Explainable machine learning with epigenomic features for insights into regulatory and functional genomics

A THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

BY

OMKAR CHANDRA R

Roll No. PHD17206

DEPARTMENT OF COMPUTATIONAL BIOLOGY

INDRAPRASTHA INSTITUTE OF INFORMATION TECHNOLOGY DELHI

NEW DELHI- 110020

31st June, 2024

THESIS CERTIFICATE

This is to certify that the thesis titled **Explainable machine learning with epigenomic features for insights into regulatory and functional genomics**, submitted by **Omkar Chandra R**, to the Indraprastha Institute of Information Technology, Delhi, for the award of the degree of **Doctor of Philosophy**, is a bona fide record of the research work done by him under our supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Dr Vibhor Kumar
Thesis Supervisor
Associate Professor
Centre for Computational Biology
IIIT Delhi, 110020

Place: New Delhi

Date: 17th of December, 2024

ACKNOWLEDGEMENTS

I wish to formally express my profound gratitude to all individuals who have contributed to the projects encompassed in this thesis and to those who have supported me throughout this journey.

I have gained invaluable insights from every person I have interacted with during this period, beginning with my Ph.D. advisor, **Dr. Vibhor Kumar**. I am deeply indebted to him for accepting me into his lab and allowing me to pursue research on a topic of my interest. His guidance has afforded me the opportunity to learn continuously and develop self-reliance. Dr. Vibhor Kumar's mentorship has instilled in me the resilience needed to persist in the right direction, even in the face of challenges, to complete a research project. The successful completion of my Ph.D. research, with its critical thinking and scientific rigor, would not have been possible without his expert input and knowledge. I am sincerely grateful for his support and encouragement, both academically and personally.

My immediate laboratory colleagues have significantly contributed to my understanding of both academic and extramural aspects of life. Since joining the lab, my interactions and collaborations with colleagues have been profoundly instructive and beneficial, facilitating the development of my philosophical and scientific perspectives. Senior colleagues **Indraprakash Jha**, **Smriti Chawala**, **Neetesh Pandey**, **Shreya Mishra**, and **Shristi Gautam** have generously shared their experiences and knowledge with me, providing invaluable guidance in completing the projects related to my thesis. Additionally, my junior colleagues, **Madhu Sharma**, **Durjay Pramanik**, **Biswarup Mahato**, **Niharika Dubey**, **Jaidev Sharma**, and **Karuna Kerketta**, have been integral to the academic research work through their collaboration and have offered steadfast support and encouragement throughout my Ph.D. journey.

I extend my sincere gratitude to the faculties of IIITD's Computational Biology department, who have provided me with essential knowledge and experience throughout my Ph.D. coursework. **Dr. GPS Raghava**, **Dr. Sriram K.** , **Dr. Debarka Sengupta**,

and **Dr. Subhadip Ray Chaudhuri** have been instrumental in my academic growth. I am particularly indebted to **Dr. GPS Raghava** for fostering a dynamic scientific community and ensuring that the resources and facilities necessary for research scholars were available.

I am also deeply thankful to **Dr. Ganesh Bagler**, **Dr. Arjun Ray**, **Dr. Gaurav Ahuja**, **Dr. Jaspreet Kaur Dhanjal** , **Dr. Arul Murugan**, and **Dr. Tarini Shankar Ghosh** for their invaluable scientific feedback during my research presentations at various stages of my Ph.D. journey.

I am grateful to have collaborated with **Dr. Kong Say Li**, whose guidance has allowed me to explore various domains of computational biology research.

Koushik Biswas, a senior colleague and friend, has been a steadfast source of support throughout the challenges and triumphs of my Ph.D. journey. His willingness to listen and assist in overcoming obstacles has been invaluable. Additionally, **Ridam Pal, Gayatri Panda, Harika G.L., Swarnava Samanta, Shreyanshu Sharma, Akshay Srinath, Sumeet Patiyal, Anjali Dhall, Sarita Poonia, Nishant Kumar, Vivek Ruhela, Sakshi Gujral, Avik Datta, Suvilesh Kanave, Siddu Naikodi, Jadav Lokesh**, and **Vishakha Gautam** have been supportive and caring friends and colleagues, sharing in both my joys and challenges.

I am also grateful to colleagues from other labs, including **Shiju S., Raghav Awasthi, Krishan Gupta, Samriddhi Gupta, Aayushi Mittal, Sanjay Mohanty, Naman Kumar Mehta, Bernadette Mathew, Abhishek Halder, Shruti, Omprakash Shete, Sourav Goswami, Shivangi Verma,** , and **Sadiyah Afroz**, for their support at various stages of this journey.

The IIITD library has been an indispensable source of information and knowledge, and I thank **Mr. Rajendra Singh** for his prompt assistance with my research article requests.

The IT and computing facilities at IIITD have been the backbone of my research, and I am thankful to **Mr. Adarsh Kumar** for his constant support and assistance with my computing needs.

My journey would have been considerably more difficult without the valuable help from the IIITD academic office, hostel, and technical support community. I am particu-

larly grateful for the assistance of the Computational Biology department manager, **Ms. Shipra Jain**, and the past Ph.D. admin, **Ms. Priti Patel**, who ensured our academic progress went smoothly. I also thank the current Ph.D. admins, Ms. **Anshu Dureja and Mr. Raju Biswas**, for consistently meeting our academic needs.

I extend my gratitude to the student affairs administration officers, **Dr. Ravi Bhasin, Mr. Jaganan Devedi**, and **Mr. Rajeev Rai**, for ensuring my stay at IIITD was comfortable.

The smooth functioning of the laboratory and hostel has been greatly facilitated by the facility management service staff, to whom I am deeply grateful.

I am thankful to the IIITD director, **Dr. Ranjan Bose**, and the founding director, **Dr. Pankaj Jalote**, for establishing and maintaining the scientific environment at IIITD. Their efforts have also ensured the availability of resources like the gymnasium and sports facilities, contributing to the overall development of the students. I extend my gratitude to the **University Grants Commission, India**, for the fellowship that offered crucial financial support throughout my Ph.D. studies.

This contribution to the scientific community through my thesis would not have been possible without the values and courage instilled in me by my late father and late mother, **Mr. Ramachandra, Mrs. M. V. Mahalakshmi**. My brother, **Mr. Manoj Kumar R.**, and my sister-in-law, **Mrs. Rashmi S. G.**, have provided significant support in my pursuit of the Ph.D. The constant care and encouragement from my uncle, **Mr. Shantaraju K. P.** , have been crucial in my research endeavors; he deserves substantial credit for always believing in me. I also appreciate all my other family members who have supported and helped me throughout this journey.

Finally, I am deeply grateful to all the others, though unmentioned, who have contributed to making this Ph.D. journey a fun, remarkable, and transformative experience.

Omkar Chandra R.
PHD17206
16th of December, 2024

ABSTRACT

KEYWORDS: functional genomics; regulatory genomics; gene function prediction; transcription factors; epigenome

There are thousands of genes with incomplete functional annotations, particularly non-coding genes. Understanding the functional roles of genes is crucial for dissecting the complex genomic regulatory mechanisms underlying biological processes, which in turn provides control over cellular processes such as the immune response and cell cycle for potential clinical interventions. Over the years, numerous computational methods have emerged to link genes with biological processes and molecular functions. However, these methods often fail to account for non-coding genes and rarely provide interpretations of their predictions.

To address this problem, a computational framework has been developed that incorporates features of non-coding genes at the promoter level using epigenome profiles, open-chromatin profiles, and transcription factor (TF) binding profiles of gene promoters. This approach allows for reliable predictions of gene functions, which are independently validated using available CRISPR screens and PubMed abstract mining.

The explainable machine learning algorithms used for the prediction of gene function allowed for post hoc analysis using the top predictors of the learned models, yielding latent clusters of functions that collectively contribute to larger cellular processes. Additionally, downstream analysis using only transcription factors as top predictors provided insights into their synergy and pleiotropy in regulating various biological functions.

The entire computational framework is built into an R package, "GFPredict," which can be used to predict biologically similar genes to user-defined query genes.

Further analysis utilizing TF binding and epigenome profiles as features identified novel disease-gene associations. The predicted associations of coding and non-coding genes with diseases were validated using GWAS data and PubMed abstract mining.

The genomic regulation analysis using top predictors of individual disease gene-sets revealed associations of divergent cell types in diseases. These association insights were validated with evidence from the literature, providing a basis for generating putative hypotheses for developing strategies for diagnosis, prognosis, and potential therapeutics.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS i

ABSTRACT iv

LIST OF TABLES ix

LIST OF FIGURES xiv

ABBREVIATIONS xv

1 INTRODUCTION 1

1.1 The human genome 2

1.2 Incomplete annotation of molecular and biological functions of the genes 4

1.3 Incomplete annotation of disease genes 5

1.4 Characteristics of the genes 5

1.4.1 Regulatory DNA elements (cis-regulatory elements) 6

1.4.1.1 Transcription start sites (TSSs) 7

1.4.1.2 Promoters 8

1.4.1.3 Enhancers 10

1.4.1.4 Silencers 12

1.4.1.5 Insulators 14

1.4.2 Biochemical signatures 15

1.4.2.1 Transcription factors and co-regulators 15

1.4.2.2 DNA methylation 17

1.4.2.3 Histone modifications 18

1.4.2.4 Euchromatin and heterochromatin 20

1.4.3 Gene expression and regulation 21

1.4.4 Non-coding genes 22

1.5 Predictive biology 24

1.6 Summary 26

2 Chapter 2: Epigenome and TF binding patterns are predictive of ontology-based functions of coding and non-coding genes **27**

2.1 Challenges and approaches of associating non-coding genes to functions . . . 27

2.2 Materials and methods . . . 30

2.2.1 Epigenome and TF-binding features score calculation for promoters . . . 30

2.2.2 Prediction method . . . 31

2.2.3 Calculating confidence score for gene-sets . . . 32

2.2.4 Other methods . . . 33

2.2.5 Availability of data and code . . . 37

2.3 Results . . . 37

2.3.1 Epigenome and TF binding patterns at promoters are predictable of ontology-based functions of genes . . . 38

2.3.2 Non-random nature and relevance of high predictability . . . 41

2.3.3 Inference from clustering of functions . . . 42

2.3.4 Independent validations and comparison with other methods 44

2.3.4.1 PubMed abstract mining of co-occurrence of gene names and function term . . . 44

2.3.4.2 Comparison of predicted results with other gene function prediction methods . . . 48

2.3.5 CRISPR-based validation of association of genes with major cellular processes of clusters of functions . . . 49

2.3.6 Explainability through insight into the association of binding patterns of TF-pairs with functions . . . 50

2.3.7 Broader applicability of GFPredict and its utility for predicting functions of non-coding RNAs . . . 55

2.3.7.1 Application of expanding small CRISPR screens for non-coding genes function prediction . . . 56

2.4 Discussion . . . 58

3 Chapter 3. Explainable models using transcription factor binding and epigenome patterns at promoters reveal disease-associated genes and their regulators in the context of cell-types **62**

3.1 Role of epigenome and TFs in diseases and approaches to identify such links . . . 62

3.2 Material and methods . . . 65

3.2.1 Epigenome and TF-binding features score calculation for promoters . . . 65
3.2.2 Prediction method . . . 66
3.2.3 Method for survival plots . . . 66
3.2.4 Method for GWAS validation . . . 67
3.3 Results . . . 67
3.3.1 Epigenome and TF binding patterns at promoters are predictive of gene-disease association . . . 67
3.3.2 Independent validations of predicted gene-disease associations 69
3.3.2.1 Validation using PubMed abstract mining . . . 69
3.3.2.2 Validation using the result of GWAS . . . 69
3.3.2.3 Survival analysis of predictive transcription factors and predicted genes for diseases . . . 70
3.3.3 Regulatory insights using the association between predictive TF and diseases . . . 71
3.3.4 Association of non-coding genes . . . 74
3.4 Discussion . . . 76

4 Conclusion **80**

4.1 Summary of contribution . . . 80
4.1.1 Chapter 2. Epigenome and TF binding patterns are predictive of ontology-based functions of coding and non-coding genes 80
4.1.2 Chapter 3. Explainable models using transcription factor binding and epigenome patterns at promoters reveal disease-associated genes and their regulators in the context of cell-types . . . 81
4.2 Future work . . . 82

LIST OF TABLES

2.1 Intersection of predicted genes with human disease-gene-sets. . . . 46

2.2 Intersection of predicted genes with mice disease-gene-sets. 47

2.3 List of predicted functions of non-coding RNAs with experimental evidence. 57

3.1 Cell type-specific association of TFs with diseases 73

LIST OF FIGURES

1.1 The distribution of human genes. 2

1.2 The distribution of human transcripts. 3

1.3 A. Represents the core promoter region, these include: BREu (B Recognition Element upstream), TATA-box (representing specific set of nucleotides), BREd (B Recognition Element downstream), Inr (initiator) region, motif ten elements (MTE), and downstream core promoter elements (DPE). B. depicts the proximal promoter region. TFBS1, TFBS2, and TFBS3 are the different transcription factor binding sites and CpG Island. 9

1.4 This figure represents how enhancer/silencer region interacts with the promoter region of a gene. 11

2.1 This flowchart depicts the workflow of the study. [Figure source: Chandra, Omkar, et al. "Patterns of transcription factor binding and epigenome at promoters allow interpretable predictability of multiple functions of non-coding and coding genes." "Computational and Structural Biotechnology Journal" 21 (2023): 3590-3603] 38

2.2 Demonstrates the robustness of epigenome profiles, specifically the predictive capacity of transcription factor binding patterns at promoters in gene function prediction. The figure includes four panels: A) Bar plot illustrating the count of gene-sets with strong predictions (80% sensitivity and 90% specificity) using five machine learning models. The upper panel highlights transcription factor ChIP-seq profiles, while the lower panel integrates five profile types. B) Box plots showing the area under the receiver operating characteristic curve (AUC-ROC) across all gene-sets, averaged over five-fold runs, with counts of gene-sets above 0.9 and between 0.8 to 0.9 indicated. C) Bar chart presenting the count of function sets with robust predictability using any of the five machine learning models. D) Plot validating the methodology, depicting the distribution of balanced accuracy achieved with false gene-sets (generated via random sampling) and experimentally annotated gene-sets, focusing on functions with balanced accuracy exceeding the 35th percentile. [Figure source: Chandra, Omkar, et al. "Patterns of transcription factor binding and epigenome at promoters allow interpretable predictability of multiple functions of non-coding and coding genes." "Computational and Structural Biotechnology Journal" 21 (2023): 3590-3603] 40

2.3 A) The count of functions (gene-sets) meeting the satisfactory prediction criteria (specificity 90%, sensitivity 70%). B) The count within the union set of gene-sets meeting the satisfactory prediction criteria across different machine learning (ML) methods. [Figure source: Chandra, Omkar, et al. "Patterns of transcription factor binding and epigenome at promoters allow interpretable predictability of multiple functions of non-coding and coding genes." "Computational and Structural Biotechnology Journal" 21 (2023): 3590-3603] 41

2.4 This figure depicts the clustering of functions based on their shared predictive TFs and cofactors derived from ChIP-seq profiles, illustrating potential overlaps in significant cellular processes. The tSNE plot and DBSCAN-based clustering visualization represent each gene-set as a point. The heatmap illustrates the similarity in the count of shared top predictors between two clusters: cluster-47, associated with cell cycle functions, and cluster-26, related to early developmental processes. The members of cluster-47 and cluster-26 are listed beneath the cluster plot. [Figure source: Chandra, Omkar, et al. "Patterns of transcription factor binding and epigenome at promoters allow interpretable predictability of multiple functions of non-coding and coding genes." "Computational and Structural Biotechnology Journal" 21 (2023): 3590-3603] 43

2.5 (Caption next page.) 45

2.5 Validation of novel predictions for function-gene associations: A) The box plot shows the frequency of co-occurrence between gene names and function terms in PubMed abstracts. On the left, it depicts the co-occurrence rates for predicted gene-function associations identified by GFPredict, while the right side illustrates the co-occurrence rates for randomly paired gene-function associations. Neither the novel predictions nor the random associations were included in the gene sets used for training or testing. B) A comparative evaluation of five methods for identifying associations between genes and functions. For the "Viability" cluster, GFPredict-predicted genes predominantly associated with DNA repair and cell cycle processes. Striped bars represent random gene scores, while solid bars represent predicted gene scores. No significant differences were observed for the clusters "chemical resistance," "pyroptosis," and "phagocytosis," which is crucial in immune responses. Similarly, for the "immune system" cluster, no significant differences were found between predicted and random associations in processes like "chemical resistance," "pyroptosis," and "peptide accumulation." [Figure source: Chandra, Omkar, et al. "Patterns of transcription factor binding and epigenome at promoters allow interpretable predictability of multiple functions of non-coding and coding genes." "Computational and Structural Biotechnology Journal" 21 (2023): 3590-3603] 46

2.6 Comparison of z-scores for newly predicted genes for gene-set "immune effector process" in various CRISPR screens. The red box represents the CRISPR z-scores in the phagocytosis CRISPR screen. The purple, blue, and green boxes represent the CRISPR z-scores of the same predicted genes in pyroptosis, resistance to chemicals, and peptide accumulation CRISPR screens, respectively. [Figure source: Chandra, Omkar, et al. "Patterns of transcription factor binding and epigenome at promoters allow interpretable predictability of multiple functions of non-coding and coding genes." "Computational and Structural Biotechnology Journal" 21 (2023): 3590-3603] 50

2.7 Provides insights into the co-occurrence and synergistic effects of Transcription Factor (TF) pairs as predictors. A) Depicts the distribution of TF ChIP-seq pairs among the top 20 predictors within the same cell type, categorized by functions (pink) and function clusters (green). The scatter plot on the right highlights these counts, with notable TF pairs such as C3: E2F4-GATA1, C4: MAZ-GATA1, F3: ZNF366-SPI1, and F4: SPI1-STAT1. B) Features a heatmap showcasing the statistical significance of TF ChIP-seq peak overlaps at promoters in GM12878 cells. C) Displays a box plot comparing the significance values (-log(P-value)) of promoter peak overlaps for TF ChIP-seq pairs that consistently appeared as top predictors across various functions in GM12878 cells. Additionally, the box plot on the right illustrates the significance of overlaps among random TF ChIP-seq profile pairs in GM12878 cells. [Figure source: Chandra, Omkar, et al. "Patterns of transcription factor binding and epigenome at promoters allow interpretable predictability of multiple functions of non-coding and coding genes." "Computational and Structural Biotechnology Journal" 21 (2023): 3590-3603] 52

2.8 The predictive pleiotropy of TFs is illustrated as follows: (A) Shows the distribution of functions where a single TF emerged as a top predictor based on GM12878 cells. The size of the dots corresponds to the feature importance score, while the color indicates the directionality of the relationship. (B) Presents a dot plot highlighting the count and feature importance of TF ChIP-seq profiles ranked as top predictors in GM12878 cells. Only functions specific to GM12878 cells are included. [Figure source: Chandra, Omkar, et al. "Patterns of transcription factor binding and epigenome at promoters allow interpretable predictability of multiple functions of non-coding and coding genes." "Computational and Structural Biotechnology Journal" 21 (2023): 3590-3603] 54

