold. To test the significance of gene set changes (up- or downregulated) among the differentially expressed (DE) genes from controls and mutant RNA-seq data, we used the gene set enrichment fgesa Bioconductor package.

H3K14 is required for gene expression in the gut

To investigate the impact of H3K14R on gene expression, we conducted RNA sequencing (RNA-seq) experiments on stage 15 embryos approximately 12-13 hours post egg laying. This specific time frame was selected to identify primary target genes of H3K14 that are affected soon after maternal H3K14ac depletion. To minimize variations due to genetic background, we compared H3K14R embryos with both H3WT and heterozygous ΔHisC/+;H3K14R animals, and looked for genes that showed differential expression between H3K14R and both controls. Consequently, although there were more genes that exhibited expression changes when comparing H3K14R to H3WT or H3K14R to ΔHisC/+;H3K14R separately, only 10 genes showed common upregulation, while 50 genes displayed decreased expression in H3K14R embryos (false discovery rate [FDR] < 0.1, fold change ≥ 2-fold; **Figure 8.1, Table 8.1-8.2**). Gene Ontology (GO) analysis of

the commonly upregulated genes showed weak enrichment for odorant binding, whereas downregulated genes were associated with sugar metabolism and signaling (Figures 8.1C and 8.1D). Additionally, a gene set enrichment analysis (GSEA) yielded comparable outcomes (Figures 8.1E and 8.1F). Notably, we observed that 46% of these downregulated genes were expressed in the midgut or hindgut which were the same tissues that exhibited enrichment of unique H3K14ac peaks. Based on these findings, we concluded that H3K14 is necessary for tissue-specific expression of a crucial set of genes in the development of embryos.

This study identified a set of genes that rely on a unique H3K14ac chromatin state for proper development in the fruit fly *Drosophila melanogaster*. While H3K14 is essential in flies, substitutions with H3K14R are viable in yeasts such as *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*. However, H3K14ac is important for regulating transcription elongation and the response to nutrient stress in these yeasts. In *Drosophila*, Gcn5 and HBO1 are known to acetylate H3K14, but H3K14 target genes lack H3K9ac, suggesting that another enzyme is responsible for H3K14ac at these genes. Inhibition of the HBO1 homolog Chameau in the fly's wing disc results in gene expression phenotypes similar to those observed in H3K14R mutant clones.

This study also found that while elimination or reduction in canonical histone marks has relatively minor effects on *Drosophila* development, reduced H3K14ac levels have more significant consequences. Mass spectrometry was used to identify potential mediators of H3K14ac, including 11 confirmed or likely subunits of the SWI/SNF chromatin remodeler complexes BAP and PBAP. Three of these subunits contain bromodomains that recognize acetylated histones, and the study demonstrates that the *Drosophila* Brm bromodomain binds to an H3K14ac peptide and acts in the same genetic pathway as H3K14R. The role of the Brm-H3K14ac interaction is not yet fully understood, but it may help recruit BAP/PBAB to H3K14 targets or modulate chromatin remodelling activity. H3K14 target genes become less accessible in the absence of H3K14ac, suggesting that this chromatin state plays a widespread role in development beyond what has been revealed by the study. The presence of a unique H3K14ac chromatin state in

human mesenchymal stem cells also suggests that it may be important for cell type-specific expression of key genes in some human cell types.

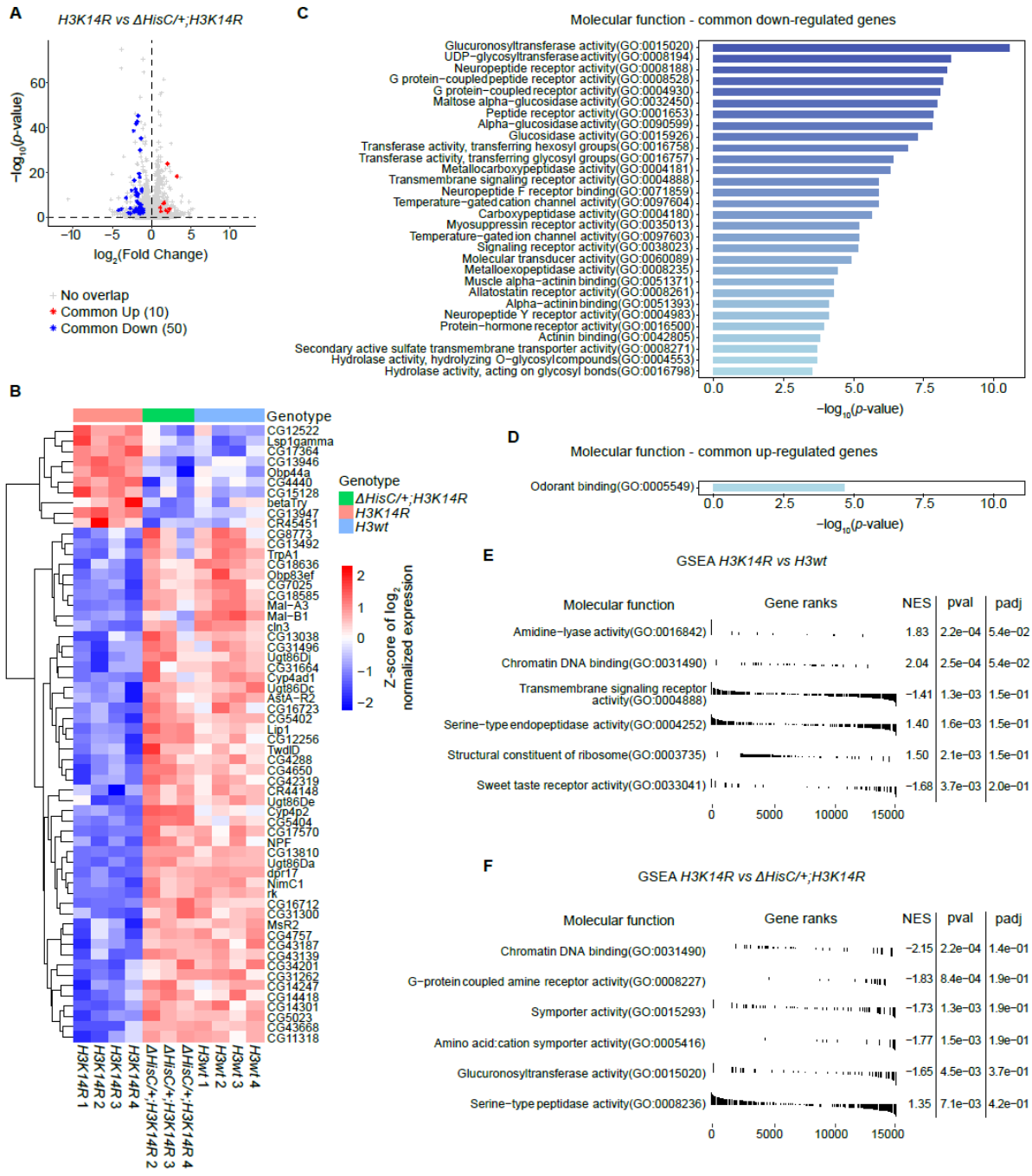


Figure 8.1: Functions of genes in Drosophila embryos that are dependent on H3K14. *The volcano plot in panel A shows the differential gene expression between H3K14R and ΔHisC/+;H3K14R identified through RNA-seq analysis. Genes that are significantly changed (fold change ≥2-fold; FDR <0.1) in comparison to*

both H3wt and to ΔHisC/+;H3K14R are highlighted in blue and red. The number of commonly upregulated and downregulated genes are 10 and 50, respectively. The heatmap in panel B shows the normalized expression in replicates that are standardized to Z-score by subtracting the mean and dividing by the standard deviation of the significantly changed genes. RNA-seq analysis was performed using four biological replicates for H3K14R and H3wt genotypes and three biological replicates for ΔHisC/+;H3K14R. Panels C and D present gene ontology analysis for significantly downregulated and upregulated genes, respectively. Panels E and F show gene set enrichment analysis (GSEA) for H3K14R vs H3wt and H3K14R vs ΔHisC/+;H3K14R.

Table 8.1: List of differentially expressed genes that were downregulated (Overlapped).

gene_name	K14R_vs_WT.log2FC	K14R_vs_WT.FDR	K14R_vs_GFP.log2FC	K14R_vs_GFP.FDR	FB ID
CG31664	-1.85	0.002	-1.904	0.002	FBgn0051664
CG16712	-1.069	0	-1.442	0	FBgn0031561
CR44148	-2.409	0.004	-2.693	0.001	FBgn0264997
CG18585	-2.472	0	-2.192	0	FBgn0031929
CG7025	-2.468	0	-2.019	0.007	FBgn0031930
Lip1	-1.169	0	-1.806	0	FBgn0023496
Mal-B1	-3.04	0	-1.052	0.043	FBgn0032381
NimC1	-1.264	0	-1.267	0	FBgn0259896
rk	-1.598	0	-1.608	0	FBgn0003255
CG4650	-2.448	0	-3.244	0	FBgn0032549
CG18636	-3.458	0.011	-2.688	0.086	FBgn0032551
CG17570	-1.28	0.001	-1.379	0.001	FBgn0260000
Cyp4ad1	-1.347	0	-1.283	0	FBgn0033292
Mal-A3	-1.477	0	-1.245	0	FBgn0002571
Cyp4p2	-1.022	0	-2.27	0	FBgn0033395
CG42319	-1.273	0	-1.717	0	FBgn0259219
CG43187	-1.595	0.039	-1.417	0.098	FBgn0262816
CG34201	-1.556	0	-1.577	0	FBgn0085230
CG43668	-3.66	0.004	-3.812	0.002	FBgn0263743
CG13492	-1.566	0	-1.268	0.004	FBgn0034662
CG13810	-1.447	0	-1.681	0	FBgn0035313
MsR2	-1.009	0.011	-1.04	0.012	FBgn0264002
TrpA1	-1.717	0	-1.254	0.021	FBgn0035934
CG13038	-1.834	0.072	-2.297	0.014	FBgn0040795
cln3	-1.968	0.001	-1.475	0.035	FBgn0036756
Obp83ef	-1.254	0	-1.037	0.002	FBgn0046876
CG31496	-1.866	0.047	-2.176	0.015	FBgn0051496
Ugt86Dc	-1.758	0.003	-1.707	0.007	FBgn0040257
Ugt86Da	-1.656	0	-1.938	0	FBgn0040259
CG4757	-2.997	0.042	-2.941	0.051	FBgn0027584
Ugt86De	-1.881	0.006	-2.057	0.003	FBgn0040255
Ugt86Dj	-1.044	0.016	-1.288	0.002	FBgn0040250

dpr17	-1.454	0	-1.345	0	FBgn0051361
CG12256	-1.065	0	-1.488	0	FBgn0038002
CG8773	-1.787	0.025	-1.673	0.053	FBgn0038135
CG5404	-1.126	0.005	-1.985	0	FBgn0038354
NPF	-3.602	0.028	-4.166	0.006	FBgn0027109
CG31262	-1.775	0.023	-1.482	0.097	FBgn0051262
CG14301	-1.226	0.012	-1.218	0.017	FBgn0038632
CG5023	-1.528	0	-1.755	0	FBgn0038774
CG4288	-1.667	0	-1.951	0	FBgn0038799
CG16723	-1.838	0.086	-2.144	0.033	FBgn0039092
CG31300	-1.806	0	-2.031	0	FBgn0051300
TwdlD	-1.088	0.011	-1.717	0	FBgn0039444
CG14247	-1.078	0.011	-1.419	0	FBgn0039454
CG43139	-1.618	0	-1.885	0	FBgn0262613
CG5402	-2.137	0	-2.533	0	FBgn0039521
AstA-R2	-1.258	0	-1.186	0.002	FBgn0039595
CG11318	-1.552	0.006	-1.756	0.001	FBgn0039818
CG14418	-1.097	0.002	-1.429	0	FBgn0040354

Table8.2. List of differentially expressed genes that were upregulated (Overlapped).