2.9 A) Comparison of CRISPR scores between the top 30 genes predicted by the GFPredict modeltrained on the top 50 genes from CRISPR screensand a set of random genes. The top 30 predicted genes were excluded from the training dataset. B) CRISPR scores of lncRNA genes among the top 30 predicted genes in the lncRNA-CRISPR screen for the cell cycle, identified by GFPredict, which was trained on the top 50 positive coding genes from a different cell-cycle CRISPR screen. Out of the top 30 predicted genes, 15 were lncRNA genes. C) Comparison of CRISPR scores between 52 lncRNA genes predicted to belong to the cell cycle cluster (cluster-47, as shown in Fig. 3) and random genes in the lncRNA-CRISPR screen for the cell cycle. [Figure source: Chandra, Omkar, et al. "Patterns of transcription factor binding and epigenome at promoters allow interpretable predictability of multiple functions of non-coding and coding genes." "Computational and Structural Biotechnology Journal" 21 (2023): 3590-3603] 56

3.1 This figure presents the outcomes of our prediction model for identifying associations between diseases and genes: A) Gene-set memberships for various diseases were predicted using five distinct machine-learning models. The panel showcases the Operating Characteristic (ROC) values derived from predictions across 3,675 disease gene-sets. B) This section illustrates the distribution of ROC-AUC values, highlighting comparisons with other methodologies. The panel emphasizes the proportion of predictions that reveal novel associations through our approach. [Figure source: Chandra, O., Pramanik, D., Gautam, S., Sharma, M., Dubey, N., Mahato, B. and Kumar, V., 2024. "Explainable models using transcription factor binding and epigenome patterns at promoters reveal disease-associated genes and their regulators in the context of cell-types." bioRxiv, pp.2024-05] 68

3.2 Validation using PubMed abstracts mining: The vertical axis of this figure depicts the frequency of disease names co-occurring with non-coding gene names in PubMed abstracts. For comparison, the figure also displays the co-occurrence frequencies of randomly paired disease names and non-coding gene names in PubMed abstracts. B) Validation Using GWAS Data: This section of the figure presents the validation outcomes utilizing GWAS-derived mutations in genes predicted to be associated with specific diseases. Only GWAS mutations corresponding to the predicted diseases were considered. As a control, the number of GWAS mutations associated with the target disease was assessed in randomly selected genes. [Figure source: Chandra, O., Pramanik, D., Gautam, S., Sharma, M., Dubey, N., Mahato, B. and Kumar, V., 2024. "Explainable models using transcription factor binding and epigenome patterns at promoters reveal disease-associated genes and their regulators in the context of cell-types." bioRxiv, pp.2024-05] 70

3.3 Survival analysis of genes associated with bladder urothelial carcinoma (TCGA-BLCA). The box plot on the left illustrates the statistical significance (-log(P-value)) of survival association. On the right, Kaplan-Meier plots depict the survival outcomes for genes predicted to be linked with TCGA-BLCA. [Figure source: Chandra, O., Pramanik, D., Gautam, S., Sharma, M., Dubey, N., Mahato, B. and Kumar, V., 2024. "Explainable models using transcription factor binding and epigenome patterns at promoters reveal disease-associated genes and their regulators in the context of cell-types." bioRxiv, pp.2024-05] 71

3.4 Five-fold cross validation result of random forest model train on disease-gene-sets using TFs as features. bioRxiv, pp.2024-05] 72

3.5 Screenshot from the UCSC Browser. The screenshot from the UCSC Browser illustrates the genomic positions of two non-coding genes whose functions were predicted using our methodology: A) Genomic location of MIR137HG gene. B) Genomic location of LINC00877 gene. [Figure source: Chandra, O., Pramanik, D., Gautam, S., Sharma, M., Dubey, N., Mahato, B. and Kumar, V., 2024. Explainable models using transcription factor binding and epigenome patterns at promoters reveal disease-associated genes and their regulators in the context of cell-types." bioRxiv, pp.2024-05] . . 76

ABBREVIATIONS

TF	Transcription Factor
ncRNAs	Non-coding RNAs
GO	Gene ontology
ML	Machine learning
lncRNA	Long non-coding RNAs
DNase-seq	DNase I hypersensitive sites sequencing
ChIP-seq	Chromatin Immunoprecipitation Sequencing
CAGE tags	Cap analysis of gene expression tags
CRISPR	Clustered regularly interspaced short palindromic repeats
hPSC	Human pluripotent stem cells
XGBoost	Extreme gradient boosting
RF	Random forest
TSS	Transcription start sites
SVM	Support vector machines
GWAS	Genome-wide association studies
AUROC	Area under the receiver operating characteristics curve

CHAPTER 1

INTRODUCTION

The genome is the library of a cell, containing all the information needed for the cell to survive and divide. The quest of genome biology has been to identify all the instructions present in the human genome and to understand how a cell uses them to synthesize biomolecules to perform functions essential for its survival. Mapping out all the functional elements of the human genome would provide the blueprint required to manipulate the biological processes of a cell, thereby gaining control over the physiology of the organism.

The complexity of the genome can be perceived by acknowledging the total number of functional units of DNA (genes) and its interactions with other biomolecules (RNA, protein). A comprehensive research into the human genome commenced with the Human Genome Project (1). This initiative uncovered that the human genome comprises a larger number of genes, domains, and protein families, in addition to paralogues, multidomain proteins with diverse functions, and various domain architectures. The empirical evidence provided by this project gave a glimpse into the complexity of the human genome.

Over the two decades (2000-2020) of research into the human genome, along with the development of novel biotechnology techniques and computational methods, we are beginning to understand the true complexity of the human genome. This includes the identification of novel genes, elucidating the structure of the genome, functional mapping of genes in different biological contexts, and the underlying molecular mechanisms by which these genes interact with other biomolecules to drive genome function. However, most of these aspects of the human genome remain unanswered.

This thesis work primarily contributes to the functional mapping of regions of the human genome in the context of biological functions and disease conditions by developing a novel computational framework using epigenome and transcription factor binding patterns at the promoters of genes.

This introductory chapter discusses the knowledge we have so far about the human genome that is relevant to this thesis work, and the critical insights that can be computationally exploited to infer novel gene-function and gene-disease associations.

1.1 The human genome

The ENCODE (Encyclopedia of DNA Elements) project was a significant endeavor by the scientific community to identify and catalog every functional element within the human genome. The subsequent project, GENCODE, aimed to extend the annotation of coding and non-coding genes, including alternatively spliced transcripts and pseudogenes (2; 3).

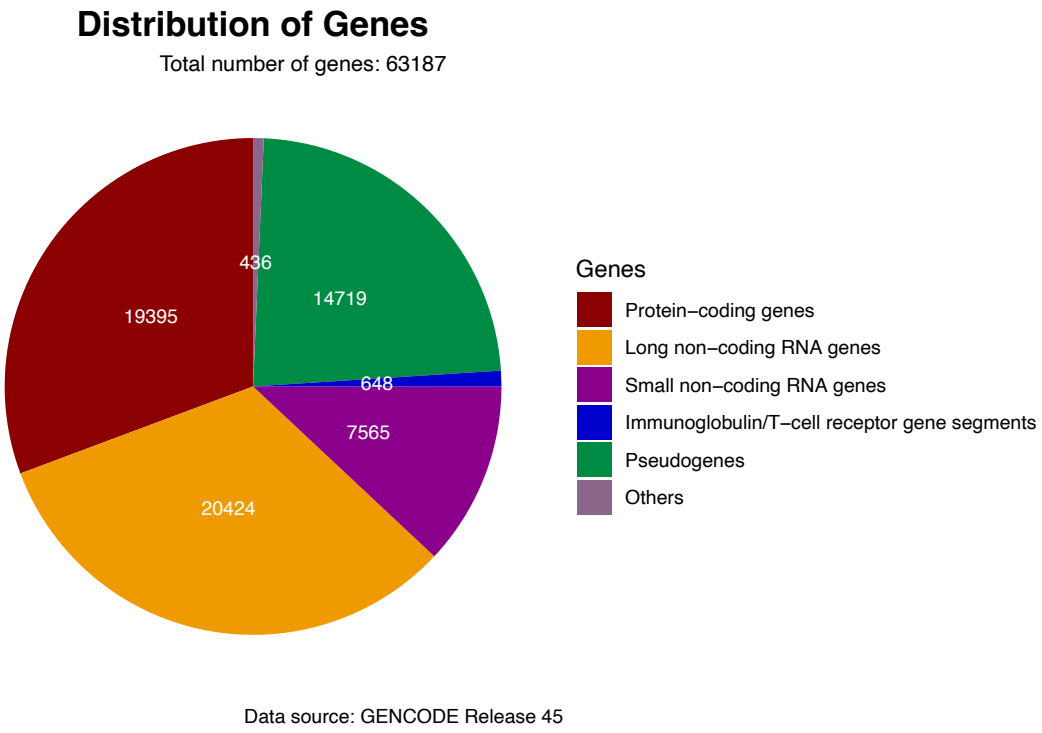

Figure 1.1: The distribution of human genes.

According to GENCODE Release 45 (Figure 1.1), the human genome comprises 63,187 genes. These genes are categorized based on their nucleotide composition and the type of product (RNA or protein) they encode. A segment of DNA is considered a gene if it is transcribed into RNA. Of these genes, 19,395 encode functional proteins, while non-coding genes' RNA does not translate into protein. The major category of non-coding genes consists of long non-coding RNA genes, totaling 20,424 genes with

sizes varying from 40 to 200 nucleotides. Additionally, there are 7,565 small non-coding genes. Pseudogenes, non-functional copies of coding genes resulting from gene duplication events and reverse transcription of coding gene RNAs, make up a considerable portion of the genome (4). Another minor class of genes is the T-cell receptor genes.

The subsequent subsections will delve into the biological functions and mechanisms of these non-coding genes in more detail.

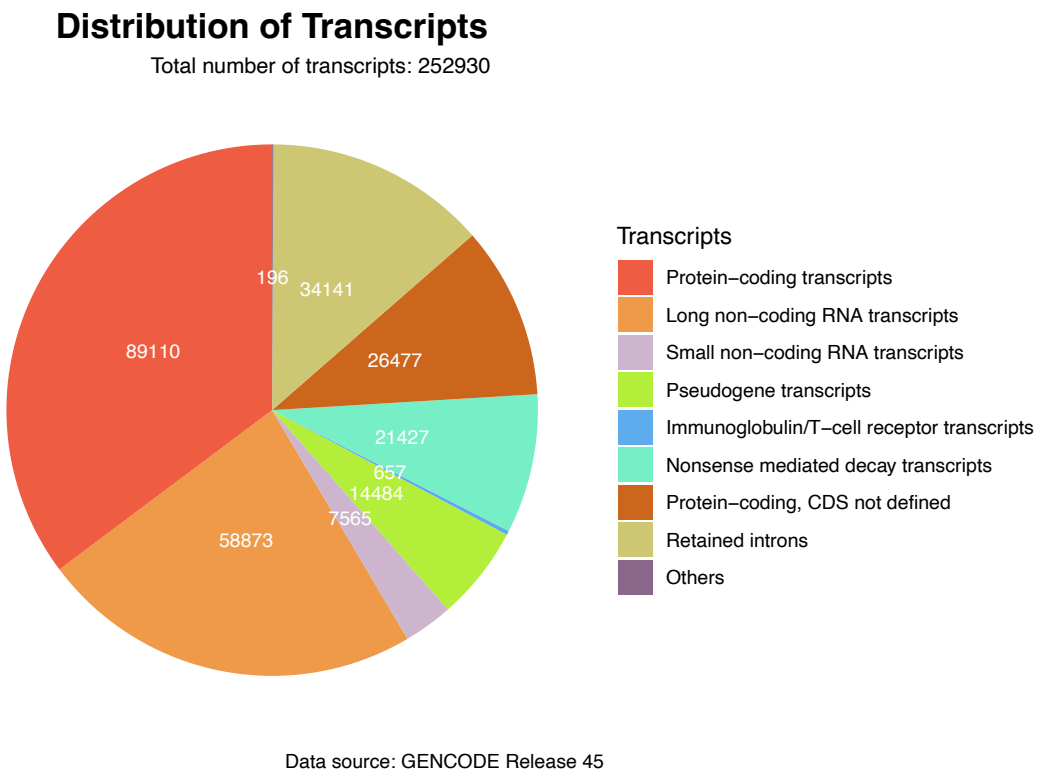

Figure 1.2: The distribution of human transcripts.

The 63,187 genes in the human genome encode up to 252,930 RNA transcripts (Figure 1.2), a number greater than the total number of genes. This discrepancy arises because a single gene can encode multiple types of RNA transcripts due to alternative transcription events, which occur when there is more than one transcriptional initiation or termination site within a gene. The largest number of RNA transcripts are encoded by protein-coding genes, followed by long non-coding genes.

1.2 Incomplete annotation of molecular and biological functions of the genes

Thousands of novel genes have been identified and annotated with their physical location in the human genome. Decades of experimental and computational research have greatly enhanced our understanding of gene function. The Gene Ontology consortium was established to systematically represent the functional knowledge of genes. The consortium aims to capture the full complexity of biology by using standardized vocabularies to describe the molecular functions of genes, their involvement in larger biological processes, and their localization within cells (5). Genes are annotated under subclasses that fall into one of three main domains:

- Biological Process: Includes terms that capture molecular events or interactions that contribute to larger biological functions, such as cell division and signal transduction.
- Molecular Function: Includes terms that describe the activities of individual genes and their products at the molecular level, such as receptor binding and enzymatic functions.
- Cellular Component: Encompasses terms describing the subcellular locations where gene products are transported and active.

Genes are annotated under these domains based on various types of evidence, including experimental data, phylogenetic analysis, computational predictions, and curatorial statements. One important feature of gene ontology is its ability to enable the innovative annotation of both coding and non-coding genes through accumulating evidence, covering terms such as receptor binding and cell division (6).

From our analysis, out of 27,989 non-coding genes, approximately 4,500 genes have been annotated in the gene ontology. Growing evidence suggests that many non-coding genes are potentially functional, playing roles in regulating other protein-coding and non-coding genes. More details regarding the non-coding genes are described in the subsequent section. Therefore, extensive functional annotation of non-coding genes into different molecular and biological functions is needed (7). For protein-coding genes, their biological functional annotation is incomplete in terms of pleiotropy. Pleiotropy is a phenomenon in which a gene can have biological roles in unrelated functions (8). Therefore, there is a need for a more comprehensive functional annotation of

genes in the gene ontology to capture the real complexity of biology.

1.3 Incomplete annotation of disease genes

Somatic mutations (changes in the DNA sequence that occur after fertilization) alter the behavior of genes, affecting their expression or molecular function in disease conditions. The deviation in the normal behavior of genes harboring such somatic mutations results in alterations in normal biological processes, leading to disease conditions in humans such as cancer and autoimmune diseases, including developmental abnormalities (9; 10; 11). Most disease conditions, especially cancers, result from mutations in multiple genes; therefore, knowledge of those causal genes is crucial to understanding the pathology of diseases and effective clinical treatment (12; 13). Additionally, there are genes whose function changes because of second-order consequences caused by the activity of causal genes in a diseased cell (14; 15).

To effectively annotate the knowledge of disease genes, it is crucial to comprehensively store them in a database, allowing researchers easy exploration, retrieval, and analysis. Several such databases exist, including the KEGG DISEASE database (16), DISEASES (17), and DisGeNET (18); the largest among them is DisGeNET (v7.0), containing 1,134,942 gene-disease associations. DisGeNET aims to facilitate researchers' easy access to the exploration and analysis of the genetic underpinnings of human diseases.

Most human diseases remain incurable due to a lack of complete genetic understanding. Therefore, there is a need to identify all the causal genetic factors of these diseases.

1.4 Characteristics of the genes

To classify genes into respective ontological classes, it is essential to understand their characteristics, such as molecular and biochemical signatures and sequence information. This understanding helps compare genes of known ontological functions to those with unknown functions. The definition of genes has evolved since the classical age of

genetics. Portin et al. extensively detailed the evolving definition of the term "gene" (19). During the classical period of genetics (1900-1930), genes were primarily described based on their hereditary properties, chromosomal location, and their impact on phenotypic traits. However, with the emergence of evidence during the neoclassical period (the 1940s) of genetics, the concept of a gene evolved to incorporate the phenomenon of one geneone mRNAone polypeptide.

The current era of molecular genetics has revealed the following features of human genes:

1. There is up to 25% overlap among protein-coding genes in the genome (20).
2. A gene can have more than one transcription start site, resulting in transcript isoforms (21) and also because of post-transcriptional splicing events (22; 23).
3. Both coding and non-coding genes are transcribed from either DNA strand (24).
4. Chimeric transcripts are encoded by 4-5% of the tandem gene pairs (25).
5. The presence and function of extrachromosomal genes (23).

In light of recent evidence detailing the characteristics of genes, a gene can be defined as "a DNA sequence that encodes an RNA transcript or polypeptide, influencing a phenotype."

Other characteristics can be attributed to a gene based on the way and types of proteins by which it is regulated. The DNA structure of a gene can be divided into two regions: the coding region and the regulatory region. The DNA elements in the regulatory regions (described in detail in the next subsection) control gene expression. The coding region contains DNA elements that encode RNA molecules. The extensive knowledge gained from analyzing the gene's regulatory region in relation to gene expression and phenotype regulation is substantial, offering valuable insights for basic genome research.

1.4.1 Regulatory DNA elements (cis-regulatory elements)

The regulatory DNA elements are located upstream of genes' coding sequences (CDS). They are nucleotide sequences that interact with transcription initiation factors and other cis-regulatory elements, such as promoters. Understanding and characterizing the regulatory elements help us to comprehend the collective roles of different genes in larger

cellular processes based on the similarity of their regulation, as described in Chapter 2. Additionally, this knowledge provides insight into the dysregulation of genes in disease conditions. Mapping out all the regulatory elements of the genes provides an anchor point for genetic manipulation using their regulatory factors.

Regulatory sequences are classified into different types based on their molecular functions as follows:

1.4.1.1 Transcription start sites (TSSs)

A gene's transcription start site (TSS) is the first nucleotide that is transcribed into the $5'$ end of its resulting RNA transcript. TSSs are present in the core promoter sequences of the genes. A gene can have more than one TSS, indicating the presence of alternative promoter sites, which result in the transcription of multiple isoforms of a gene (26). Therefore, identifying TSSs in the human genome helps pinpoint promoters and enhancer sites for investigating gene regulation. The identification of TSSs is valuable for exploring the various transcript and protein isoforms of a gene within the contexts of development and disease (27).

Biotechnology Assays to Detect Global TSSs

Oligo-Capping

Oligo-capping was initially developed by Maruyama et al. for capturing complete mRNAs with the $5'$ end (28). Oligo-capping allows for the adapter ligation that the polymerase enzyme requires for the reverse transcription for cDNA synthesis and sequencing. $5'$ Serial gene expression analysis ($5'$ SAGE) was the first technique developed for analyzing TSSs genome-wide (29). Later, high-throughput sequencing protocols were adopted to sequence the $5'$ end-tagged mRNAs on different sequencing platforms (30). Ni et al. developed the Paired-End Analysis of TSSs (PEAT) technique for oligo-capping in paired-end sequencing (31). However, oligo-capping methods are prone to sequence and structure biases in RNA ligases when adding adapters (32).