gene_name	K14R_vs_WT.log2FC	K14R_vs_WT.FDR	K14R_vs_GFP.log2FC	K14R_vs_GFP.FDR	FB ID
CG13946	2.026	0	3.226	0	FBgn0040725
CG13947	1.327	0	2.025	0	FBgn0031277
CR45451	1.524	0.098	2.067	0.028	FBgn0267007
CG4440	1.035	0.002	1.606	0	FBgn0040984
Obp44a	1.032	0.003	1.529	0	FBgn0033268
betaTry	1.328	0.075	2.218	0.002	FBgn0010357
CG15128	2.126	0.003	2.313	0.002	FBgn0034467
Lsp1gamma	1.431	0	1.091	0.001	FBgn0002564
CG12522	1.619	0.01	1.834	0.007	FBgn0036131
CG17364	1.538	0.001	1.202	0.016	FBgn0036391

Chapter 9: Summary of Projects

Throughout my research journey, I embarked on a series of projects aimed at unravelling the intricate mechanisms underlying gene regulation and cellular dynamics. The first two projects, detailed extensively in this thesis, stand as the pillars of my exploration into the world of genomics and its implications for understanding biological complexity.

1: scTFNet: Predicting cell-type specific combinatorial binding of neuronal transcription factor network by deep learning

In this project, I dived into the fascinating realm of gene regulation by focusing on transcription factor (TF) binding patterns and their combinatorial interactions. By harnessing the power of deep learning, I developed scTFNet, an innovative model that effectively predicted cell-specific TF binding profiles and the resulting gene expression patterns. This work shed light on the complex orchestration of gene regulation, particularly within the context of neuronal development. The accuracy achieved by scTFNet not only underscored its potential as a predictive tool but also provided insights into the nuanced dynamics of gene expression across various cell types.

2: Using Single-cell RNA seq to explore bone marrow immune landscape by IL-2 therapy

In the second project, I ventured into the immune landscape of the bone marrow, investigating how interleukin-2 (IL-2) therapy influences immune cell dynamics. Employing cutting-edge single-cell RNA sequencing, I unravelled the intricate changes induced by IL-2 treatment within the bone marrow microenvironment. This study illuminated the behaviour of diverse immune cell subsets and their responses to therapy, painting a detailed picture of immune modulation. The project's significance lies in its contribution to our understanding of the immune system's interplay in both health and disease, as well as the potential for targeted therapeutic interventions.

While the other projects discussed in this thesis provided valuable insights into specific aspects of gene regulation and cellular processes, the focus and impact of Projects 1 and 2 are particularly noteworthy.

3: Identification of Alternative Splicing and Polyadenylation in RNA-seq Data

Delving into the intricacies of post-transcriptional gene regulation, this project focused on identifying alternative splicing and polyadenylation events using RNA-seq data. By meticulously analyzing exon-level expression patterns, I uncovered the dynamic nature of gene expression and the role these modifications play in shaping functional diversity. Through this project, the complex interplay between transcription and post-transcriptional modifications emerged, underscoring the layers of regulation that drive cellular complexity.

4: CRMnet: a deep learning model for predicting gene expression from large regulatory sequence datasets

Building upon the power of deep learning, this project introduced CRMnet—a remarkable model designed to predict gene expression from regulatory sequence data. By harnessing the potential of neural networks, this project exemplified the importance of computational approaches in deciphering the intricate regulatory codes that underlie gene expression. The potential of CRMnet to unlock insights into gene regulation is evident, promising to shape our understanding of genetic control in new and profound ways.

5: A unique histone 3 lysine 14 chromatin signature underlies tissue-specific gene regulation

The final project, characterized by the discovery of a distinct histone modification—H3K14ac—offers a glimpse into the complex world of chromatin dynamics and its role in tissue-specific gene regulation. By uncovering the functional significance of this unconventional mark, this project provided a deeper understanding of gene expression control mechanisms. This insight into the epigenetic landscape further enriches our understanding of developmental processes and the orchestration of gene expression.

Chapter 10: References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., & Isard, M. (2016). Tensorflow: a system for large-scale machine learning. Osdi,

Adelman, K., & Lis, J. T. (2012). Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nature Reviews Genetics*, *13*(10), 720-731.

Alipanahi, B., Delong, A., Weirauch, M. T., & Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol*, *33*(8), 831-838. https://doi.org/10.1038/nbt.3300

Alipanahi, B., Delong, A., Weirauch, M. T., & Frey, B. J. (2015). Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. *Nature biotechnology*, *33*(8), 831-838.

Allis, C. D., & Jenuwein, T. (2016). The molecular hallmarks of epigenetic control. *Nat Rev Genet*, *17*(8), 487-500. https://doi.org/10.1038/nrg.2016.59

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol*, *11*(10), R106. https://doi.org/10.1186/gb-2010-11-10-r106

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. In: Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom.

Angermueller, C., Pärnamaa, T., Parts, L., & Stegle, O. (2016). Deep learning for computational biology. Molecular systems biology, 12(7), 878.

Aran, D., Looney, A. P., Liu, L., Wu, E., Fong, V., Hsu, A., Chak, S., Naikawadi, R. P., Wolters, P. J., & Abate, A. R. (2019). Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nature immunology*, *20*(2), 163-172.

Artavanis-Tsakonas, S., Rand, M. D., & Lake, R. J. (1999). Notch signaling: cell fate control and signal integration in development. *Science*, *284*(5415), 770-776.

Atlasi, Y., & Stunnenberg, H. G. (2017). The interplay of epigenetic marks during stem cell differentiation and development. *Nature Reviews Genetics*, *18*(11), 643-658.

Avsec, Ž., Weilert, M., Shrikumar, A., Krueger, S., Alexandari, A., Dalal, K., Fropf, R., McAnany, C., Gagneur, J., & Kundaje, A. (2021). Base-resolution models of

transcription-factor binding reveal soft motif syntax. *Nature genetics*, *53*(3), 354-366.

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W., & Noble, W. S. (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic acids research*, *37*(suppl_2), W202-W208.

Barolo, S., Walker, R. G., Polyanovsky, A. D., Freschi, G., Keil, T., & Posakony, J. W. (2000). A notch-independent activity of suppressor of hairless is required for normal mechanoreceptor physiology. *Cell*, *103*(6), 957-970.

Batra, R., Charizanis, K., Manchanda, M., Mohan, A., Li, M., Finn, D. J., Goodwin, M., Zhang, C., Sobczak, K., & Thornton, C. A. (2014). Loss of MBNL leads to disruption of developmentally regulated alternative polyadenylation in RNA-mediated disease. *Molecular cell*, *56*(2), 311-322.

Bier, E. (2005). Drosophila, the golden bug, emerges as a tool for human genetics. *Nature Reviews Genetics*, *6*(1), 9-23. https://www.nature.com/articles/nrg1503

Black, J. C., Van Rechem, C., & Whetstine, J. R. (2012). Histone lysine methylation dynamics: establishment, regulation, and biological impact. *Molecular cell*, *48*(4), 491-507.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114-2120. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4103590/pdf/btu170.pdf

Bongartz, P., & Schloissnig, S. (2019). Deep repeat resolution—the assembly of the Drosophila Histone Complex. *Nucleic acids research*, *47*(3), e18-e18.

Boyer, L. A., Lee, T. I., Cole, M. F., Johnstone, S. E., Levine, S. S., Zucker, J. P., Guenther, M. G., Kumar, R. M., Murray, H. L., Jenner, R. G., Gifford, D. K., Melton, D. A., Jaenisch, R., & Young, R. A. (2005). Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell*, *122*(6), 947-956. https://doi.org/10.1016/j.cell.2005.08.020

Boyman, O., & Sprent, J. (2012). The role of interleukin-2 during homeostasis and activation of the immune system. *Nature Reviews Immunology*, *12*(3), 180-190.

Boyman, O., Surh, C. D., & Sprent, J. (2006). Potential use of IL-2/anti-IL-2 antibody immune complexes for the treatment of cancer and autoimmune disease. *Expert opinion on biological therapy*, *6*(12), 1323-1331.

Brand, M., Jarman, A. P., Jan, L. Y., & Jan, Y. N. (1993). asense is a Drosophila neural precursor gene and is capable of initiating sense organ formation. *Development*, *119*(1), 1-17.

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., & Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature methods*, *10*(12), 1213-1218. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3959825/pdf/nihms554473.pdf

Buenrostro, J. D., Wu, B., Chang, H. Y., & Greenleaf, W. J. (2015). ATAC-seq: a method for assaying chromatin accessibility genome-wide. *Current protocols in molecular biology*, *109*(1), 21.29. 21-21.29. 29.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., & Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology*, *36*(5), 411-420.

Calderon, D., Blecher-Gonen, R., Huang, X., Secchia, S., Kentro, J., Daza, R. M., Martin, B., Dulja, A., Schaub, C., & Trapnell, C. (2022). The continuum of Drosophila embryonic development at single-cell resolution. *Science*, *377*(6606), eabn5800.

Campos-Ortega, J. A., Hartenstein, V., Campos-Ortega, J. A., & Hartenstein, V. (1997). *Stages of Drosophila embryogenesis*.

Cheng, T., Rodrigues, N., Shen, H., Yang, Y.-g., Dombkowski, D., Sykes, M., & Scadden, D. T. (2000). Hematopoietic stem cell quiescence maintained by p21cip1/waf1. *Science*, *287*(5459), 1804-1808.

Cleary, M. D., & Doe, C. Q. (2006). Regulation of neuroblast competence: multiple temporal identity factors specify distinct neuronal fates within a single early competence window. *Genes & development*, *20*(4), 429-434.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., & Zhang, X. (2016). A survey of best practices for RNA-seq data analysis. *Genome biology*, *17*(1), 1-19.

Core, L. J., & Lis, J. T. (2008). Transcription regulation through promoter-proximal pausing of RNA polymerase II. *Science*, *319*(5871), 1791-1792.

Core, L. J., Waterfall, J. J., & Lis, J. T. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*, *322*(5909), 1845-1848.

Cowden, J., & Levine, M. (2003). Ventral dominance governs sequential patterns of gene expression across the dorsal–ventral axis of the neuroectoderm in the Drosophila embryo. *Developmental biology*, *262*(2), 335-349.

Crews, S. T. (2019). Drosophila embryonic CNS development: neurogenesis, gliogenesis, cell fate, and differentiation. *Genetics*, *213*(4), 1111-1144.

Cubas, P., De Celis, J., Campuzano, S., & Modolell, J. (1991). Proneural clusters of achaete-scute expression and the generation of sensory organs in the Drosophila imaginal wing disc. *Genes & development*, *5*(6), 996-1008.

Cui, Y., Dong, Q., Hong, D., & Wang, X. (2019). Predicting protein-ligand binding residues with deep convolutional neural networks. BMC Bioinformatics, 20(1), 1-12.

Cusanovich, D. A., Daza, R., Adey, A., Pliner, H. A., Christiansen, L., Gunderson, K. L., Steemers, F. J., Trapnell, C., & Shendure, J. (2015). Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*, *348*(6237), 910-914. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4836442/pdf/nihms776051.pdf

Cusanovich, D. A., Pavlovic, B., Pritchard, J. K., & Gilad, Y. (2014). The functional consequences of variation in transcription factor binding. *PLoS genetics*, *10*(3), e1004226.

Cusanovich, D. A., Reddington, J. P., Garfield, D. A., Daza, R. M., Aghamirzaie, D., Marco-Ferreres, R., Pliner, H. A., Christiansen, L., Qiu, X., & Steemers, F. J. (2018). The cis-regulatory dynamics of embryonic development at single-cell resolution. *Nature*, *555*(7697), 538-542.

Davidson, E. H., & Erwin, D. H. (2006). Gene regulatory networks and the evolution of animal body plans. *Science*, *311*(5762), 796-800.

de Boer, C. G., Vaishnav, E. D., Sadeh, R., Abeyta, E. L., Friedman, N., & Regev, A. (2020). Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. *Nature biotechnology*, *38*(1), 56-65.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15-21.