Cap-Trapping

Cap-trapping protocols were originally developed to create libraries of $5'$-complete cDNAs, where the $5'$ cap is oxidized and biotinylated to enable streptavidin purifica-

tion of these cDNAs following reverse transcription (33). One widely used application of cap-trapping is cap analysis of gene expression (CAGE). In CAGE, the initial step involves reverse transcription and cap-trapping, followed by ligation of a $5'$ linker containing restriction sites for XmaJI and MmeI enzymes. Next, the second strand is produced, and the resulting double-stranded cDNA is digested using MmeI. This enzyme creates an overhang at the $3'$ end of its recognition site, allowing ligation with another adapter containing an XmaJI site. Subsequent cleavage by XmaJI releases the ligated cDNA fragments (CAGE tags), which are then concatenated, cloned, and sequenced. (34).

Mapping TSSs from nascent RNA

Profiling TSSs from nascent transcripts is essential because many of the transcripts are short-lived, have a rapid turnover rate, and are underrepresented in the mature RNA pool of a cell. Nuclear run-on reactions are carried out to extend the incompletely transcribed RNAs where RNA polymerase enzymes are also binding (35; 36). The global run-on sequencing (GRO-seq) protocol employs nuclear run-on sequencing technology, where nuclei are isolated, nascent RNA is extended with 5-bromouridine $5'$-triphosphate (BrUTP) incorporation, and nascent RNA is captured using anti-BRUTP antibodies (37). The precision nuclear run-on and sequencing (PRO-seq) protocol improve upon GRO-seq, where biotin-labeled ribonucleotide triphosphate is used for the elongation of the incompletely transcribed RNAs. Additionally, only one of the four nucleotides is supplied in four separate individual run-on reactions, allowing for base-pair resolution (38).

1.4.1.2 Promoters

The promoter of a gene serves as the regulatory region located immediately upstream of the CDS region of the gene. The core promoter region is typically considered to span approximately 50 base pairs (bp) downstream and/or 50 bp upstream of the TSS (39). Core promoter elements usually include a combination of the following DNA elements (Figure 1.3A): initiator (Inr), BREd (downstream), BREu (upstream), motif ten elements (MTE), CpG islands, TATA-box, transcription factor IIB (TFIIB) recognition elements, and downstream core promoter elements (DPE) (40; 41; 42; 43; 44).

Additionally, the proximal promoter region, situated 40-250 base pairs upstream of the transcription start site (TSS), contains sites where transcription factors bind (illustrated in Figure 1.3B). These transcription factors can exhibit widespread distribution, cell-type specificity, or binding that depends on the developmental stage (45).

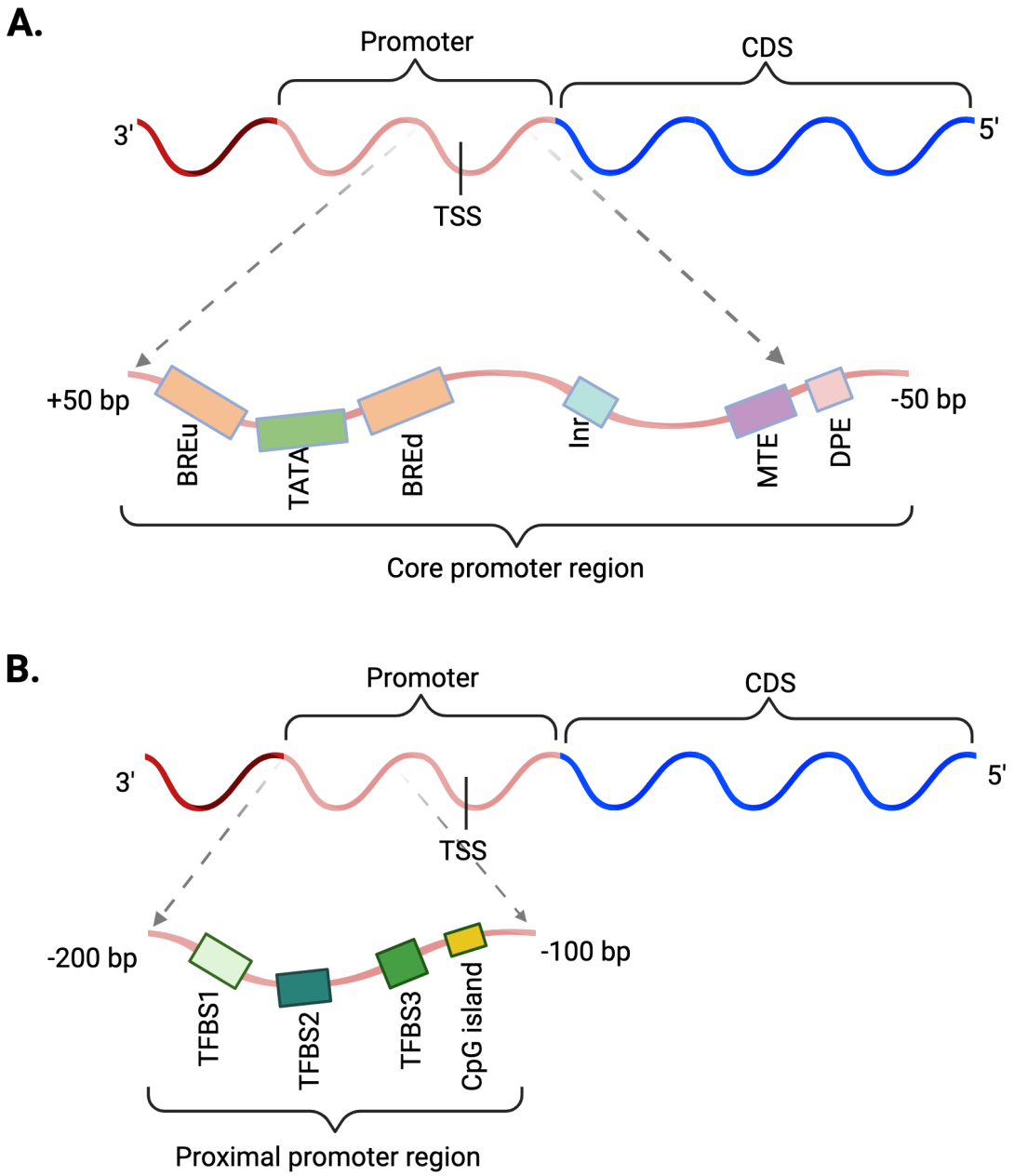

Figure 1.3: A. Represents the core promoter region, these include: BREu (B Recognition Element upstream), TATA-box (representing specific set of nucleotides), BREd (B Recognition Element downstream), Inr (initiator) region, motif ten elements (MTE), and downstream core promoter elements (DPE). B. depicts the proximal promoter region. TFBS1, TFBS2, and TFBS3 are the different transcription factor binding sites and CpG Island.

Core promoter elements are pivotal in recruiting general transcription factors (GTFs)

and facilitating the assembly of the pre-initiation complex (PIC) (44). In metazoans, core promoters are classified into three types based on their initiation patterns, motifs, chromatin structure, and gene function (46).

1. Focused Promoters: These promoters are predominant in terminally differentiated adult cells and are enriched with key regulatory elements near their TSS. Typically, they feature a single TSS embedded in their sequence, leading to the formation of a 'focused' or 'sharp' peak, resulting in the production of a single mRNA from the downstream CDS region (46; 47).
2. Dispersed Promoters: Found in housekeeping genes expressed across many cell types, dispersed promoters exhibit 'dispersed' transcription initiation, meaning they contain multiple closely spaced TSSs (48; 49).
3. Developmental Promoters: These core promoters are linked to essential transcription factor genes that play roles in early developmental processes such as morphogenesis and patterning. They are distinguished by bivalent marks including both H3 Lys 4 trimethylation (H3K4me3) and H3 Lys 27 trimethylation (H3K27me3) (50).

Additionally, up to 70% of genes' proximal promoters contain CpG islands (51), with the DNA sequence (1-2 kb) being rich in cytosine and guanosine. The presence of epigenomic marks, especially H3K4me3, at CpG sites is associated with regulating the transcription process (52).

1.4.1.3 Enhancers

Enhancers, distal cis-regulatory elements, bind to transcription factors and co-factors, thereby enhancing the activity at the core promoter region of target genes, regardless of their position relative to the target genes, whether upstream or downstream. Unlike promoters, enhancers can be located as far as one megabase pair from their target genes in eukaryotes (53). Promoters show basal activity necessary to initiate transcription, whereas enhancers recruit transcription factors and cofactors to augment transcription from core promoters (54; 55). Enhancers exhibit tissue-specific activity (56).

Transcription from enhancers results in the production of enhancer-RNAs (eRNAs) (57; 58). General and context-specific transcription factors bind to enhancer loci, facilitating enhancer sequence transcription. The transcriptional activity or the transcripts (eRNAs) from enhancers regulate the expression of target genes (59). There exists a

correlation between eRNA expression levels and the mRNA expression levels of target genes (58; 60). Knockdown experiments have demonstrated the significant role of eRNAs in regulating their target genes (61; 62).

The interaction looping between enhancers and promoters is a crucial mechanism by which enhancers regulate target genes (Figure 1.4). Chromosome conformation capture (3C) assays offer evidence for identifying these loops between promoters and enhancers. Transcription factors binding at enhancers and promoters are crucial for forming these loops, facilitating proximity interactions (63; 64). While some studies suggest that enhancer-promoter proximity is essential for target gene transcription (65; 66), others indicate that ectopic eRNA expression can increase target gene expression without requiring enhancer-promoter proximity (65). This diversity in mechanisms contributes to the difficulty in associating enhancer sequences with putative target genes, especially considering that enhancers can act on genes located thousands of bases away (53).

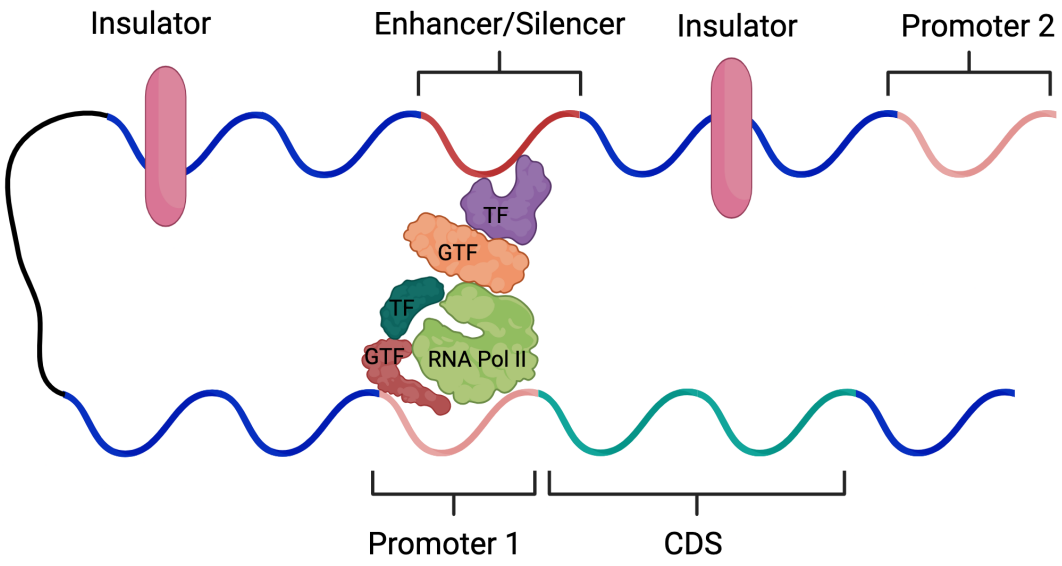

Figure 1.4: This figure represents how enhancer/silencer region interacts with the promoter region of a gene.

Super enhancers are large genomic segments where dense clusters of enhancers located on a single DNA strand. Individual enhancers within super enhancers work synergistically, acting as a single regulatory unit that governs the transcription of target genes. For further information on super enhancers, the review article by Tang et al. (67).

Identifying enhancers and understanding their function present significant chal-

lenges for several reasons. Enhancers are distributed throughout 98% of the genome, and their positions relative to target genes vary widely, appearing upstream or downstream of target genes, in intergenic and intronic regions, or even within unrelated gene introns (68). Current approaches to identifying enhancer regions include characterizing eRNA transcription regions (60), assessing the presence of histone acetyltransferase p300/CBP (69; 60), and examining the presence of RNA Pol II along with H3K4me1/2 and H3K27Ac marks (70; 71; 72).

Enhancers can exert regulatory effects on genes that extend beyond immediate neighboring coding genes, occasionally skipping several neighboring genes to regulate distant ones (73). A single enhancer has the capability to regulate multiple genes, while a single target gene can be regulated by multiple enhancers (74). The complexity of these interactions presents significant challenges in associating enhancers with their putative target genes.

Some research groups have utilized clustered regularly interspaced short palindromic repeats (CRISPR) interference (CRISPRi) to establish connections between enhancers and gene promoters (75). Genome-wide Hi-C experiments have also become popular for mapping enhancers to target regions (76; 74). For deeper insights into the functional characterization of enhancer-promoter interactions, check the review article by Mach et al. (77).

1.4.1.4 Silencers

Like enhancers, silencers are distal regulatory regions; however, they contrastingly repress the activity of the target promoter (78). Silencers exert their action on promoters independently of their orientation (downstream or upstream of the target gene) and position, even if located thousands of base pairs away (79). They embed sequences to bind regulatory factors that are transcriptionally repressive in nature (Bruce et al., 2004). Silencers induce the repression of enhancer-targeting promoters. For example, immature T cells express both CD8 and CD4 cell surface glycoproteins, but as the cells mature, only CD8 is expressed, with CD4 expression suppressed by the silencer elements (80).

Silencers are identified using the biochemical signature they harbor. Trimethylation of histone H3 lysine 27 (H3K27me3) serves as a transcriptional repressive mark.

Silencers with an open-chromatin conformation marked by H3K27me3 are considered active silencers by Huang et al. (81). Researchers also correlate nearby gene expression with the combined signal of H3K27me3 ChIP-seq and DNase I hypersensitivity sites (81). While silencers are typically found in nucleosome-depleted regions (open-chromatin regions) as active regulatory elements binding to repressive transcription factors, distinguishing them from other cis-regulatory elements like promoters and enhancers remains challenging. To identify silencers, researchers have developed methods that combine histone markers with open-chromatin profiles. Pang and Snyder, for instance, employed a parallel reporter assay for this purpose. They reasoned that silencers should be located within open chromatin regions. To test this, they cloned chromatin-accessible segments of around 200 base pairs from K562 cells into a lentiviral plasmid, which carried a modified FKBP-Casp9 gene designed to trigger apoptosis. This plasmid was subsequently introduced into K562 cells. If the candidate regions contained a silencer, it would repress the expression of the FKBP-Casp9 gene. Conversely, if the candidate region lacked a silencer element, it would result in the expression of FKBP-Casp9 and apoptosis of the cell (82).

Histone 4 Lysine 20 monomethylation (H4K20me1) is notably concentrated at silencers and functions as a cell cycle-related mark that regulates gene expression (83). Furthermore, certain types of silencers are linked with polycomb repressive complexes (PRC1 and 2) (83; 84).

Doni et al., in their approach to identifying and characterizing silencers, employed simple subtractive analysis. Open-chromatin profiles that did not overlap with already known cis-regulatory elements were considered as uncharacterized CREs, hypothesizing that these uncharacterized CREs should contain silencers. Silencers were determined based on the presence of repressor TFs signal and validated using massively parallel reporter assays (85).

The bifunctional activity of silencers, exhibiting properties of enhancers, has been reported in the literature. Sequences previously identified as silencers in yeast were later found to exhibit transcription-enhancing activity (86; 87; 88; 84). This dual functionality of silencers, acting both as silencers and enhancers, could be facilitated by transcription factors functioning as activators and repressors depending on different contexts (89) (Figure 1.4). Therefore, identifying and characterizing silencers is challenging, partic-

ularly given that silencers lack signature chromatin marks. For more information on the detailed mechanism of action and evolution of silencers, the review article by Segert et al. (90).

1.4.1.5 Insulators

The regulation of target gene transcription depends on gene promoters and interacting cis-regulatory enhancer elements, regardless of their relative position and orientation to the target genes (Enhancers 1.4.1.3). Given that these cis-regulatory elements, such as enhancers, are spread across the genome (91), cells must precisely control interactions between gene promoters and their regulatory enhancers to ensure appropriate function. This mechanism, known as insulation, works by blocking enhancer-promoter interactions. Insulation is facilitated by cis-regulatory DNA elements called insulators, along with their associated proteins, termed insulator-binding proteins (IBPs) (92; 93). A key characteristic of insulators is their barrier function, which protects genes from the silencing effects of neighboring condensed chromatin. Thus, unlike silencers, insulators safeguard genes from unnecessary activation and deactivation mechanisms by regulatory elements (94; 95).

Insulators in mammalian cells were initially identified by Chung et al. in human erythroid cells, positioned near the 5 boundary region of the chicken beta-globin gene. They observed that this DNA element provides directional insulation and alters the chromatin structure around the gene's promoter (96). The first identified insulator-binding protein (IBP) was the CCCTC-binding factor (CTCF), discovered in purified chicken cell extracts (97). The concept of DNA element insulators is closely tied to their associated binding proteins.

CTCF is the major class of IBPs found across insulators in the human genome. Cohesin proteins are IBPs that associate with CTCF at their binding sites (98). CTCF physically interacts with cohesin to keep them bound at the target site (99). CTCF, in partnership with cohesin, regulates inter- and intra-chromosomal interactions involved in transcriptional regulation mediated by insulators (99; 100; 101).

CTCF interacts with members of chromodomain helicases, such as CHD8, and is implicated in regulating target genes like BRCA1 and C-MYC (102). CTCF also in-

teracts with various molecules, including transcription factors like YY1, YB1, Oct4, Kaiso, and thyroid hormone receptors, which collectively regulate the transcription of target genes (103; 104; 105; 106). Additionally, CTCF binds to nucleolar proteins such as nucleophosmin and lamin to localize target regions toward the nucleolus (107; 108). Thus, IBPs interact with various protein molecules to regulate transcription by reorganizing the chromatin structure.

Insulators are chromatin remodelers. IBP CTCF is used as a proxy for identifying insulator sites in the human genome. ChIP-seq data of the CTCF protein indicates that CTCF binds at chromosome transition regions crucial for X inactivation during early development (109). Further high-throughput experiments showed that CTCF binds at the flanking sites of active chromosomal domains, marked by H3K27 trimethylation (110). Insulator-binding CTCF participates in segregating repressive chromatin domains, with this CTCF activity pattern being specific to cell types (111; 112). Chromosome conformation capture assays have demonstrated that CTCF facilitates chromatin looping, thereby regulating gene expression (100; 113). The binding of CTCF in regulating gene expression depends on the DNA sequence's methylation status (114). For more detailed insights into the mechanisms of insulator, enhancer, and IBP regulation, as well as their implications in disease, consult the review by Yang and Corces (115) and the research article by Ribeiro-Dos-Santos et al. (116).