Doe, C. Q., Chu-LaGraff, Q., Wright, D. M., & Scott, M. P. (1991). The prospero gene specifies cell fates in the Drosophila central nervous system. *Cell*, *65*(3), 451-464. https://www.cell.com/cell/pdf/0092-8674(91)90463-9.pdf?_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2F0092867491904639%3Fshowall%3Dtrue

Elkon, R., Ugalde, A. P., & Agami, R. (2013). Alternative cleavage and polyadenylation: extent, regulation and function. *Nature Reviews Genetics*, *14*(7), 496-506.

Eraslan, G., Avsec, Ž., Gagneur, J., & Theis, F. J. (2019). Deep learning: new computational modelling techniques for genomics. *Nature Reviews Genetics*, *20*(7), 389-403.

Farnham, P. J. (2009). Insights from genomic profiling of transcription factors. *Nature Reviews Genetics*, *10*(9), 605-616.

Gaspar, J. M. (2018). Improved peak-calling with MACS2. *BioRxiv*, 496521.

Gaulton, K. J., Nammo, T., Pasquali, L., Simon, J. M., Giresi, P. G., Fogarty, M. P., Panhuis, T. M., Mieczkowski, P., Secchi, A., & Bosco, D. (2010). A map of open chromatin in human pancreatic islets. *Nature genetics*, *42*(3), 255-259.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., & Gentry, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, *5*(10), 1-16.

Golowasch, J., Thomas, G., Taylor, A. L., Patel, A., Pineda, A., Khalil, C., & Nadim, F. (2009). Membrane capacitance measurements revisited: dependence of capacitance value on measurement method in nonisopotential neurons. *Journal of neurophysiology*, *102*(4), 2161-2175.

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, *17*(6), 333-351. https://www.nature.com/articles/nrg.2016.49

Grant, C. E., Bailey, T. L., & Noble, W. S. (2011). FIMO: scanning for occurrences of a given motif. *Bioinformatics*, *27*(7), 1017-1018.

Greig, L. C., Woodworth, M. B., Galazo, M. J., Padmanabhan, H., & Macklis, J. D. (2013). Molecular logic of neocortical projection neuron specification, development and diversity. *Nature Reviews Neuroscience*, *14*(11), 755-769.

Grosskortenhaus, R., Pearson, B. J., Marusich, A., & Doe, C. Q. (2005). Regulation of temporal identity transitions in Drosophila neuroblasts. *Developmental cell*, *8*(2), 193-202. https://www.cell.com/developmental-cell/pdf/S1534-5807(04)00456-3.pdf

Guo, Y., Mahony, S., & Gifford, D. K. (2012). High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints.

Hartemann, A., Bensimon, G., Payan, C. A., Jacqueminet, S., Bourron, O., Nicolas, N., Fonfrede, M., Rosenzwajg, M., Bernard, C., & Klatzmann, D. (2013). Low-dose interleukin 2 in patients with type 1 diabetes: a phase 1/2 randomised, double-blind, placebo-controlled trial. *The lancet Diabetes & endocrinology*, *1*(4), 295-305.

Hartenstein, V., & Campos-Ortega, J. A. (1984). Early neurogenesis in wild-type Drosophila melanogaster. *Wilhelm Roux's archives of developmental biology*, *193*, 308-325.

Hattori, D., Millard, S. S., Wojtowicz, W. M., & Zipursky, S. L. (2008). Dscam-mediated cell recognition regulates neural circuit formation. *Annual review of cell and developmental biology*, *24*, 597-620.

He, Q., Johnston, J., & Zeitlinger, J. (2015). ChIP-nexus enables improved detection of in vivo transcription factor binding footprints. *Nature biotechnology*, *33*(4), 395-401.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., & Glass, C. K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B

cell identities. *Molecular cell*, *38*(4), 576-589. https://www.cell.com/molecular-cell/pdf/S1097-2765(10)00366-7.pdf

Heng, T. S., Painter, M. W., Immunological Genome Project Consortium, t., Elpek, K., Lukacs-Kornek, V., Mauermann, N., Turley, S. J., Koller, D., Kim, F. S., & Wagers, A. J. (2008). The Immunological Genome Project: networks of gene expression in immune cells. *Nature immunology*, *9*(10), 1091-1094.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in science & engineering*, *9*(03), 90-95.

Ihaka, R., & Gentleman, R. (1996). R: a language for data analysis and graphics. *Journal of computational and graphical statistics*, *5*(3), 299-314.

Isshiki, T., Pearson, B., Holbrook, S., & Doe, C. Q. (2001). Drosophila neuroblasts sequentially express transcription factors which specify the temporal identity of their neuronal progeny. *Cell*, *106*(4), 511-521. https://www.cell.com/cell/pdf/S0092-8674(01)00465-2.pdf

Jiang, C., & Pugh, B. F. (2009). Nucleosome positioning and gene regulation: advances through genomics. *Nat Rev Genet*, *10*(3), 161-172. https://doi.org/10.1038/nrg2522

Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K. R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., & Wei, G. (2013). DNA-binding specificities of human transcription factors. *Cell*, *152*(1-2), 327-339.

Kalia, V., Sarkar, S., Subramaniam, S., Haining, W. N., Smith, K. A., & Ahmed, R. (2010). Prolonged interleukin-2Rα expression on virus-specific CD8+ T cells favors terminal-effector differentiation in vivo. *Immunity*, *32*(1), 91-103.

Kanai, M. I., Okabe, M., & Hiromi, Y. (2005). seven-up Controls switching of transcription factors that specify temporal identities of Drosophila neuroblasts. *Developmental cell*, *8*(2), 203-213. https://www.cell.com/developmental-cell/pdf/S1534-5807(04)00466-6.pdf

Karolchik, D., Baertsch, R., Diekhans, M., Furey, T. S., Hinrichs, A., Lu, Y., Roskin, K. M., Schwartz, M., Sugnet, C. W., & Thomas, D. J. (2003). The UCSC genome browser database. *Nucleic acids research*, *31*(1), 51-54.

Keller, R. (2002). Shaping the vertebrate body plan by polarized embryonic cell movements. *Science*, *298*(5600), 1950-1954.

Kelley, D. R., Reshef, Y. A., Bileschi, M., Belanger, D., McLean, C. Y., & Snoek, J. (2018). Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome research*, *28*(5), 739-750.

Kelley, D. R., Snoek, J., & Rinn, J. L. (2016). Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Research*, *26*(7), 990-999.

Khan, A., & Mathelier, A. (2017). Intervene: a tool for intersection and visualization of multiple gene or genomic region sets. *BMC bioinformatics*, *18*(1), 1-8.

Klapper, J. A., Downey, S. G., Smith, F. O., Yang, J. C., Hughes, M. S., Kammula, U. S., Sherry, R. M., Royal, R. E., Steinberg, S. M., & Rosenberg, S. (2008). High-dose interleukin-2 for the treatment of metastatic renal cell carcinoma: a retrospective analysis of response and survival in patients treated in the surgery branch at the National Cancer Institute between 1986 and 2006. *Cancer*, *113*(2), 293-301.

Klinedinst, S. L., & Bodmer, R. (2003). Gata factor Pannier is required to establish competence for heart progenitor formation.

Knoblich, J. A. (2008). Mechanisms of asymmetric stem cell division. *Cell*, *132*(4), 583-597. https://www.cell.com/cell/pdf/S0092-8674(08)00208-0.pdf

Kwak, H., Fuda, N. J., Core, L. J., & Lis, J. T. (2013). Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science*, *339*(6122), 950-953.

La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M. E., Lönnerberg, P., & Furlan, A. (2018). RNA velocity of single cells. *Nature*, *560*(7719), 494-498.

Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T. R., & Weirauch, M. T. (2018). The human transcription factors. *Cell*, *172*(4), 650-665.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods*, *9*(4), 357-359.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

Li, H., & Guan, Y. (2021). Fast decoding cell type–specific transcription factor binding landscape at single-nucleotide resolution. *Genome research*, *31*(4), 721-731.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & Subgroup, G. P. D. P. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078-2079.

Li, H., Quang, D., & Guan, Y. (2019). Anchor: trans-cell type prediction of transcription factor binding sites. *Genome research*, *29*(2), 281-292.

Li, Z., Gao, E., Zhou, J., Han, W., Xu, X., & Gao, X. (2023). Applications of deep learning in understanding gene regulation. Cell Reports Methods.

Lianoglou, S., Garg, V., Yang, J. L., Leslie, C. S., & Mayr, C. (2013). Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes & development*, *27*(21), 2380-2396.

Liao, W., Lin, J.-X., & Leonard, W. J. (2013). Interleukin-2 at the crossroads of effector responses, tolerance, and immunotherapy. *Immunity*, *38*(1), 13-25.

Lifton, R., Goldberg, M., Karp, R., & Hogness, D. (1978). The organization of the histone genes in Drosophila melanogaster: functional and evolutionary implications. Cold Spring Harbor symposia on quantitative biology,

Lin, D. M., & Goodman, C. S. (1994). Ectopic and increased expression of Fasciclin II alters motoneuron growth cone guidance. *Neuron*, *13*(3), 507-523.

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, *15*(12), 1-21.

Ma, S., & Zhang, Y. (2020). Profiling chromatin regulatory landscape: insights into the development of ChIP-seq and ATAC-seq. *Molecular biomedicine*, *1*, 1-13.

Mackay, F., & Browning, J. L. (2002). BAFF: a fundamental survival factor for B cells. *Nature Reviews Immunology*, 2(7), 465-475.

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., & Martersteck, E. M. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, *161*(5), 1202-1214. https://www.cell.com/cell/pdf/S0092-8674(15)00549-8.pdf

Mahat, D. B., Kwak, H., Booth, G. T., Jonkers, I. H., Danko, C. G., Patel, R. K., Waters, C. T., Munson, K., Core, L. J., & Lis, J. T. (2016). Base-pair-resolution genome-

wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). *Nature protocols*, *11*(8), 1455-1476.

Mantovani, A., Allavena, P., Sica, A. and Balkwill, F., (2008). Cancer-related inflammation. *Nature*, 454(7203), 436-444.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, *17*(1), 10-12.

Massaia, A., Chaves, P., Samari, S., Miragaia, R. J., Meyer, K., Teichmann, S. A., & Noseda, M. (2018). Single cell gene expression to understand the dynamic architecture of the heart. *Frontiers in cardiovascular medicine*, *5*, 167.

Mathelier, A., Shi, W., & Wasserman, W. W. (2015). Identification of altered cis-regulatory elements in human disease. *Trends in Genetics*, *31*(2), 67-76.

Maurange, C., Cheng, L., & Gould, A. P. (2008). Temporal transcription factors and their targets schedule the end of neural proliferation in Drosophila. *Cell*, *133*(5), 891-902. https://www.cell.com/cell/pdf/S0092-8674(08)00499-6.pdf

Mayr, C., & Bartel, D. P. (2009). Widespread shortening of 3′ UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell*, *138*(4), 673-684.

Mbalaviele, G., Novack, D. V., Schett, G., & Teitelbaum, S. L. (2017). Inflammatory osteolysis: a conspiracy against bone. *The Journal of clinical investigation*, *127*(6), 2030-2039.

McKay, D. J., Klusza, S., Penke, T. J., Meers, M. P., Curry, K. P., McDaniel, S. L., Malek, P. Y., Cooper, S. W., Tatomer, D. C., & Lieb, J. D. (2015). Interrogating the function of metazoan histones using engineered gene clusters. *Developmental cell*, *32*(3), 373-386.

Melnick, M. B., Perkins, L. A., Lee, M., Ambrosio, L., & Perrimon, N. (1993). Developmental and molecular characterization of mutations in the Drosophila-raf serine/threonine protein kinase. *Development*, *118*(1), 127-138.

Merkin, J., Russell, C., Chen, P., & Burge, C. B. (2012). Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science*, *338*(6114), 1593-1599.