1.4.2 Biochemical signatures

The cis-regulatory DNA elements described above interact with various proteins and are characterized by the methylation of specific nucleotides in certain regions, depending on their functional context. These biochemical features at genomic regions play active roles in modulating transcription and replication. The following subsections discuss some of the biochemical signatures present across different regulatory DNA elements.

1.4.2.1 Transcription factors and co-regulators

Transcription factors (TFs) are proteins that selectively bind to specific regions of the DNA sequence (117). A comprehensive catalog includes over 1,600 TFs, most of which have been extensively studied, including their binding motifs (the specific DNA se-

quences they recognize) (118). These factors play a fundamental role in interpreting the DNA sequence by binding to specific regions and influencing chromatin structure and gene expression (119; 120).

Studying transcription factors involves identifying and elucidating the DNA sequences (motifs) to which they bind. Determining the location of these motifs is the first step in characterizing a TF functionally. However, establishing functional connections between transcription factors (TFs) and potential target genes is challenging, since TF binding at a gene's promoter or enhancer site does not always imply direct regulation of the gene's expression by that TF (121). Since the size of motifs for a TF is typically 6-12 nucleotides long, the chances of such sequences being present for different TFs within a gene are high, leading to non-functional binding of TFs (122).

This problem of functional characterization of TFs, where only the functional binding at DNA regions is considered, can be addressed by taking into account the fact that TFs work in synergy and cooperativity with other TFs and co-factors (123).

Transcription factors (TFs) are categorized into major classes or families according to the type of DNA-binding domain they possess:

1. Basic domain TFs: These transcription factors contain basic amino acids in their DNA-binding domain and include subfamilies such as Basic leucine zipper factors (bZIP) and Basic helix-loop-helix factors (bHLH) (124).
2. Zinc-coordinating DNA-binding domain TFs: These TFs contain repetitive patterns of cysteine and histidine residues and have zinc-dependent properties (125).

Similarly, there are other families of transcription factors. For more information, the review by Wingender et al. (126).

Various methods are employed to identify transcription factors (TFs), including one-hybrid assays (127) and DNA affinity purification-mass spectrometry (128). Typically, hidden Markov models (HMMs) are utilized to detect the protein sequences' DNA-binding domains in TFs and categorize them. For further details, consult databases such as Pfam, SMART, and InterPro (129; 130; 131). Various biotechnological assays like ChIP-seq and microfluidics are employed to identify the motifs of TFs. For detailed information on such assays, the review article by Stormo and Zhao (132; 133; 134).

One of the most important and complex properties of TFs is their ability to cooperatively interact with each other and other co-regulators/co-factors to form multiprotein

complexes that transduce signals regulating the transcription of specific genes (135). The formation of multiprotein complexes is crucial to activate and transcribe only the relevant genes for a phenotypic outcome without activating other phenotypically unrelated genes.

Co-regulators are classified into two types, co-activators and co-repressors, based on their intrinsic properties to either aid in the activation or repression of transcription, respectively (136). Co-activators modulate biochemical events that favor the transcription of a gene, such as acetylation and demethylation (137; 138). Conversely, co-repressors induce events that repress gene transcription, such as deacetylation and methylation (139; 140). For more information on the biological roles of co-regulators, the review by Talukdar and Chatterji (141).

Transcription factors and co-regulators collaborate in various ways, including forming homodimers, trimers, and higher-order structures (142; 143). This synergistic interaction between TFs and co-regulators regulates gene transcription, thereby influencing phenotypic outcomes.

1.4.2.2 DNA methylation

DNA methylation in the human genome entails the covalent addition of a methyl group to the 5th carbon atom of cytosine residues at cytosine and guanine (CpG) sites by DNA methyltransferase (DNMT) enzymes (144). CpG islands, which are regions abundant in CpG sites, are predominantly located in gene promoter regions (145). While the function of methylated non-CpG sites is still under investigation, there is evidence suggesting their involvement in the epigenetic regulation of genes associated with mitochondrial function and fuel utilization (146; 147). The methylation of CpG islands is biologically significant for gene silencing, genomic imprinting, X-chromosome inactivation, and transcriptional regulation (148; 149; 150; 151; 152). The aspect of transcriptional regulation by DNA methylation is particularly pertinent to this thesis.

DNA methylation occurs approximately at 1% of the human genome (153). Rather than directly inhibiting gene transcription, methylated DNA is bound by methyl-CpG binding proteins, which hinder transcription by competing with transcriptional activators (154). The primary classes of methyl-CpG binding proteins include MBD pro-

teins, zinc-finger proteins, and UHRF proteins (155; 156; 157). MBD proteins possess domains for transcriptional repression, enabling them to interact with various transcriptional repressor complexes and thereby inhibit transcription initiation (158). DNA methylation interacts significantly with other epigenetic mechanisms, such as histone modifications, microRNAs, and transcription factors, in the regulation of transcription (159; 160; 161; 52). For a more comprehensive understanding of DNA methylation and its implications in disease, consult the review by Moore et al. (162). Consequently, DNA methylation at CpG islands plays a crucial role in transcriptional regulation.

1.4.2.3 Histone modifications

The fundamental unit of chromatin structure is the nucleosome, composed of a histone octamer. Each octamer includes two copies each of H2A, H2B, H3, and H4, around which approximately 150 base pairs of DNA are wrapped (163). Histones undergo post-translational modifications with various chemical groups on their lysine chains, known as epigenomic marks.

The initial evidence linking histone modification to RNA synthesis was presented by Allfrey et al. in 1964. Their work demonstrated that acetylation of histone proteins influences the process of RNA synthesis (164). The intricate nature of histone methylation's effect on transcription became clearer much later, following the discovery of methyltransferase enzymes. Histone methylation's impact on transcription varies depending on the specific lysine position that undergoes methylation (165; 166), illustrating intricate regulatory mechanisms.

Another crucial aspect of histone modification involves regulatory effector proteins that bind to acetylated or methylated histone sites (167). Bromodomain and YEATS-domain proteins recognize acetylated lysine residues on histones, whereas chromodomain, MBT, PWWP, and Tudor domains identify methylated lysine residues on histones (168; 169).

The transcriptional state is directly influenced by the combination of histone modifications on the histone proteins of the nucleosome structure near the target gene. This combination recruits downstream effector proteins, leading to the formation of less compact chromatin structures. This relaxed chromatin state allows transcriptional ma-

chinery to bind at the promoter regions of genes (170; 171). These downstream effector proteins are sensitive to the metabolic state of the cell (172; 173). The histone modifications are specific towards cell types and histone modification ChIP-seq profiles can be utilized to infer particular cell types (174).

Histone marks are dynamic, with the post-translational landscape on histones changing according to the cell's biological context (175; 176). Eraser and writer enzymes modify histones, allowing for transient histone marks that align with cell maturation events and biological functions (177; 178).

The ChIP-seq protocol has long been the gold standard for studying histone modifications. However, recent advancements in high-throughput techniques, such as Cleavage Under Targets and Release Using Nuclease (CUT&RUN), now enable the analysis of histone modifications at single-cell resolution, including the co-occupancy of histone modifications (110; 179). For further details on biotechnological approaches to study histone modifications, the review article by Chen et al. (180).

Each type of hallmark post-translational histone modification has a specific effect on downstream processes important for the transcriptional regulation of genes. Some of these post-translational modifications include:

- **Histone H3 Lys 4 trimethylation (H3K4me3)**

  H3K4me3 is a hallmark modification found at active promoters and enhancers (181). Experiments over the years have revealed that while H3K4me3 is not a critical regulator of transcription, it might be involved in overall chromatin remodeling, with the resultant transcription from H3K4me3 being context-dependent (181; 182; 183). Perturbation of the enzyme complex that adds methyl groups to H3K4 (Complex of Proteins Associated with Set1, COMPASS) in mice embryonic stem cells affected gene expression but not cell differentiation. This indicates that H3K4 trimethylation is not necessary for transcription activation but is required for optimal gene expression (184; 185). The H3K4me mark in sperm cells plays a crucial role in setting up normal gene expression patterns and developmental potential in the resulting embryo (186; 187). Multiple studies indicate a strongly conserved negative relationship between H3K4me3 and DNA methylation (188; 189).

- **Histone H3 Lys 36 methylation (H3K36me)**

  H3K36me3 is associated with actively transcribed regions because the methyltransferase SETD2 is recruited by elongation RNA polymerase II (190). While H3K36me3 is not essential for transcriptional elongation, it helps prevent spurious transcription events by promoting deacetylation and DNA methylation (191). Additionally, H3K36me1 and H3K36me2 are enriched across various genes and are linked with DNA methylation (192; 193).

- **Histone H3 Tyr 41 phosphorylation (H3Y41ph)**

  Phosphorylation of histone H3 at the 41st tryptophan residue is similar to acetylation in that it reduces the basic charge on the histone, allowing the chromatin conformation to become more accessible to the transcriptional machinery. This mark is observed across a subset of actively transcribing genes (194; 195). Histone phosphorylation is known to facilitate histone acetylation (196; 197). Similarly, histone H3 Serine 10 phosphorylation (H3S10ph) and histone H3 Serine 28 phosphorylation (H3S28ph) are modifications that promote an open chromatin conformation by preventing heterochromatin and polycomb group proteins from binding to the repressive histone modifications H3K9me3 and H3K27me3, respectively (198; 199).

- **Histone H3 lys 27 trimethylation (H3K27me3) and Histone H2 Alanine ubiquitylation (H2Aub)**

  Both of these marks are associated with transcriptional repression and are produced by the polycomb repressive complexes PRC2 and PRC1, respectively (200). Although these marks are linked to transcriptional repression, their precise role remains unclear (201). There are evidences of non-linear relationships between transcription of genes and H3K27me3, it is noticed that there is enrichment of H3K27me at the target genes of polycomb complex after the inhibition of the transcription, indicating that this mark does not always has to precede the transciptional repression (202).

- **Histone H3 lysine trimethylation (H3K9me3)**

  H3K9me3 is one of the classical markers of compact chromosomal conformation, enriched at transcriptionally repressed genomic regions (203). It is primarily present at the telomere, centromere, repetitive sequences, as well as at silenced genes (204). Heterochromatin proteins bind to the H3K9me3 regions and recruit other histone writers for H3K56me3, H3K64me3, and H3K20m3 modifications (205; 206). Experiments in mice confirm that H3K9me3 does not have a direct role in the regulation of transcription; rather, it is primarily responsible for repressed compact chromatin state (207).

  For further information on different types of histone modifications, check the review articles by Kouzarides T (169) and Bannister et al. (208).

  Thus, histone modifications have both direct and indirect roles in the regulation of transcription processes.

1.4.2.4 Euchromatin and heterochromatin

As described in the section above, nucleosomes comprise the fundamental units of chromosome organization. Since most DNA is encased inside nucleosomes, it is physically inaccessible for DNA binding proteins or enzymes. Therefore, nucleosomes play an important role in DNA metabolism such as transcription and replication (209).

Chromatin can be categorized into heterochromatin and euchromatin based on its

physical structure. Euchromatin regions feature decondensed and loosely packed nucleosomes, facilitating accessibility to transcription machinery and related proteins. In contrast, heterochromatin regions are densely packed with nucleosomes, limiting accessibility and containing genes that are transcriptionally less active. These variable chromatin conformations are mainly due to electrostatic interactions between the positively charged lysine and arginine residues of histone proteins and the negatively charged phosphate backbone of DNA (210). The single-cell open-chromatin profiles generated by ATAC-seq can be utilized to predict the chromatin interactions (211).

Chromatin structure undergoes dynamic remodeling through the addition of chemicals to histone protein residues as well as interactions with transcription factors, as discussed in previous sections (212). The review article by Grewal and Moazed discusses the role of histone modification enzymes, structural proteins, and non-coding RNAs in the formation and regulation of heterochromatin regions (213).

The chromatin interaction has wide array of biological effect on the functioning of the cells. It has been shown that interactions in the chromatin regions can be utilized to infer the drug response in cancer cells (214).

Thus, the open and closed conformation of chromatin can serve as a good indicator of regions of active transcription. Genome-wide high-throughput assays using DNase I digestion enzymes serve as a good proxy for the presence of open chromatin regions with active transcription (215).

1.4.3 Gene expression and regulation

In the preceding sections, various cis- and trans-regulatory elements involved in gene expression regulation were elucidated. Here, we provide a concise overview and highlight key aspects of transcriptional regulation relevant to this thesis work.

Gene expression is a dynamic process, not simply "on" or "off", but finely regulated according to cellular needs by precise regulatory mechanisms involving various regulatory elements. Dysregulation of gene expression can lead to disease conditions (216; 217). The molecular mechanisms underlying transcription processes have been elucidated over the years, focusing extensively on the RNA polymerase II enzyme. A combination of trans-regulatory factors and cis-regulatory elements interact to modulate

chromatin structure, facilitating RNA polymerase II binding at enhancers and promoters for subsequent transcription of gene coding regions (218; 219; 220). For a comprehensive exploration of the current mechanistic understanding of cis-regulatory elements like promoters, enhancers, and their interplay with transcription factors, co-factors, and RNA polymerase enzymes, please consult the review by Kim and Wysocka (221).

The role of non-coding genes in transcriptional regulation has gained prominence in the past decade. Long non-coding RNAs (lncRNAs) influence chromatin structure by affecting chromatin folding, thereby impacting promoter-enhancer interactions (222; 223). LncRNAs interact with nuclear proteins like CTCF to modulate histone protein methylation, thereby affecting the regulation of target genes (224; 225). Certain lncRNAs are also known to interact with transcription factors to stabilize their interactions at specific motifs (226).

One critical aspect of gene transcriptional regulation is that genes involved in the same biological process are often controlled by a limited number of transcription factors. This temporal co-regulation ensures genes are synchronously regulated to achieve specific biological outcomes. For instance, Liu et al. showed that the KLF4 transcription factor regulates genes crucial for the G1/S phase transition in the cell cycle (227). This coordinated gene expression is vital for efficiently executing biological functions, ensuring the expression of all necessary genes whose products actively contribute to that function.

Numerous computational and experimental analyses have been conducted to identify regulatory networks for different phenotypes (228; 229; 230; 231). However, to date, identifying the exact combination of regulatory elements for a given biological context remains an active research area. Deciphering such context-specific regulatory networks for different phenotypes holds promise for effectively manipulating gene expression externally, potentially leading to the development of genomic medicine strategies against disease conditions (232).

1.4.4 Non-coding genes

According to findings from the ENCODE project, less than 1.2% of the human genome encodes RNA that is translated into proteins, whereas approximately 80% of the genome

is transcribed into non-coding RNA (ncRNA) (2).

ncRNAs have garnered significant research interest due to their diverse roles in biology and pathophysiology, ranging from early development to autoimmune disorders (233; 234). Moreover, most disease-linked mutations are found in the non-coding regions of the human genome (235).

Non-coding RNAs (ncRNAs) can be broadly categorized into housekeeping ncRNAs and regulatory ncRNAs, with a focus on regulatory ncRNAs being more pertinent to this thesis work. These regulatory ncRNAs can be further classified based on their size and functional roles:

1. MicroRNA (miRNA): Typically 21-23 nucleotides long, miRNAs are the most abundant class of small ncRNAs, playing a crucial role in gene silencing across the nucleus and cytoplasm (236).
2. Small interfering RNA (siRNA): Ranging from 20-25 nucleotides in length, siRNAs are key players in the RNA interference (RNAi) pathway, facilitating mRNA cleavage in a sequence-dependent manner (237; 238).
3. PIWI-interacting RNA (piRNA): Typically 26-32 nucleotides long, piRNAs interact with PIWI proteins to execute their functions (239).
4. Enhancer RNA (eRNA): With lengths ranging from 50 to 2000 nucleotides, eRNAs are primarily implicated in the transcriptional regulation of target protein-coding genes (66).
5. Long non-coding RNA (lncRNA): Typically spanning 150-200 nucleotides, lncRNAs are the most prevalent form of non-coding RNA. They are classified based on their transcriptional origin: (i) lincRNAs (long intergenic ncRNAs) transcribed from intergenic regions, (ii) intronic lncRNAs transcribed from gene introns, (iii) sense lncRNAs transcribed from the sense strand of protein-coding genes, often containing exons, and (iv) anti-sense lncRNAs transcribed from the anti-sense strand of coding genes (240).
6. Circular RNA (circRNA): Ranging from 100-10,000 nucleotides, circRNAs possess a circular structure and often act as miRNA sponges by containing miRNA binding sites (241).
7. Y RNA: Found in the cytoplasm, Y RNAs are distinct from nuclear RNAs in mammalian cells. They are involved in RNA stability and DNA replication (242; 243).

Despite the annotation of numerous non-coding RNAs (ncRNAs), only a fraction of them have been functionally characterized. These ncRNAs participate in diverse nuclear and cytoplasmic functions, prompting the development of various molecular

biology assays to study their interactions with mRNA, DNA, and proteins. For detailed information on such assays, the review article by Sun and Chen (244).

While these binding assays illuminate interactions between ncRNAs and other molecules, they often fail to provide direct or indirect insights into the involvement of ncRNAs in biological processes. As a result, the functional characterization of many ncRNA genes remains elusive.

1.5 Predictive biology

Over the past century, advancements in biochemical and molecular biology experimental methods have provided detailed insights into cellular biology at the molecular level. High-throughput sequencing techniques in genome biology research have generated vast amounts of genomic data, presenting a significant big data challenge in analyzing and interpreting this information (245). Various computational models utilizing different features have been developed to assign roles to DNA's fundamental units, genes, linking them to biological functions and diseases, although many genes remain uncharacterized, as discussed earlier.

One of the important approaches in predictive biology is integrating multi-omics data to derive fundamental insights into the complex mechanisms of the cells. The multi-omics data includes RNA-seq, DNA-seq, ATAC-seq, ChIP-seq etc. Over the years, researchers have utilized different RNA-seq data extensively to understand the regulation of genes (246; 247; 248; 249). Still, each approach seems to give different results, agreeing with the saying all models are wrong, but some models are useful by George Box (250). Therefore, it is necessary to approach the problem while utilizing biological data of the cells to reflect the true biology. The particular dataset must capture the ground truth of the cell in order to infer their biological mechanism for biotechnological and clinical applications. In multi-omics integrative approaches, the data derived at different layers, such as genomic and transcriptomic, must complement each other to understand how each layer interacts and influences the other layer in a given biological function, like cell division. This thesis work takes this holistic approach of integrating multiple genomic data to understand the role of genes in different biological functions and diseases. A detailed overview and previous work done in the problem of associat-

ing gene function and gene regulation has been discussed in the introduction sections of Chapters 2 and 3.

One of the important approaches in predictive biology is integrating multi-omics data to derive fundamental insights into the complex mechanisms of the cells. The multi-omics data includes RNA-seq, DNA-seq, ATAC-seq, ChIP-seq etc. Over the years, researchers have utilized different RNA-seq data extensively to understand the regulation of genes (251). Still, each approach seems to give different results, agreeing with a quote: all models are wrong, but some models are useful by George Box. Therefore, it is necessary to approach the problem while utilizing biological data of the cells to reflect the true biology. The particular dataset must capture the ground truth of the cell to infer its biological mechanism for biotechnological and clinical applications. In multi-omics integrative approaches, the data derived at different layers, such as genomic and transcriptomic, must complement each other to understand how each layer interacts and influences the other layer in a given biological function, like cell division. This thesis work takes this holistic approach of integrating multiple genomic data to understand the role of genes in different biological functions and diseases. A detailed overview and previous work on associating gene function and gene regulation has been discussed in the introduction sections of Chapters 2 and 3.

In this thesis work, explainable machine learning algorithms have been employed to learn critical gene features involved in transcriptional regulation, aiming to predict novel gene-function and gene-disease associations. The construction of features for modeling genes is a crucial initial step in machine learning classification, ensuring that data points (genes) are explained and distinguished effectively.

Predictive biology, as exemplified in this thesis, is a subdomain of computational biology that synergizes with experimental approaches, offering powerful insights into biology through predictive modeling. This integration of computational and experimental methodologies contributes to a deeper understanding of biological systems and processes.

1.6 Summary

In the preceding subsections, it is apparent that many genes remain uncharacterized regarding their roles in biological functions and disease conditions. Comprehensive annotation of all genes in the human genome is essential for a thorough understanding of biology.

The interaction between cis- and trans-regulatory elements is intricate, orchestrating gene expression. It's empirically established that genes functionally involved in the same biological processes are regulated by common subsets of regulatory elements (227; 252). This synchronization of cis- and trans-regulatory elements indicates their temporal coordination in regulating gene expression for products involved in specific biological processes (253; 254; 255). Molecular biology assays like ChIP-seq and ATAC-seq capture signals across the genome, freezing cells to trap temporally related biomolecules at specific biologically relevant regions. The hidden interdependence of these elements in regulating groups of functionally-related genes can be modeled and exploited to predict the belongingness of other genes to these groups.

Subsequent chapters of this thesis delve into how epigenomic and transcription factor binding patterns at gene promoters are utilized to predict novel associations between gene function and disease, followed by post hoc analysis of these predictions.

CHAPTER 2

Chapter 2: Epigenome and TF binding patterns are predictive of ontology-based functions of coding and non-coding genes

In this chapter, the computational framework developed to predict ontology-based gene functions is described. The reliability of the predicted results through independent validation using CRISPR screens and PubMed abstract mining is outlined. Additionally, post-hoc analysis results are presented, including clustering biological and molecular functions using their top predictors from the random forest model, and analyzing the synergy and pleiotropy of transcription factors in relation to ontological functions. Finally, the utility of the framework is demonstrated.

2.1 Challenges and approaches of associating non-coding genes to functions

Both coding and non-coding RNA (ncRNA) play a part in a cell's metabolic pathway. The main purpose of non-coding RNA (ncRNA) is to cis- or trans-regulate the coding genes, the products of which form the framework of metabolic pathways (256). It is more challenging to investigate the roles of ncRNAs experimentally due to the diverse molecular processes by which they perform their roles in a multitude of biological and molecular functions at numerous regulatory levels (257; 258). Moreover, identifying homologs for non-coding genes is challenging because their sequences are rarely conserved across species, unlike protein-coding genes (259). Therefore, the role of ncRNAs in human cells may differ from that of model organisms where their roles have been experimentally validated. Another challenge in sequence-based function prediction, commonly used by scientific groups, is the low homology and sequence conservation observed in many genes, including non-coding RNAs.

Computational analysis, which makes use of the current understanding of the links between genes and functions or diseases, offers a potential method for deciphering the roles of the genes. Gene ontologies show the links between functions, genes, and illness that have been empirically annotated. These ontologies have already been used by numerous research teams to forecast the relationships between genes and illnesses and functions (260).

To enhance clarity, ontological gene-sets encompassing molecular activities and biological processes are referred to here as "functions." Predictive models play a crucial role in identifying gene functions by discerning patterns between known and unknown gene properties, leveraging their effectiveness in data analysis. Still, it's critical to train a prediction model using the most pertinent biological signals. To reliably predict gene functions, it is crucial that the characteristics of functionally related genes accurately reflect their functional classes. One effective strategy involves comparing the amino acid and nucleotide sequences of genes and proteins with known functions to those of genes whose functions are not yet determined (261; 262; 260; 263). However, alternative isoforms have been shown to diverge in function (264). Because non-coding genes lack established biological roles, relying solely on primary sequence comparison for them would offer limited utility. In order to find ncRNA genes associated with diseases, some researchers have employed the ontological linkages between genes (265). However, there are fewer ncRNA gene annotations in the ontologies, which would lead to poorer coverage. Few research have employed features derived from gene expression data to infer the roles of non-coding genes from the co-expression of coding genes (266). Liao et al. developed a co-expression network involving both coding and non-coding genes to identify the functions of long non-coding RNAs (lncRNAs). (267). However, genes that share a function might not always show correlated expression, while many functionally unrelated genes can exhibit co-expression at certain time points (268). Hence, current methods do not effectively leverage genomic features to predict the functions of non-coding genes.

Non-coding RNAs (ncRNAs) are known to regulate the transcription of genes participating in the same biological processes by interacting with chromatin, RNA, and proteins (269). Numerous computational approaches have also been suggested for predicting the functions of non-coding genes, employing diverse combinations of features. Zhang et al. introduced BiRWLGO, a bi-random walk model, for predicting the func-

tions of long non-coding RNAs (lncRNAs) by leveraging protein-protein interaction networks and lncRNA-protein interaction (270). PLAIDOH, another computational approach, integrates transcriptome data with enhancer landscape, subcellular localization, genome architecture, RNA-binding profiles, and chromatin interaction. It calculates statistically determined scores for each lncRNA, functionally linking them with coding genes across various cancer conditions (271). However, the investigation of epigenomic patterns and transcription factor (TF) binding at promoters for predicting non-coding RNA (ncRNA) function remains limited. Epigenomic features, TF binding, and cofactor interactions are regulatory elements found at both coding and non-coding genes, playing pivotal roles in gene expression regulation (272; 273). Epigenetic marks and chromatin structure work in concert with TFs to modulate gene expression dynamics (274). Previously, epigenome profiles have been utilized for predicting gene expression (275) and linking diseases with single nucleotide polymorphisms (SNPs) (276). Meanwhile, various methods and studies leveraging TF ChIP-seq profiles have aimed to correlate TF binding patterns with specific gene functions (277; 278; 279). Integrating TFs as features can offer insights into the synergistic and cooperative regulation involved in diverse functions (280). Nevertheless, an extensive analysis of the combinatorial binding patterns of multiple transcription factors (TFs) at promoters and their associations with gene function remains largely unexplored.

In this study, we developed a new method for predicting gene ontology functions by integrating signals from TF-binding patterns, epigenomic profiles, and CAGE-tag distributions at gene promoters. To capture comprehensive transcriptional regulatory signatures, we utilized a diverse array of publicly available datasets including ChIP-seq data for histone modifications, transcription factors (TFs), and DNase I hypersensitivity sites, along with cap analysis of gene expression (CAGE) tags to account for both polyadenylated and non-polyadenylated gene expression. To validate the robustness of our approach, we conducted downstream analyses that emphasized the most predictive profiles of TF and cofactor binding. These analyses included clustering functions and associating genes with these functional clusters. Furthermore, we investigated the specificity of basic TF combinations (e.g., TF pairs) in terms of their functional roles.

2.2 Materials and methods

2.2.1 Epigenome and TF-binding features score calculation for promoters

In this study, each gene ontology was approached as a distinct class, where genes with empirical annotations served as positive labels, shaping the problem of gene function prediction into a classification task. The method utilized read-count data from epigenomic and transcriptomic binding assays (ChIP-seq, DNase-seq, CAGE-tags) as the primary features. All TF, cofactor, and epigenome ChIP-seq profiles, along with CAGE-tags, were obtained from diverse human tissues and cells, with reads mapped to the hg19 genome version. To assess binding activities at promoters, the quantification of the number of DNA fragments (reads) within a 1 kbp region around gene transcription start sites (TSS) was performed.

i) ChIP-seq and open-chromatin profile

From the ChIP-atlas database, ChIP-seq profiles of transcription factor bindings and histone modifications in bigWig format, across various human cell types and tissues were downloaded. Similarly, open-chromatin profiles, predominantly DNAse-seq data in bigWig format, were obtained from the same database (281). Using the bigWigToBedgraph tool, the bigWig files were converted to bedGraph format (282). The bedGraph files contain scaled read-count scores for genomic bins. To ensure uniformity, the read-counts in each bin were normalized by the mean read-count across all genome-wide bins. The normalized read-count scores from genomic bins within 1 kbp of transcription start sites (TSS) were summed for each ChIP-seq profile to obtain the read-count score at the TSS for every gene. This procedure was performed individually for each ChIP-seq profile.

ii) CAGE-tag profile

CAGE-tag profiles for multiple human cell types were downloaded in .bam file format from the FANTOM (RIKEN, Japan) database (283). The .bam files were converted into BedGraph format, generating read-count scores in 200 bp bins across the genome. To ensure consistency, read counts were standardized by normalizing the values in each bin using the mean read count. Scores of bins located within 1 kbp of the transcription

start site (TSS) in each CAGE-tag profile were combined to calculate the read-count score at the TSS for each gene.

Transcription start sites (TSS) of non-coding genes were obtained from Gencode (V30) and RefSeq gene transcripts (284; 3). Multiple transcripts for each gene were included, ensuring their TSS locations were separated by at least 500 base pairs. In total, 89,747 promoter sites were analyzed.

2.2.2 Prediction method

For each gene set within the gene ontology, annotated genes were designated as positive data points, and an equal number of non-annotated genes were randomly selected as negative data points, framing gene function prediction as a classification problem. With a pool of 50,000 potential genes and fewer than 100 expected unknown positive genes for a function, the background probability of a false negative for any randomly selected gene was calculated to be less than 0.002 (100/50,000).

For each gene set, negatives and positives were balanced. For instance, if a function included 50 known positive genes in the training set and 50 randomly chosen genes in the negative set, the background false negative probability of 0.002 resulted in a likelihood of including one or more false negatives (positives) among the 50 randomly selected negative genes being less than 0.005 (as determined by the Binomial test). Based on these calculations, randomly selected genes not linked to the gene set were used as the negative set for that specific function.

The positive and negative examples were split into a training set (75%) and a test set (25%). Five distinct machine learning models were applied for each gene set: Random Forest, XGBoost, Support Vector Machine (SVM), Lasso regression, and Ridge regression (L2-regularized logistic regression).

Logistic regression with L2-regularization (ridge regression) was implemented using ‘cv.glmnet‘ with ‘alpha = 0‘ and ‘family = "binomial"‘ from the ‘glmnet‘ R package. The Random Forest algorithm was applied using the ‘randomForest‘ function from the ‘randomForest‘ R package. SVM models were implemented using the ‘svm‘ function from the ‘e1071‘ R package. Lasso regression was performed using ‘cv.glmnet‘ with ‘alpha = 1‘ from the ‘glmnet‘ R package. The XGBoost algorithm was executed using

the ‘xgb.train‘ function from the ‘xgboost‘ R package.

The following types of machine learning algorithms were utilized to predict the belongingness of the genes to ontological gene sets:

1. **Linear regression:** This algorithm serves as a foundation approach to get the preliminary idea of the data and to fit more complex models. It is computationally inexpensive and easy to implement.
2. **Logistic regression:** This algorithm also serves as a foundation to fit more complex models. It differs from linear regression in that it utilizes L2/Ridge regularization, avoiding over-fitting of the model.
3. **Random forest:** It is one of the most efficient and powerful algorithms because it uses multiple decision trees to arrive at a consensus outcome that enables the model to overcome noise in the data. Another advantage of using a random forest algorithm is that it provides feature importance scores that can be utilized to derive novel insights into the biology of the genes and their regulators (features).
4. **Support Vector Machine (SVM):** The SVM model has strong generalizable capabilities; it can perform well in predicting data points even if they fall away from the trained distribution. SVM can work even when the number of training data points is less.
5. **XGBoost:** XGBoost is one of the best-performing models, outperforming other algorithms on numerous benchmark tests. It is fast, scalable, and has L1 and L2 regularization to avoid over-fitting. It also gives feature importance scores.

Evaluate using these metrics: accuracy, balanced accuracy, F1-score, error rate and Mathews correlation coefficient (MCC (Supplementary File 2).

After evaluating the test set, the trained model was employed to predict associations between gene functions and promoters within our dataset, aiming to discover new connections. To ensure the robust identification of novel gene-function links, we computed a confidence score for each function (gene-set).

2.2.3 Calculating confidence score for gene-sets

For robust predictions, we computed the confidence score for each function.

For each gene-set, the trained random forest algorithms were utilized to predict probability scores indicating gene-set membership across all 89,747 promoters. The confidence score of a gene-set is determined by its maximum precision, calculated as the highest ratio of true positive genes to the total predicted genes (false positive +

true positive), achieved by varying the threshold for classifying genes as negative or positive based on predicted probabilities. Gene-sets with a confidence score exceeding 60% were prioritized for downstream analysis.

Evaluation metrics such as sensitivity, specificity, balanced accuracy, and accuracy were computed using balanced datasets (equal numbers of negatives and positives) in the training and test sets. However, probabilities of the positives and all non-positives (unbalanced datasets) were utilized to estimate the confidence score.

2.2.4 Other methods

Balanced accuracy calculation

To evaluate the performance of a binary classifier balanced accuracy is a metric used, especially in situations where there is an unequal distribution between positive and negative instances (imbalanced classes). It is calculated as the average of sensitivity and specificity:

$$\text{Balanced Accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2} \tag{2.1}$$

Method for five-fold cross-validation method

The R package "sperrorest" was utilized to implement the five-fold cross-validation (285).

Method to make inferences about top regulators

We identified key regulators by analyzing feature importance during the training of random forest algorithms. This approach is akin to the methodology employed by GENIE3, which excelled in gene-network inference during the DREAM 5 challenge (286; 287; 288). Instead of relying solely on transcription factor gene expression, binding affinity to promoters was used as feature scores to predict gene associations with specific classes. Accordingly, for each function, the top 20 predictors with significant feature importance were identified by the random forest method.

Method for clustering functions

To delineate clusters of gene functions (gene-sets), initially computed similarity metrics between functions were based on assessing the presence of transcription factor (TF) and cofactor ChIP-seq profiles (SRX IDs) among their respective top 20 predictors. Matching directionality was considered, penalizing cases where common TFs and cofactors (within the top 20) exhibited opposite directions. Subsequently, these similarity scores were inverted and converted into distances to facilitate dimension reduction using tSNE. Thus, the distance between gene-sets A and B was defined as:

$$d(A, B) = 10 - \text{closeness}(A, B) \tag{2.2}$$

Where,

$$\text{closeness}(A, B) = \sum_{\text{TF}_i \in \text{top 20}(A \text{ and } B)} \text{TF}_n \cdot \text{sign}(\text{cor}(\text{TF}_i, A)) \cdot \text{sign}(\text{cor}(\text{TF}_i, B)) \tag{2.3}$$

Where,

TFi represents a transcription factor (TF) that is part of the common TFs (TFn) identified within the top 20 predictors by the random forest algorithm.The function cor(TFi, A) measures the correlation between the read-count score of TFi at promoters and the genes' association with function A. A positive cor(TFi, A) suggests that TFi is likely enriched at the promoters of genes associated with function A. Thus, the directionality of the relationship between function and the presence of top predictive TFs is utilized to calculate a stringent proximity score. Additionally, the cohesion index for each cluster was computed as the average distance among its individual members (Supplementary File 3).

Dimensionality reduction was conducted using the distance matrix, followed by density-based clustering. The 'Rtsne' R package was employed with the 'is_distance' option set to TRUE for generating the low-dimensional embedding. Subsequently, DBSCAN was applied to identify clusters of functions based on the 2D embedding coordinates produced by 'Rtnse'.

Method for PubMed abstract mining as independent validations

To validate new predictions and benchmark the methodology against alternative approaches, validation based on PubMed abstracts was conducted. In this process,

the ontology term and its associated predicted gene term were utilized as inputs. The 'Bio.Entrez' package was employed to query the PubMed database for instances where the ontology term and its corresponding predicted gene term co-occurred within the abstracts of research articles. For negative control for this methodology, gene terms were randomly paired with ontology function terms, and their co-occurrence was investigated under the same parameters.

To ensure effective matching of the ontology term within potentially relevant abstracts, stop words were removed from the ontology terms during processing.

Stop words 1: 'small', 'type', 'the', 'or', 'from', 'he', 'into', 'on', 'like', 'cell', 'is', 'groups', 'to', 'layer', 'in', 'ii', 'by', 'of'.

Stop words 2: 'response', 'environment', 'pathway', 'cellular', 'species', 'protein', 'coupled', 'activity', 'mrna', 'negative', 'gene', 'process', 'interaction', 'regulation', 'formation', 'particle', 'receptor', 'sensory', 'acid', 'positive', 'complex', 'maintenance', 'dependent', 'movement', 'chemical', 'involved', 'signaling', 'binding', 'group', 'modified', 'other', 'organism', 'nucleotide', 'compound', 'left', 'right'.

The table of novel predicted associations between genes and functions used for PubMed abstract mining to generate Figure 2.5A, is available by clicking this link.

Method for comparison of predicted results of different methods using PubMed abstract mining

Fifty genes were chosen from a selection of 20 gene-sets randomly picked from a pool of 1423 gene-sets, each with a functional confidence score exceeding 60%. These genes were subsequently analyzed using four prediction tools: NetGO 2.0, DeepGO, Correlation AnalyzeR, and GenetICA-Network (289; 290; 291; 292). Using the 'correlationAnalyzeR' R package, the 'analyzeSingleGenes' function was employed to predict ontology-based functions for the list of 50 genes. The highest-scoring label from these predictions was selected as the final prediction. Genes without a generated prediction were left unlabeled.

GenetICA-Network, NetGO 2.0, and DeepGO web servers were used to predict functions for the list of 50 genes by inputting the corresponding protein sequence FASTA files. The top isoform from the UniProt database was selected (293). The final label chosen was the highest-scoring one with specific terms from either molec-

ular functions or biological processes. Genes with prediction scores below 50% or non-coding genes were excluded from labeling.

The table of gene names and corresponding predicted ontology terms by different prediction methods used for PubMed abstract mining comparison to generate Figure 2.5B is available by clicking here.

Methodology for transcription factors synergy and pleiotropy analysis

The occurrence of transcription factors (TFs) and cofactors among the top 20 predictors was counted across various and identical cell types, concentrating on functions with a confidence score exceeding 60%. Furthermore, a list of TF pairs was compiled using the set of TFs as features, and the frequency of each TF pair among the top 20 predictors across all functions was recorded.

Methodology for evaluation using CRISPR screens

Validation of genes involved in the "cell cycle process" and "immune system" clusters was carried out using CRISPR screens, evaluating gene function via viability (emphasizing cell cycle and DNA repair-related genes) and phagocytosis (294; 295). Genes with p-values greater than 0.05 were excluded from the analysis. The z-scores of the predicted genes within these clusters were compared to those of randomly selected genes. This validation approach was also applied to additional control CRISPR datasets, as illustrated in Figure 2.5C-D.