Metzker, M. L. (2010). Sequencing technologies—the next generation. *Nature Reviews Genetics*, *11*(1), 31-46. https://www.nature.com/articles/nrg2626

Morgan, D. A., Ruscetti, F. W., & Gallo, R. (1976). Selective in vitro growth of T lymphocytes from normal human bone marrows. *Science*, *193*(4257), 1007-1008.

Ni, P., & Su, Z. (2021). Accurate prediction of cis-regulatory modules reveals a prevalent regulatory genome of humans. *NAR Genomics and Bioinformatics*, *3*(2), lqab052.

Nolo, R., Abbott, L. A., & Bellen, H. J. (2000). Senseless, a Zn Finger Transcription Factor, Is Necessary and Sufficient for Sensory Organ Development in Drosophila. *Cell*, *102*(3), 349-362. https://doi.org/10.1016/s0092-8674(00)00040-4

Novotny, T., Eiselt, R., & Urban, J. (2002). Hunchback is required for the specification of the early sublineage of neuroblast 7-3 in the Drosophila central nervous system.

Nüsslein-Volhard, C., & Wieschaus, E. (1980). Mutations affecting segment number and polarity in Drosophila. *Nature*, *287*(5785), 795-801. https://www.nature.com/articles/287795a0

Ou, J., Liu, H., Yu, J., Kelliher, M. A., Castilla, L. H., Lawson, N. D., & Zhu, L. J. (2018). ATACseqQC: a Bioconductor package for post-alignment quality assessment of ATAC-seq data. *BMC genomics*, *19*, 1-13.

Paul, W. E., & Zhu, J. (2010). How are TH2-type immune responses initiated and amplified? *Nature Reviews Immunology*, *10*(4), 225-235.

Pearson, B. J., & Doe, C. Q. (2003). Regulation of neuroblast competence in Drosophila. *Nature*, *425*(6958), 624-628. https://www.nature.com/articles/nature01910

Pearson, B. J., & Doe, C. Q. (2004). Specification of temporal identity in the developing nervous system. *Annu. Rev. Cell Dev. Biol.*, *20*, 619-647. https://www.annualreviews.org/doi/10.1146/annurev.cellbio.19.111301.115142?url_ver=Z39.88-2003&rfr_id=ori%3Arid%3Acrossref.org&rfr_dat=cr_pub++0pubmed

Pérez-Lluch, S., Blanco, E., Tilgner, H., Curado, J., Ruiz-Romero, M., Corominas, M., & Guigó, R. (2015). Absence of canonical marks of active chromatin in developmentally regulated genes. *Nature genetics*, *47*(10), 1158-1167.

Peter, I. S., & Davidson, E. H. (2015). *Genomic control process: development and evolution*. Academic Press.

Pliner, H. A., Packer, J. S., McFaline-Figueroa, J. L., Cusanovich, D. A., Daza, R. M., Aghamirzaie, D., Srivatsan, S., Qiu, X., Jackson, D., & Minkina, A. (2018). Cicero

predicts cis-regulatory DNA interactions from single-cell chromatin accessibility data. Molecular cell, 71(5), 858-871. e858.

Poulin, J.-F., Tasic, B., Hjerling-Leffler, J., Trimarchi, J. M., & Awatramani, R. (2016). Disentangling neural cell diversity using single-cell transcriptomics. *Nature neuroscience*, *19*(9), 1131-1141.

Powell, L. M., & Jarman, A. P. (2008). Context dependence of proneural bHLH proteins. *Current opinion in genetics & development*, *18*(5), 411-417.

Qian, L., Liu, J., & Bodmer, R. (2007). Heart development in Drosophila. *Advances in Developmental Biology*, *18*, 1-29.

Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H. A., & Trapnell, C. (2017). Reversed graph embedding resolves complex single-cell trajectories. *Nature methods*, *14*(10), 979-982. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5764547/pdf/nihms896903.pdf

Quang, D., & Xie, X. (2016). DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic acids research*, *44*(11), e107-e107.

Quang, D., & Xie, X. (2019). FactorNet: a deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data. *Methods*, *166*, 40-47.

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841-842.

Raggatt, L. J., & Partridge, N. C. (2010). Cellular and molecular mechanisms of bone remodeling. *Journal of biological chemistry*, *285*(33), 25103-25108.

Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A., & Manke, T. (2014). deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic acids research*, *42*(W1), W187-W191.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research*, *43*(7), e47-e47.

Roadmap, E. C., Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., & Wang, J. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, *518*(7539), 317-330.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18,

Rosenberg, S. A., Yang, J. C., Sherry, R. M., Kammula, U. S., Hughes, M. S., Phan, G. Q., Citrin, D. E., Restifo, N. P., Robbins, P. F., & Wunderlich, J. R. (2011). Durable complete responses in heavily pretreated patients with metastatic melanoma using T-cell transfer immunotherapycomplete regressions in melanoma. *Clinical cancer research*, *17*(13), 4550-4557.

Rothenberg, E. V. (2014). Transcriptional control of early T and B cell developmental choices. *Annual review of immunology*, *32*, 283-321.

Rubin, G. M., Yandell, M. D., Wortman, J. R., Gabor, G. L., Miklos, Nelson, C. R., Hariharan, I. K., Fortini, M. E., Li, P. W., & Apweiler, R. (2000). Comparative genomics of the eukaryotes. *Science*, *287*(5461), 2204-2215. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2754258/pdf/nihms142596.pdf

Sakaguchi, S., Yamaguchi, T., Nomura, T., & Ono, M. (2008). Regulatory T cells and immune tolerance. *Cell*, *133*(5), 775-787.

Sapoval, N., Aghazadeh, A., Nute, M. G., Antunes, D. A., Balaji, A., Baraniuk, R., Barberan, C., Dannenfelser, R., Dun, C., & Edrisi, M. (2022). Current progress and open challenges for applying deep learning across the biosciences. Nature Communications, 13(1), 1728.

Shen, S., Park, J. W., Lu, Z.-x., Lin, L., Henry, M. D., Wu, Y. N., Zhou, Q., & Xing, Y. (2014). rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proceedings of the National Academy of Sciences*, *111*(51), E5593-E5601.

Sim, G. C., & Radvanyi, L. (2014). The IL-2 cytokine family in cancer immunotherapy. *Cytokine & growth factor reviews*, *25*(4), 377-390.

Simon, J. M., Giresi, P. G., Davis, I. J., & Lieb, J. D. (2012). Using formaldehyde-assisted isolation of regulatory elements (FAIRE) to isolate active regulatory DNA. *Nature protocols*, *7*(2), 256-267. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3784247/pdf/nihms509724.pdf

Slattery, M., Zhou, T., Yang, L., Machado, A. C. D., Gordân, R., & Rohs, R. (2014). Absence of a simple code: how transcription factors read the genome. *Trends in biochemical sciences*, *39*(9), 381-399.

Song, L., & Crawford, G. E. (2010). DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harbor Protocols*, *2010*(2), pdb. prot5384. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3627383/pdf/nihms456416.pdf

Spitz, F., & Furlong, E. E. (2012). Transcription factors: from enhancer binding to developmental control. *Nature Reviews Genetics*, *13*(9), 613-626. https://www.nature.com/articles/nrg3207

St Johnston, D. (2002). The art and design of genetic screens: Drosophila melanogaster. *Nature Reviews Genetics*, *3*(3), 176-188. https://www.nature.com/articles/nrg751

Stormo, G. D. (2000). DNA binding sites: representation and discovery. *Bioinformatics*, *16*(1), 16-23.

Stormo, G. D., & Hartzell 3rd, G. (1989). Identifying protein-binding sites from unaligned DNA fragments. *Proceedings of the National Academy of Sciences*, *86*(4), 1183-1187.

Suvà, M. L., & Tirosh, I. (2019). Single-cell RNA sequencing in cancer: lessons learned and emerging challenges. *Molecular cell*, *75*(1), 7-12. https://www.cell.com/molecular-cell/pdf/S1097-2765(19)30356-9.pdf

Sznajder, Ł. J., Scotti, M. M., Shin, J., Taylor, K., Ivankovic, F., Nutter, C. A., Aslam, F. N., Subramony, S., Ranum, L. P., & Swanson, M. S. (2020). Loss of MBNL1 induces RNA misprocessing in the thymus and peripheral blood. *Nature communications*, *11*(1), 2022.

Tsukasaki, M., & Takayanagi, H. (2019). Osteoimmunology: Evolving concepts in bone–immune interactions in health and disease. *Nature Reviews Immunology*, *19*(10), 626-642.

Ueberschär, M., Wang, H., Zhang, C., Kondo, S., Aoki, T., Schedl, P., Lai, E. C., Wen, J., & Dai, Q. (2019). BEN-solo factors partition active chromatin to ensure proper gene activation in Drosophila. *Nature communications*, *10*(1), 5700.

Vaishnav, E. D., de Boer, C. G., Molinet, J., Yassour, M., Fan, L., Adiconis, X., Thompson, D. A., Levin, J. Z., Cubillos, F. A., & Regev, A. (2022). The evolution, evolvability and engineering of gene regulatory DNA. *Nature*, *603*(7901), 455-463.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Vaz, J. M., & Balaji, S. (2021). Convolutional neural networks (CNNs): Concepts and applications in pharmacogenomics. Molecular diversity, 25(3), 1569-1584.

Verdin, E., & Ott, M. (2015). 50 years of protein acetylation: from gene regulation to epigenetics, metabolism and beyond. *Nature reviews Molecular cell biology*, *16*(4), 258-264.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., & Bright, J. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature methods*, *17*(3), 261-272.

Von Ohlen, T., & Doe, C. Q. (2000). Convergence of dorsal, dpp, and egfr signaling pathways subdivides the drosophila neuroectoderm into three dorsal-ventral columns. *Developmental biology*, *224*(2), 362-372.

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S. F., Schroth, G. P., & Burge, C. B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature*, *456*(7221), 470-476.

Wang, Y., Wang, Q., Shi, S., He, X., Tang, Z., Zhao, K., & Chu, X. (2020). Benchmarking the performance and energy efficiency of AI accelerators for AI training. 2020 20th IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing (CCGRID),

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, *10*(1), 57-63. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2949280/pdf/nihms229948.pdf

Waskom, M., Botvinnik, O., Ostblom, J., Gelbart, M., Lukauskas, S., Hobson, P., Gemperline, D. C., Augspurger, T., Halchenko, Y., & Cole, J. B. (2020). mwaskom/seaborn: v0. 10.1 (April 2020). *zenodo*.

Welch, J. D., Kozareva, V., Ferreira, A., Vanderburg, C., Martin, C., & Macosko, E. Z. (2019). Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell*, *177*(7), 1873-1887. e1817.

Wysoker, A., Tibbetts, K., & Fennell, T. (2013). Picard tools version 1.90. *Jhpsn*, *107*, 308.

Yu, W., Uzun, Y., Zhu, Q., Chen, C., & Tan, K. (2020). scATAC-pro: a comprehensive workbench for single-cell chromatin accessibility sequencing data. *Genome biology*, *21*(1), 1-17.

Zeng, Y., Gong, M., Lin, M., Gao, D., & Zhang, Y. (2020). A review about transcription factor binding sites prediction based on deep learning. *IEEE Access*, *8*, 219256-219274.

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., & Li, W. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome biology*, *9*(9), 1-9.

Zheng, R., Dong, X., Wan, C., Shi, X., Zhang, X., & Meyer, C. A. (2020). Cistrome Data Browser and Toolkit: analyzing human and mouse genomic data using compendia of ChIP-seq and chromatin accessibility data. *Quantitative Biology*, *8*, 267-276.

Zhou, J., & Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning–based sequence model. *Nature methods*, *12*(10), 931-934.

Zhu, F., Farnung, L., Kaasinen, E., Sahu, B., Yin, Y., Wei, B., Dodonova, S. O., Nitta, K. R., Morgunova, E., & Taipale, M. (2018). The interaction landscape between transcription factors and the nucleosome. *Nature*, *562*(7725), 76-81.