Demonstrating the flexibility of GFPredict, the "predict_related_genes" function utilized random forest machine learning models to train on the top 50 genes from diverse CRISPR screens (227; 296; 297; 298; 299). The n_bootstrap parameter was varied from 3 to 20 to optimize the models, enhancing negative point sampling and aiming for improved balanced accuracy. GFPredict was used to train the model, after which the top 30 predicted genes were selected. Their CRISPR scores were then evaluated and compared to those of randomly chosen genes.

To verify the functional predictions of non-coding genes, lncRNA CRISPR screens were employed. Specifically, the non-coding genes predicted within cluster-47 (associated with cell cycle functions) were overlapped with the genes from CRISPR screens, and their scores were compared to those of randomly selected genes (300). The validation process utilized two R scripts: "lncrna_crispr_validation.R" and "package_test_crispr.R",

accessible via the GitHub link to the R package GFPredict.

The scoring method for lncRNA genes, as outlined by Liu et al., was computed using the following formula: screen score = scaled (-log10(adjusted P)) + absolute value of scaled (log2(sgRNA fold change)) (300).

2.2.5 Availability of data and code

Transcription factor, histone mark, and DNase-seq profiles in bedGraph format were sourced from the ChIP-Atlas database and processed using the DFilter tool. Additionally, CAGE-tags profiles were retrieved from the FANTOM database. The read counts for these epigenome profiles are accessible via DFilter, which can be accessed here.

Our tool, GFpredict, enables the prediction of genes that are functionally related to a list of genes provided by the user, which is biologically relevant. This tool is available as an R package called 'GFPredict'.

The code and documentation for GFpredict can be accessed at https://github.com/reggenlab/GFPredict.

Additionally, predictions made using GFpredict can be accessed at http://reggen.iiitd.edu.in:1207/gfpredict_server_script/

2.3 Results

This approach was developed on the hypothesis that the coordinated expression of functionally related genes is regulated by a distinct set of factors, including both coding and non-coding genes. To execute this, we obtained ChIP-seq profiles for transcription factors (TFs), histone modification marks, DNase-seq data, and CAGE-tags from multiple repositories. We then quantified their read counts within a one Kbp vicinity of gene transcription start sites (TSS), thereby creating read counts across a two Kbp-wide promoter region. The workflow outlining our methodology, named GFPredict, is illustrated in Figure 2.1.

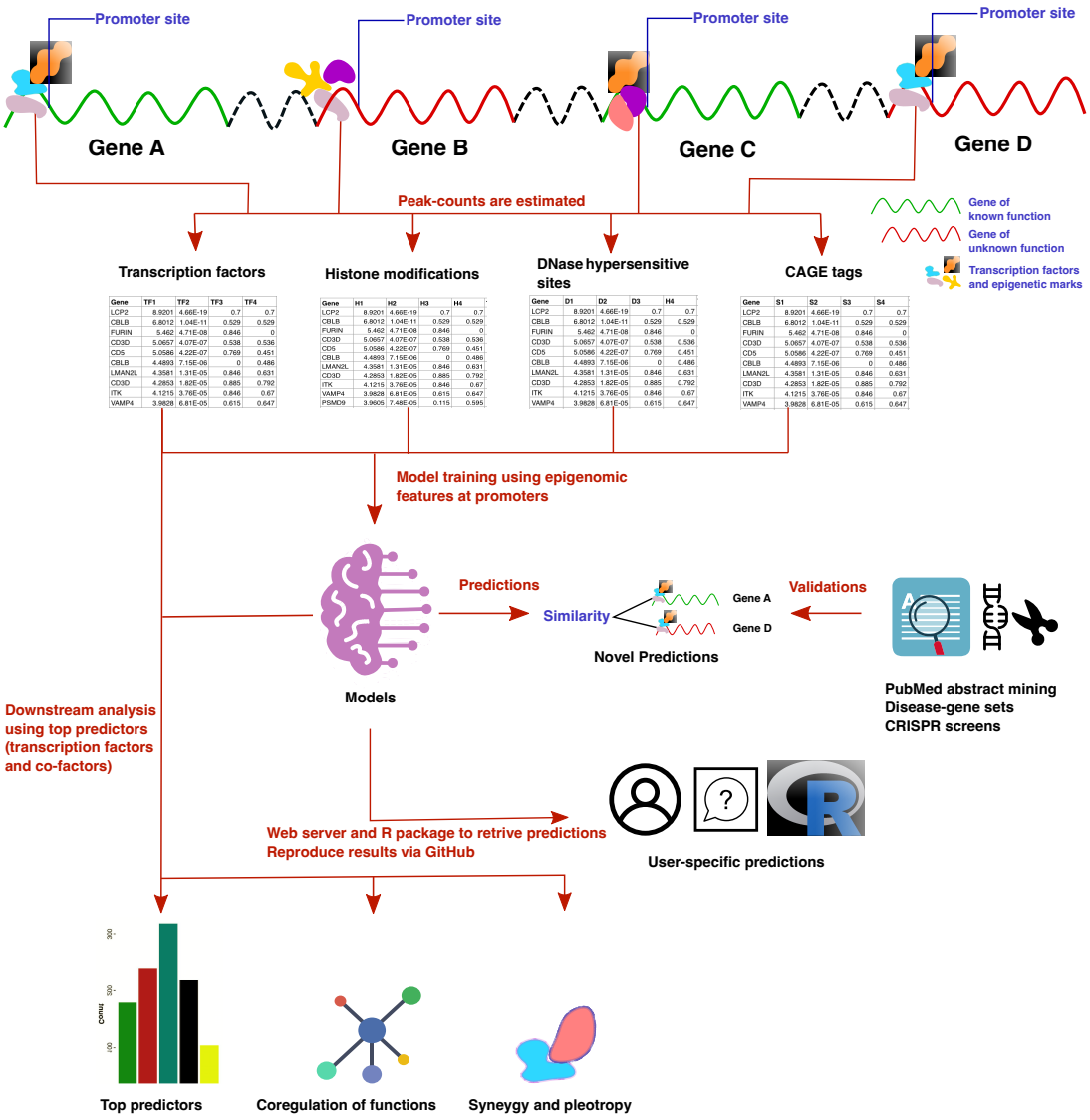

Figure 2.1: This flowchart depicts the workflow of the study.
[Figure source: Chandra, Omkar, et al. "Patterns of transcription factor binding and epigenome at promoters allow interpretable predictability of multiple functions of non-coding and coding genes." "Computational and Structural Biotechnology Journal" 21 (2023): 3590-3603]

2.3.1 Epigenome and TF binding patterns at promoters are predictable of ontology-based functions of genes

Machine learning algorithms were utilized to train models for each biological function within the ontologies. Five different ML models were employed, incorporating TF binding patterns along with ChIP-seq data for cofactors, DNase hypersensitivity, histone modifications, profiles, and CAGE tags. The MSigDB database provided the

training dataset, which consisted of 9559 function gene-sets (301). Two approaches were employed for predictive modeling. In the first approach, ChIP-seq profiles of TFs and cofactors (n=823), along with ChIP-seq data for DNase-seq (n=255), histone modifications (n=621), and CAGE tags (n=255), were utilized from a total of 1954 non-diseased samples. This method yielded highly accurate predictions for numerous functions. For example, using random forest, sensitivity exceeded 80% and specificity reached at least 90% for 425 gene-sets. The remaining four ML models (logistic regression with L2-regularization (ridge), Lasso-based linear regression, XGBoost, and SVM) showed sensitivities of 80% and specificities of 90% for 100-300 gene-sets (Figure 2.2A). Furthermore, it was observed that the AUROC (area under the receiver operating characteristics curve) surpassed 0.9 (considered excellent) for 555 gene-sets when employing the random forest model with a balanced test set (Fig. 2B). Using the random forest model, 4467 functions (gene-sets) exhibited favorable AUC values (ranging between 0.8 and 0.9) on the balanced test sets (302).

In the second approach, feature scores were estimated using 823 TF and cofactor ChIP-seq profiles (encompassing 736 TFs and 87 cofactors) from normal (non-diseased) samples. Despite this adjustment, there was no significant decrease in the number of functions demonstrating similar predictability. By applying the threshold criteria of 80% sensitivity and 90% specificity, 318 functions were identified using the random forest model. Subsequently, functions with very good predictability (sensitivity > 80%, specificity > 90%) from the five ML models were combined. Using TFs and cofactors, a total of 670 functions showed very good predictability when using at least one of the five ML models.

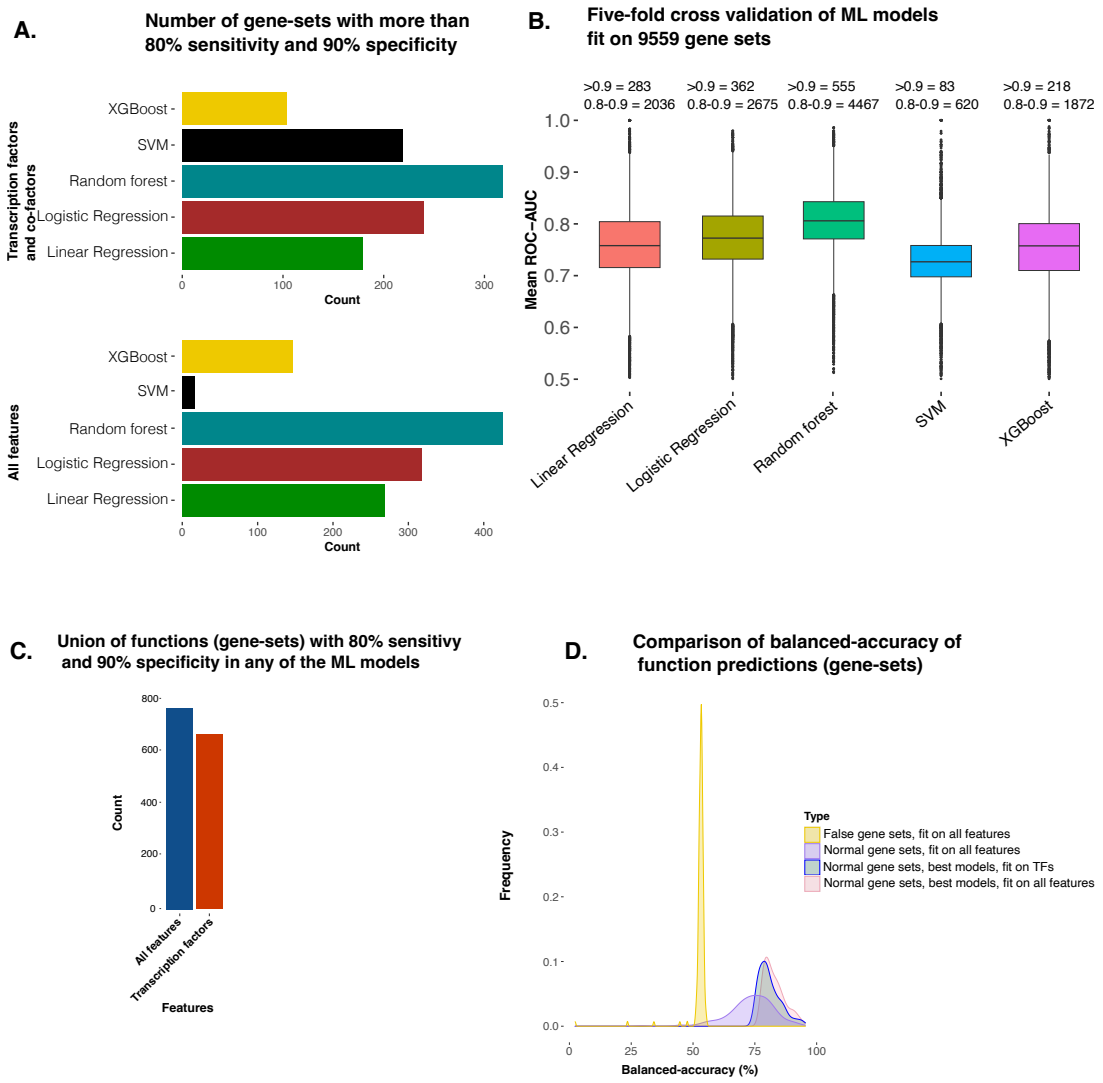

Figure 2.2: Demonstrates the robustness of epigenome profiles, specifically the predictive capacity of transcription factor binding patterns at promoters in gene function prediction. The figure includes four panels: A) Bar plot illustrating the count of gene-sets with strong predictions (80% sensitivity and 90% specificity) using five machine learning models. The upper panel highlights transcription factor ChIP-seq profiles, while the lower panel integrates five profile types. B) Box plots showing the area under the receiver operating characteristic curve (AUC-ROC) across all gene-sets, averaged over five-fold runs, with counts of gene-sets above 0.9 and between 0.8 to 0.9 indicated. C) Bar chart presenting the count of function sets with robust predictability using any of the five machine learning models. D) Plot validating the methodology, depicting the distribution of balanced accuracy achieved with false gene-sets (generated via random sampling) and experimentally annotated gene-sets, focusing on functions with balanced accuracy exceeding the 35th percentile.
[Figure source: Chandra, Omkar, et al. "Patterns of transcription factor binding and epigenome at promoters allow interpretable predictability of multiple functions of non-coding and coding genes." "Computational and Structural Biotechnology Journal" 21 (2023): 3590-3603]

Despite incorporating all features (TFs, DNase-seq, CAGE-tags, and histone modification) in the first approach, there was only a 15% increase in the number of functions (totaling 773) demonstrating excellent predictability (sensitivity > 80%, specificity > 90%) using at least one machine learning model (Figure 2.2C). However, employing a more lenient criterion of sensitivity > 70% (with specificity > 90%) in the second approach, which focused solely on transcription factor and cofactor ChIP-seq features, led to a substantial increase to over 1300 functions (Figure 2.3B). In Supplementary File 2, the evaluation metrics for the machine learning model fitted on each gene-set are provided.

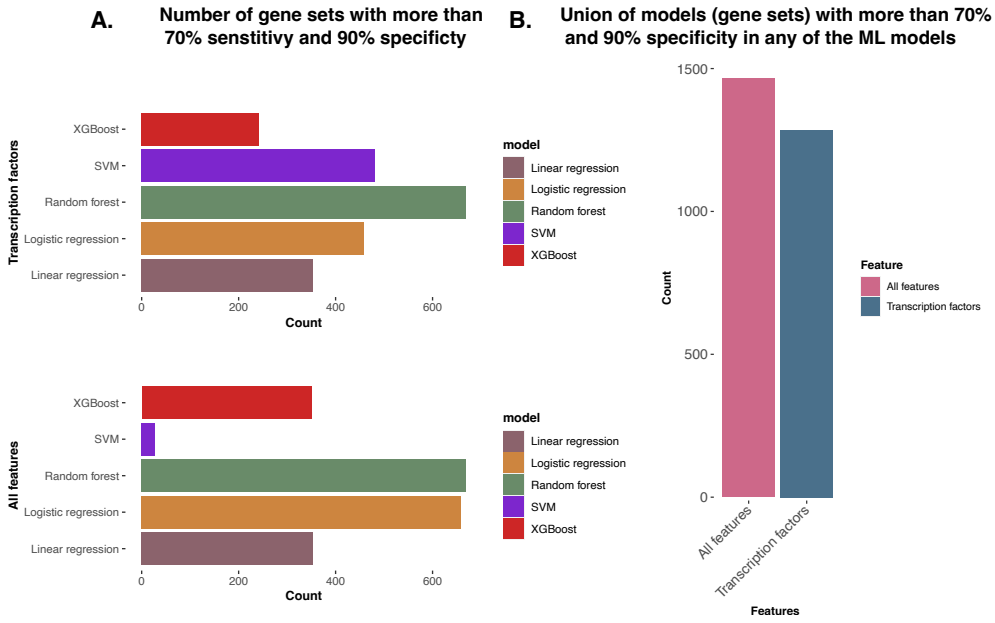

Figure 2.3: A) The count of functions (gene-sets) meeting the satisfactory prediction criteria (specificity 90%, sensitivity 70%). B) The count within the union set of gene-sets meeting the satisfactory prediction criteria across different machine learning (ML) methods.
[Figure source: Chandra, Omkar, et al. "Patterns of transcription factor binding and epigenome at promoters allow interpretable predictability of multiple functions of non-coding and coding genes." "Computational and Structural Biotechnology Journal" 21 (2023): 3590-3603]

2.3.2 Non-random nature and relevance of high predictability

To confirm that the high predictability observed using this method is not incidental, a null model was established for comparison. Genes from existing gene-sets were randomly shuffled to create 200 "false gene-sets," and random forest models were trained

on these sets. Results revealed that the best-performing model achieved an average balanced accuracy of less than 55%. In contrast, models trained on empirically annotated gene-sets achieved an average balanced accuracy of 75%, as depicted in Figure 2.2D. This underscores that robust predictability is specific to biologically relevant gene-sets, indicating a discernible regulatory pattern involving common regulators at gene promoter sites linked to similar biological functions.

2.3.3 Inference from clustering of functions

Furthermore, a direct investigation was conducted to explore the coregulation of functions resulting from the combined interactions of TFs, aiming to understand the principal functional groups encompassing both coding and non-coding RNA. Based on shared top predictive TFs and cofactors, 1423 gene-sets were clustered with a confidence score above 60%, following the procedure outlined in Methods 2.2.4.

Fifty distinct clusters of functions (Supplementary File 3) were identified based on shared top predictors, as depicted in Figure 2.4. Subsequent analysis revealed that many clusters were enriched with functions associated with similar major cellular activities (Supplementary File 3). Each cluster was manually curated and labeled with a term representing a major cellular process. For instance, one significant cluster (cluster-47) is associated with the cell cycle process and includes functions like 'microtubule-organizing center', 'regulation of cell cycle process', 'cytokinesis', 'nucleolus', and 'regulation of cellular protein localization' (Figure 2.4). Notably, the primary transcription factors and cofactors predicting functions within cluster-47 include XRN2, BRD4, SMARCA4, CTCF, and PARP1 (Supplementary File 4). Earlier research has identified MYC, CTCF, PARP-1, and SMARCA4 as key regulators of the cell cycle (303; 304; 305; 306). Another cluster (cluster-26), depicted in Figure 2.4, was notably enriched with terms associated with early development and morphogenesis. Noteworthy among the top predictive factors shared within cluster-26 were POU5F1 (307), RNF2, and SMARCB1 (Supplementary File 4) (308; 309), along with SIX1 (309), all recognized for their roles in regulating early developmental genes.

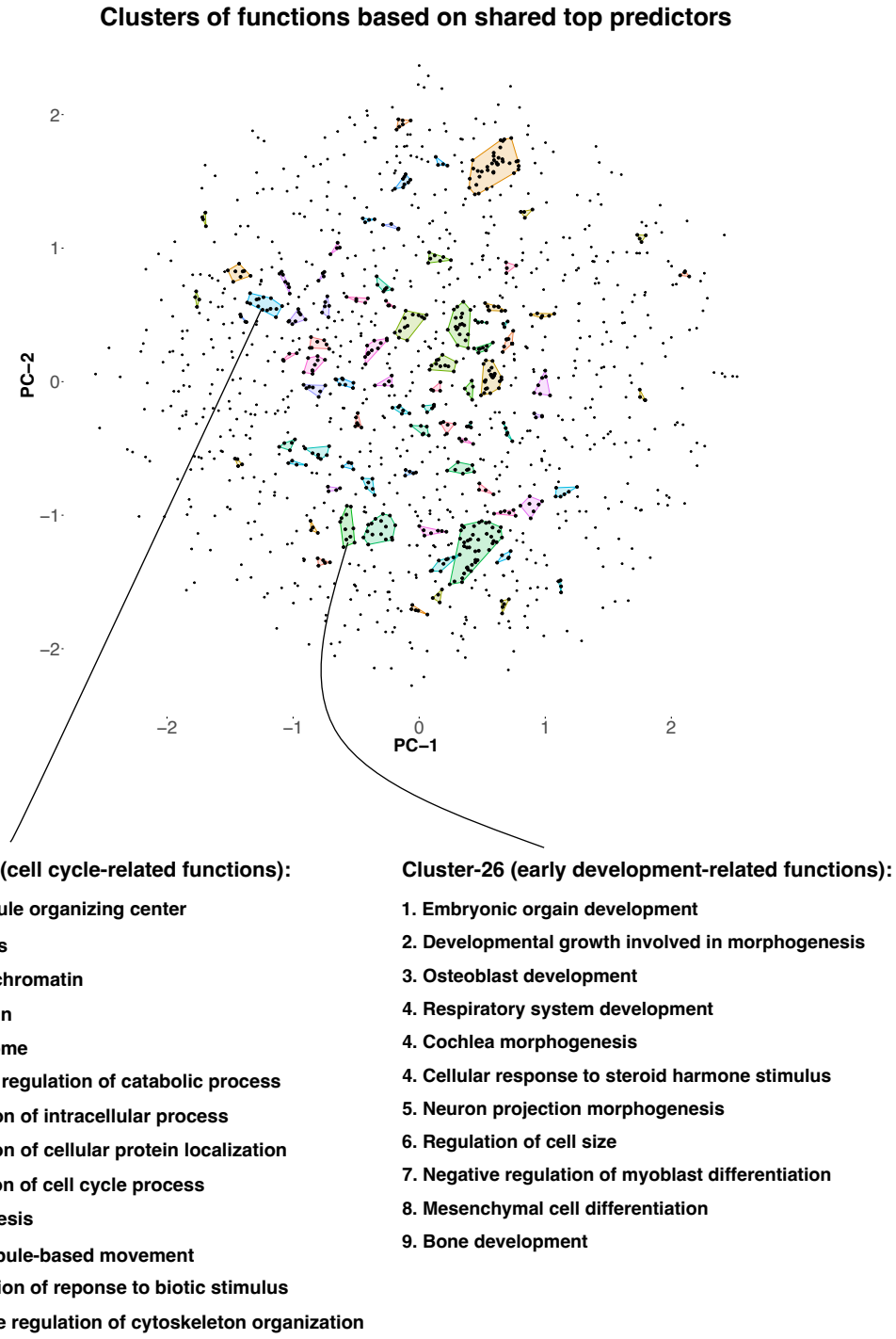

Figure 2.4: This figure depicts the clustering of functions based on their shared predictive TFs and cofactors derived from ChIP-seq profiles, illustrating potential overlaps in significant cellular processes. The tSNE plot and DBSCAN-based clustering visualization represent each gene-set as a point. The heatmap illustrates the similarity in the count of shared top predictors between two clusters: cluster-47, associated with cell cycle functions, and cluster-26, related to early developmental processes. The members of cluster-47 and cluster-26 are listed beneath the cluster plot.
[Figure source: Chandra, Omkar, et al. "Patterns of transcription factor binding and epigenome at promoters allow interpretable predictability of multiple functions of non-coding and coding genes." "Computational and Structural Biotechnology Journal" 21 (2023): 3590-3603]

Within certain clusters, members may encompass diverse ontological function terms that do not necessarily contribute prominently to the overarching major cellular process. For example, in cluster-47, most members are associated with the labeled cellular process of cell cycle activity, whereas terms like 'negative regulation of catabolic process' may appear atypical. A detailed examination reveals that during the cell cycle, there is increased metabolic activity in biosynthetic pathways that produce large molecules like DNA and structural components crucial for cell proliferation. Simultaneously, there is a decrease in degradative processes that involve breaking down protein complexes. Similar indirect roles of cluster members in major cellular processes, such as reproduction (cluster-44) and the immune system (cluster-7), are also observed in other clusters (Supplementary File 3). Therefore, the formation of these functional clusters expands our ability to associate gene-sets with important cellular processes and offers a chance to investigate the specificity of binding patterns among regulators (TFs and cofactors) in the context of systems biology.

2.3.4 Independent validations and comparison with other methods

The predicted results can be reliable only if they can be validated independently, for this purpose we utilized PubMed abstract mining, known disease-gene association database, along with CRISPR perturbation screens generated in different phenotypic contexts.

2.3.4.1 PubMed abstract mining of co-occurrence of gene names and function term

Abstracts of PubMed articles published between 1990 and 2021 were examined to assess the biological significance of our predictions by analyzing the co-occurrence of predicted gene terms and their corresponding biological function terms from the ontology. The boxplot shown in Figure 2.5A compares the total co-occurrence of predicted gene term-function term pairs with randomly paired gene-function terms used as controls. This analysis reinforces our confidence in the accuracy of our predictions. Additionally, we investigated whether the predicted genes were associated with known human disease gene-sets. Our analysis identified 23 instances where the predicted genes overlapped with disease gene-sets obtained from the DisGeNET database (18) (Table

2.1).

Moreover, a total of 15 overlaps were identified between the predicted genes and mouse disease-phenotype gene-sets available through Mouse Genome Informatics (MGD) (310) (Table 2.2).

The methodology involved initially matching function terms from the ontology with disease-phenotype terms to which novel predictions were made. Subsequently, the predicted gene terms were intersected with annotated genes. If there was a convergence between the function and gene terms with disease phenotypes, the prediction was considered validated.

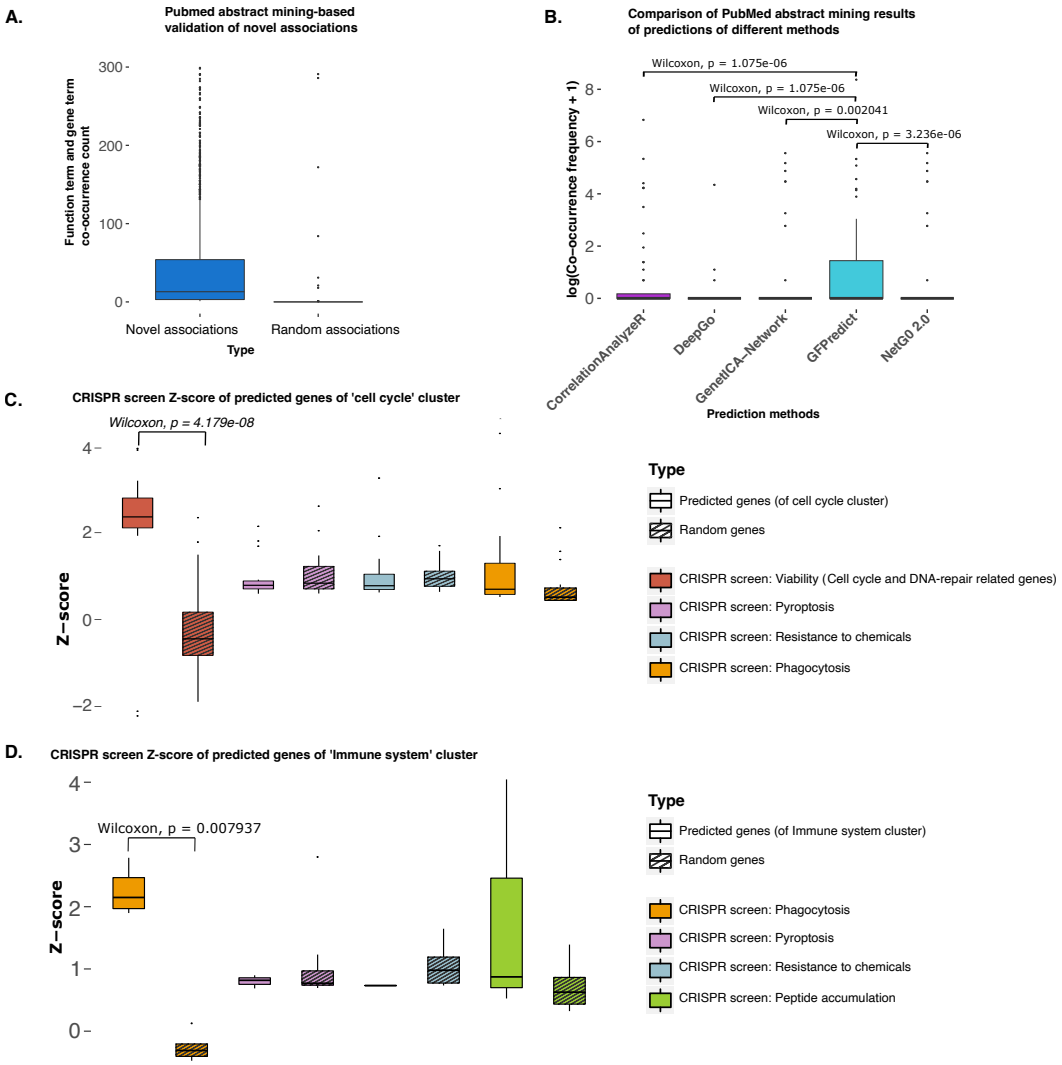

Figure 2.5: (Caption next page.)

Figure 2.5: Validation of novel predictions for function-gene associations: A) The box plot shows the frequency of co-occurrence between gene names and function terms in PubMed abstracts. On the left, it depicts the co-occurrence rates for predicted gene-function associations identified by GFPredict, while the right side illustrates the co-occurrence rates for randomly paired gene-function associations. Neither the novel predictions nor the random associations were included in the gene sets used for training or testing. B) A comparative evaluation of five methods for identifying associations between genes and functions. For the "Viability" cluster, GFPredict-predicted genes predominantly associated with DNA repair and cell cycle processes. Striped bars represent random gene scores, while solid bars represent predicted gene scores. No significant differences were observed for the clusters "chemical resistance," "pyroptosis," and "phagocytosis," which is crucial in immune responses. Similarly, for the "immune system" cluster, no significant differences were found between predicted and random associations in processes like "chemical resistance," "pyroptosis," and "peptide accumulation."
[Figure source: Chandra, Omkar, et al. "Patterns of transcription factor binding and epigenome at promoters allow interpretable predictability of multiple functions of non-coding and coding genes." "Computational and Structural Biotechnology Journal" 21 (2023): 3590-3603]

Table 2.1: Intersection of predicted genes with human disease-gene-sets.

Disease term	**Ontology term**	**Predicted Genes**
leukocyte disorders	leukocyte mediated immunity	ITGB2
organic mental disorders substance induced	organic catabolic	F12
peripheral nervous system diseases	nervous	AR
cardiac hypertrophy	cardiac muscle apoptotic	NPR
male sterility due chromosome deletions	male gamete generation	RBMY1A1
global developmental delay	developmental maturation	NTNG2
metabolic myopathy	nucleoside monophosphate metabolic	ACADVL
disorder eye	eye morphogenesis	CDH3
brain diseases metabolic inherited	dna metabolic	NDUFAF2
chromosome 16p112 deletion syndrome kb	nuclear chromosome	SH2B1
chromosomal instability	chromosomal region	KIF11

Continued on next page

Table 2.1 – continued from previous page

Disease term	Ontology term	Predicted Genes
acute myeloid leukemia m1	myeloid leukocyte mediated immunity	CAPG
acute coronary syndrome	acute inflammatory	PON1
malignant lymphoma lymphocytic intermediate differentiation diffuse	lymphocyte differentiation	PIK3CD
neural tube defects	tube size	PYY
metabolic bone disorder	steroid metabolic	C2
lymphoma extranodal nk t cells	T cells	JAK3
cardiac arrhythmia	cardiac muscle tissue	KCNJ2
hodgkin lymphoma lymphocyte depletion	lymphocyte activation	TNF
platelet abnormalities eosinophilia immune mediated inflammatory disease	activation immune	ARPC1B
organic mental disorders substance induced	organic hydroxy metabolic	MSRA
B cells expansion nfkb anergy	B cells	CARD11
abdominal obesity metabolic syndrome	organic hydroxy metabolic	MTTP

Table 2.2: Intersection of predicted genes with mice disease-gene-sets.

Ontology terms	Disease terms	Genes
Antigen processing and presentation via mhc class ib	thyroid gland Hurthle cell carcinoma	PSMB9
Negative regulation of leukocyte mediated immunity	leukocyte adhesion deficiency 1	ITGB2
Positive regulation of myeloid leukocyte differentiation	acute myeloid leukemia	DLEC1

Continued on next page

Table 2.2 – continued from previous page

Ontology terms	Disease terms	Genes
Innate immune response	immune dysregulation-polyendocrinopathy-enteropathy-X-linked syndrome	DOCK8
Homologous chromosome pairing at meiosis	hepatocellular carcinoma	ASPM
Peptide metabolic process	inherited metabolic disorder	NDUFS1
Regulation of lipid metabolic process	inherited metabolic disorder	MMAB
Nucleoside triphosphate metabolic process	bilirubin metabolic disorder	SOD2
Cerebral cortex development	transient cerebral ischemia	NEFM
Acute inflammatory response	acute lymphoblastic leukemia	PON1
Ear development	congestive heart failure	ERBB4
Cellular ketone metabolic process	bilirubin metabolic disorder	UGT1A1
Organic hydroxy compound metabolic process	abdominal obesity-metabolic syndrome 1	MTTP
Regulation of trans synaptic signaling	transient cerebral ischemia	KCNK2
Regulation of vascular associated smooth muscle cell differentiation	renovascular hypertension	FN1

2.3.4.2 Comparison of predicted results with other gene function prediction methods

Predicting gene functions has been a persistent challenge in computational biology. Recent approaches utilize diverse features such as primary amino acid sequences (DeepGo, NetGo 2.0), gene expression (correlation AnalyzeR), and network inference from transcriptomic profiles (GenetICA-Network) (289; 290; 291; 292). In Figure 2.5B, we com-

pared the abstract mining results of these methods with the novel associations inferred by our approach. The co-occurrence of the input ontology term and predicted gene term at least once in PubMed abstracts for 20 randomly selected gene-sets (Methods 2.2.4) is notably higher with our method compared to DeepGo, NetGo 2.0, GenetICA-Network, and AnalyzeR.

2.3.5 CRISPR-based validation of association of genes with major cellular processes of clusters of functions

Our method of clustering functions based on shared top predictors (TFs or cofactors) introduces innovative approaches to identify both direct and indirect associations among coding and non-coding genes with major cellular processes. To evaluate these novel links, an analysis of available CRISPR screens was conducted. Initially, the 'viability' CRISPR screen in human pluripotent stem cells (hPSC) highlighted crucial genes enriched in hPSCs, focusing on transcription factors and proteins associated with the cell cycle and DNA repair (295). Genes identified through our method within cluster-47, primarily associated with the cell cycle process, demonstrated notably higher z-scores in comparison to an equivalent number of randomly selected genes in the same CRISPR screen assessing hPSC viability (Figure 2.5C). However, these predicted genes for cluster-47 displayed comparatively lower z-scores in other CRISPR screens, such as 'phagocytosis,' 'resistance to chemicals,' and 'pyroptosis' (311; 294; 312).

In another validation method, a significant disparity was noted in the z-scores between predicted genes linked to the gene ontology term 'immune effector process' and randomly selected genes in the CRISPR screen for phagocytosis compared to other CRISPR screens (Figure 2.6). Furthermore, we validated cluster-7, which includes functions associated with the major cellular process 'immune system' (Supplementary File 3). The newly predicted genes in cluster-7 showed elevated z-scores compared to an equivalent number of randomly selected genes in the phagocytosis screen, which is a prominent process in immune response (313). However, these newly predicted genes in cluster-7 displayed relatively lower z-scores in other CRISPR screens (Figure 2.5D) (311; 312; 297). Validations using CRISPR screens underscore the associations of novel genes with major cellular processes and link the underlying regulatory factors (top predictors) to these cellular processes.

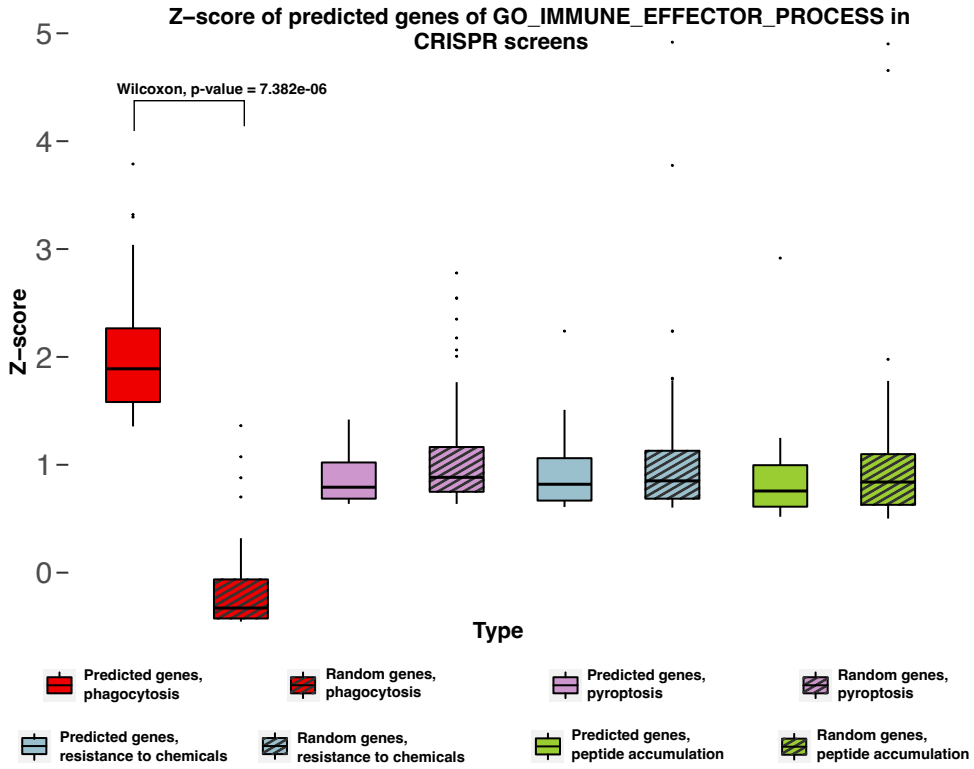

Figure 2.6: Comparison of z-scores for newly predicted genes for gene-set "immune effector process" in various CRISPR screens. The red box represents the CRISPR z-scores in the phagocytosis CRISPR screen. The purple, blue, and green boxes represent the CRISPR z-scores of the same predicted genes in pyroptosis, resistance to chemicals, and peptide accumulation CRISPR screens, respectively.
[Figure source: Chandra, Omkar, et al. "Patterns of transcription factor binding and epigenome at promoters allow interpretable predictability of multiple functions of non-coding and coding genes." "Computational and Structural Biotechnology Journal" 21 (2023): 3590-3603]

2.3.6 Explainability through insight into the association of binding patterns of TF-pairs with functions

The results from abstract mining in PubMed and the performance metrics of our model underscore the reliability of our predictive approach. However, enhancing interpretability and reliability requires a deeper exploration of TF-binding combinatorics. Investigating the basic combinatorial patterns of TF-pair bindings is crucial. TFs are known for their pleiotropic effects, influencing multiple biological functions simultaneously (314). As expected, certain TFs demonstrated high importance scores across a broad spectrum of functions (315). To assess the predictive pleiotropy of TF-pairs, TF ChIP-

seq pairs from identical cell types that emerged as primary predictors for diverse functions were analyzed. Several TF ChIP-seq pairs were identified as key predictors for multiple functions (Figure 2.7A).

The presence of TF-pairs among the top important features across multiple biological functions highlights their predictive pleiotropy (316) (Figure 2.7A). Further assessment was conducted to gauge the diversity of functions predicted by each TF pair. This analysis involved counting the occurrences of TF-pairs within clusters of co-regulated functions (Supplementary File 4). While TF-pairs showed predictive pleiotropy across numerous functions, their diversity within clusters of co-regulated functions was limited. This reduced diversity may be attributed to the clustering method's emphasis on common top predictors. Nevertheless, these clusters highlighted the cohesive nature of member functions in major cellular processes. Therefore, analyzing the diversity of TF-pairs' pleiotropic predictive power through clusters provided valuable insights into their regulatory impacts.

The co-occurrence of RUNX3 and BATF ChIP-seq patterns at promoters in B cells (GM12878) was observed together among the top 20 predictors for 11 functional gene-sets. These gene-sets were associated with only 2 clusters of functions primarily involved in immune cell activation and differentiation (Supplementary File 5). Similarly, the DNA-binding profiles at promoters in adipocytes by E2F4 and CEBPA were identified as top predictors for 8 gene-sets within a single cluster, predominantly associated with responses to stimuli such as insulin peptides, monosaccharides, and related metabolic functions. Therefore, our analysis underscores certain TF-pairs' specificity towards distinct clusters of functions, reinforcing predictions concerning the involvement of coding and non-coding genes in major cellular processes.

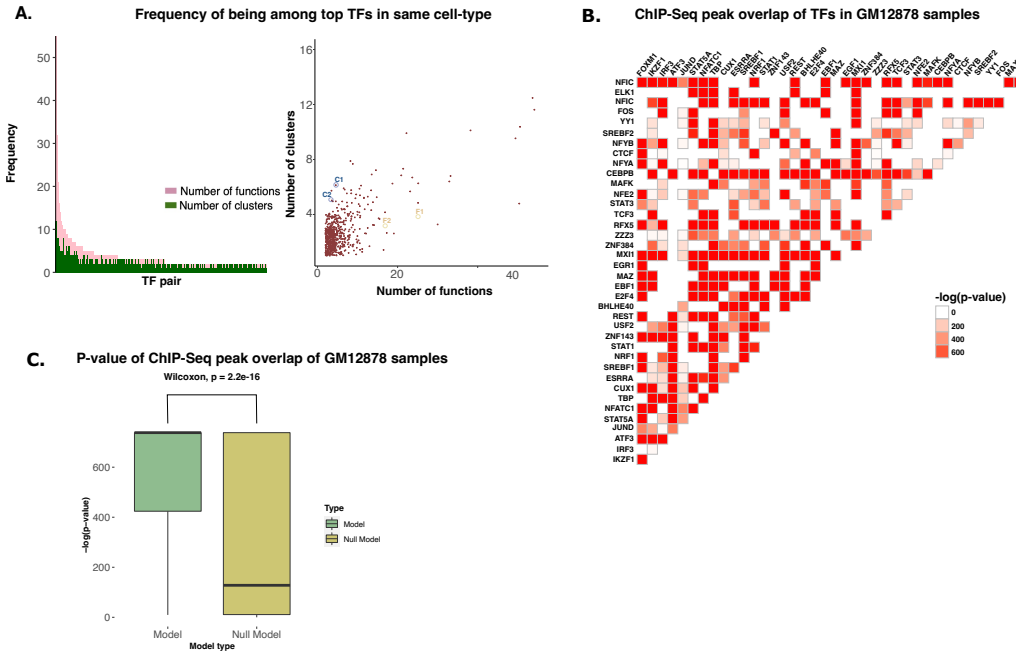

Figure 2.7: Provides insights into the co-occurrence and synergistic effects of Transcription Factor (TF) pairs as predictors. A) Depicts the distribution of TF ChIP-seq pairs among the top 20 predictors within the same cell type, categorized by functions (pink) and function clusters (green). The scatter plot on the right highlights these counts, with notable TF pairs such as C3: E2F4-GATA1, C4: MAZ-GATA1, F3: ZNF366-SPI1, and F4: SPI1-STAT1. B) Features a heatmap showcasing the statistical significance of TF ChIP-seq peak overlaps at promoters in GM12878 cells. C) Displays a box plot comparing the significance values (-log(P-value)) of promoter peak overlaps for TF ChIP-seq pairs that consistently appeared as top predictors across various functions in GM12878 cells. Additionally, the box plot on the right illustrates the significance of overlaps among random TF ChIP-seq profile pairs in GM12878 cells.
[Figure source: Chandra, Omkar, et al. "Patterns of transcription factor binding and epigenome at promoters allow interpretable predictability of multiple functions of non-coding and coding genes." "Computational and Structural Biotechnology Journal" 21 (2023): 3590-3603]

Certain TF pairs were identified as top predictors for gene-sets linked to a wider array of co-regulated function clusters. Specifically, pairs involving CTCF exhibited greater diversity within these clusters. For example, ZCAN5FB and CTCF were top predictors for functions spanning 12 different co-regulated clusters. Similarly, TET3 and CTCF were top predictors for functions from six distinct clusters. CTCF is widely recognized for its extensive influence beyond its traditional role as an insulator (317). The co-occurrence of CTCF with specific TFs as top predictors implies a significant role in various cellular processes. Additionally, the same analysis assessing pleiotropy and diversity was applied to TF-pair ChIP-seq profiles from different cell types (Figure

2.8A).

Certain TF pairs from the same family show specificity towards similar functions like immune response; for example, a pair of TFs, GATA1-GATA, belongs to the GATA family of several transcription factors that share a conserved DNA-binding domain known as the GATA domain. appear to be among the top 20 predictors for the following functions (Supplementary File 4):

1. GO NUCLEAR MEMBRANE
2. GO REGULATION OF PROTEIN POLYMERIZATION
3. GO PHAGOCYTOSIS
4. GO TRANSCRIPTIONAL REPRESSOR COMPLEX
5. GO ACTIN NUCLEATION

Another example is the specificity of the TF pair FOXA2-FOXF1, which are both members of the Forkhead box (FOX) family, characterized by a conserved winged-helix DNA-binding domain. They show specificity toward certain gene functions (Supplementary File 4):

1. GO REGULATION OF ENDOCYTOSIS
2. GO POSITIVE CHEMOTAXIS
3. GO CYCLIC NUCLEOTIDE BINDING

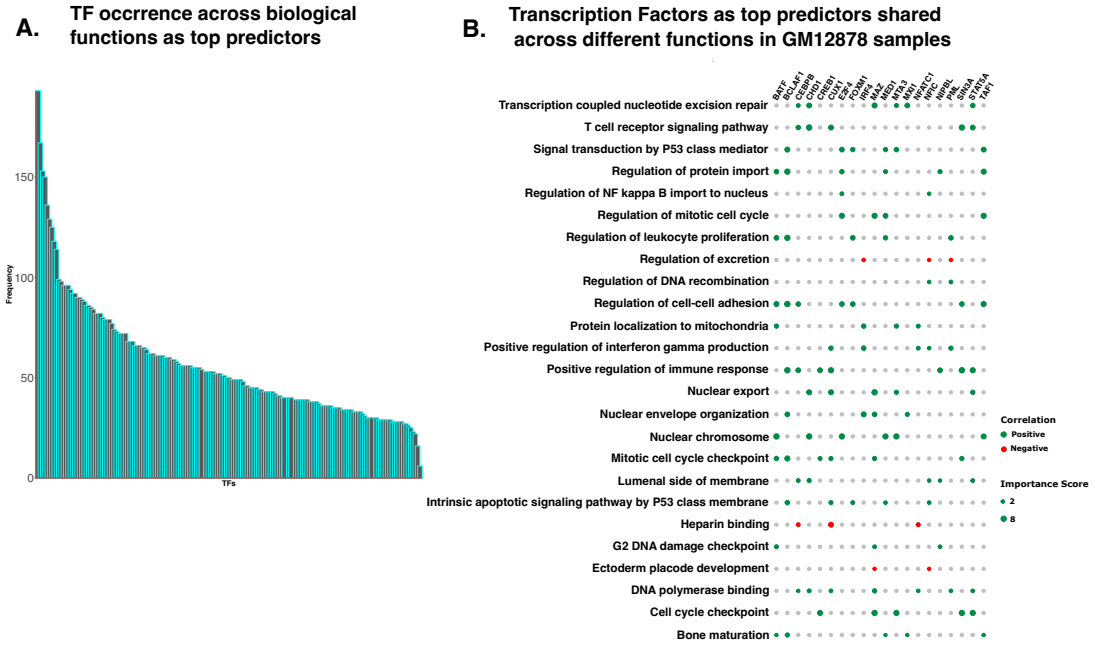

Figure 2.8: The predictive pleiotropy of TFs is illustrated as follows: (A) Shows the distribution of functions where a single TF emerged as a top predictor based on GM12878 cells. The size of the dots corresponds to the feature importance score, while the color indicates the directionality of the relationship. (B) Presents a dot plot highlighting the count and feature importance of TF ChIP-seq profiles ranked as top predictors in GM12878 cells. Only functions specific to GM12878 cells are included.
[Figure source: Chandra, Omkar, et al. "Patterns of transcription factor binding and epigenome at promoters allow interpretable predictability of multiple functions of non-coding and coding genes." "Computational and Structural Biotechnology Journal" 21 (2023): 3590-3603]

To validate the non-random occurrence of TF-pairs as top predictors (Figure 2.7B), the overlap of their ChIP-seq profile peaks in the GM12878 cell line was examined. This analysis was premised on the idea that if TF-pairs appeared as top predictors purely by chance, their peak overlaps would be random. The R package ChIPpeakAnno (318) was employed to analyze the TF ChIP-seq peaks in the GM12878 cell line. The overlap of peaks from co-predictive TF-pairs within the same cell type was compared with that of random TF-pairs as a control. Co-predictive TF-pairs were defined as pairs of TF ChIP-seq profiles in GM12878 cells that appeared among the top 20 predictors for any function (Figure 2.8B). The results showed significantly higher enrichment of peak overlap at promoters in GM12878 for co-predictive TF-pairs compared to random TF ChIP-seq pairs (Figure 2.7C). These results affirm the credibility of the method and suggest that examining top predictors can uncover insights into TF-TF cooperation by their frequent co-binding at gene promoters associated with same biological functions.

2.3.7 Broader applicability of GFPredict and its utility for predicting functions of non-coding RNAs

Our findings underscore the robustness of our method for ontology-based function prediction of the genes, which hinges on TF binding patterns. The GFPredict framework demonstrates the potential for broader application across various biologically relevant gene-sets. Numerous studies have identified different gene-sets linked to diverse phenotypes and biological functions. We aimed to address a practical need by reliably predicting and validating gene-function associations efficiently, minimizing experimental efforts. To demonstrate the effectiveness of our method, we utilized publicly available CRISPR screen datasets. From each dataset, the top 50 genes were selected as positives, while non-positive random genes from the training data served as negatives. After training GFPredict, the top 30 predicted genes were validated using their respective CRISPR screen scores. For instance, employing the top 50 positives from the CRISPR screen for resistance to chemicals (in fibroblasts) (299) to train GFPredict resulted in the top 30 predicted genes for this function exhibiting significantly higher scores (P-value < 0.004) compared to 30 randomly selected genes from the same CRISPR screen (299). Similarly, in the cell cycle CRISPR screen, the top 30 predicted genes displayed significantly higher scores (P-value < 1e-4) compared to 30 randomly selected genes (295). Results from two additional CRISPR screen-based analyses are presented in Figure 2.9. These findings collectively indicate that GFPredict effectively predicts genes associated with various biologically relevant gene-sets, in addition to its utility in typical ontological functions.

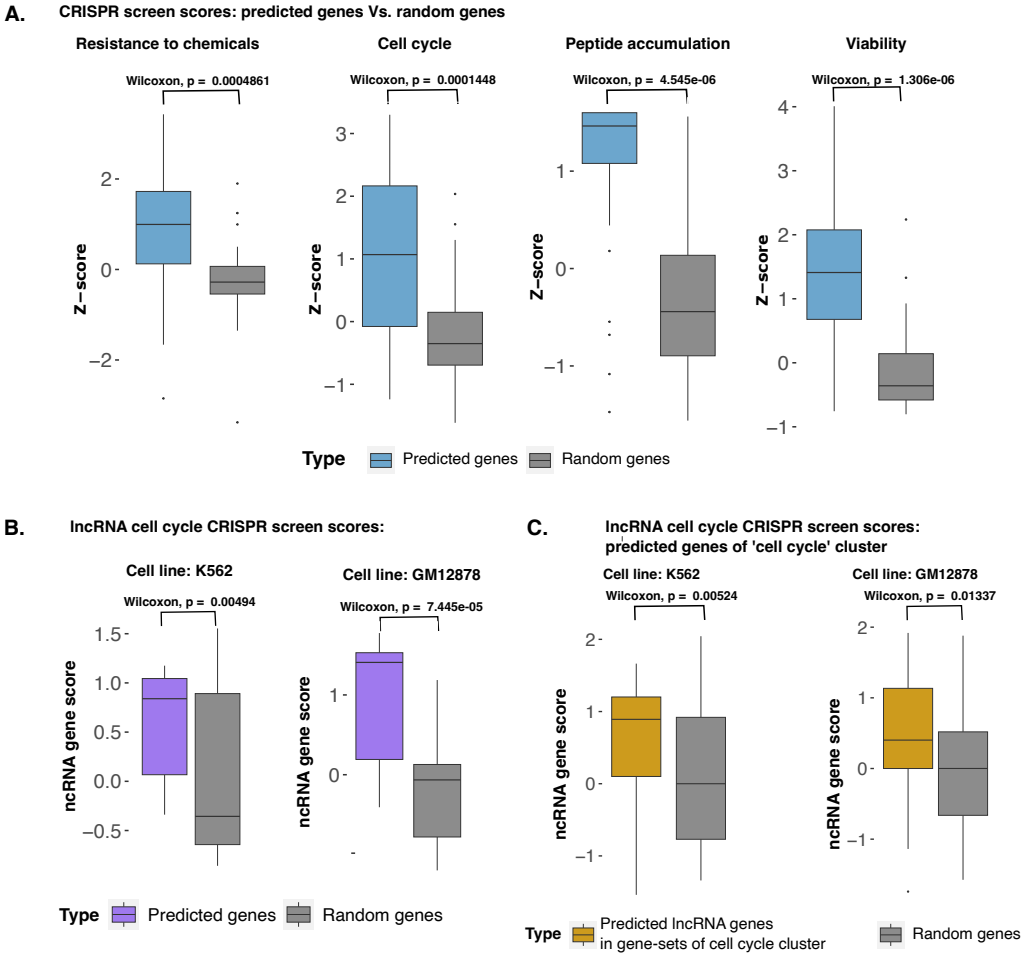

Figure 2.9: A) Comparison of CRISPR scores between the top 30 genes predicted by the GFPredict modeltrained on the top 50 genes from CRISPR screensand a set of random genes. The top 30 predicted genes were excluded from the training dataset. B) CRISPR scores of lncRNA genes among the top 30 predicted genes in the lncRNA-CRISPR screen for the cell cycle, identified by GFPredict, which was trained on the top 50 positive coding genes from a different cell-cycle CRISPR screen. Out of the top 30 predicted genes, 15 were lncRNA genes. C) Comparison of CRISPR scores between 52 lncRNA genes predicted to belong to the cell cycle cluster (cluster-47, as shown in Fig. 3) and random genes in the lncRNA-CRISPR screen for the cell cycle.
[Figure source: Chandra, Omkar, et al. "Patterns of transcription factor binding and epigenome at promoters allow interpretable predictability of multiple functions of non-coding and coding genes." "Computational and Structural Biotechnology Journal" 21 (2023): 3590-3603]

2.3.7.1 Application of expanding small CRISPR screens for non-coding genes function prediction

GFPredict facilitated the prediction of relationships between 1200 long non-coding genes and a wide range of biological processes and molecular functions (Table 2.3). To

assess the accuracy of ncRNA function predictions, we employed a rigorous evaluation strategy. GFPredict was initially trained on the top 50 genes identified in a cell cycle CRISPR screen, which mainly comprised coding genes (295). Subsequently, we validated these predictions using a separate CRISPR screen specifically designed to identify lncRNA genes associated with the cell cycle (300). The top 30 lncRNAs predicted by GFPredict to be linked with the cell cycle exhibited significantly higher CRISPR screen scores compared to randomly selected sets of the same size in two cell lines (K562 and GM12878) (Figure 2.9B).

The clustering of functions based on co-predictive TFs and cofactors enabled linking non-coding RNAs to various major cellular processes (Supplementary File 6). To validate these associations, ncRNA CRISPR screens were used (Figure 2.4A). Specifically, ncRNAs predicted to be part of gene-sets within cluster-47, associated with the cell cycle, were selected (Supplementary File 3 and Figure 2.4). These ncRNA genes exhibited significantly higher CRISPR screen scores for the cell cycle compared to random genes (p-value < 0.01) (Figure 2.9C). This validation highlights the robustness of the model, showing that it can predict the functions of non-coding genes effectively, even when non-coding gene CRISPR screens are limited.

Table 2.3: List of predicted functions of non-coding RNAs with experimental evidence.

Ontology terms	**Predicted non-coding gene**	**Literature evidence**
Heart development	ENSG00000280339	(319)
Sterol homeostasis	LINC02356	(320)
Phospholipid metabolic process	ENSG00000257023	(321)
Eye development	AC078909	(322)
Neuron maturation	ENSG00000274367	(323)
Synapse organization	MIR4281	(324)
Regulation of homeostatic process	MIR658	(325)
Negative regulation of interleukin 6 production	LINC00528	(326)

Continued on next page

Table 2.3 – continued from previous page

Ontology terms	Predicted non-coding gene	Literature evidence
Utero Embryonic development	MIR5001	(327)
Keratinocyte differentiation	PAUPAR	(328)

2.4 Discussion

Traditionally, predicting the functions of non-coding RNAs requires new assays or unique genomic features. Our research, however, shows that TF-binding profiles can be effectively used for this purpose. TFs act as universal regulators for both coding and non-coding genes associated with the same functions, as demonstrated by our results (Figure 2.9B) (329).

Through the use of various machine learning models, we achieved robust predictive performance, with sensitivity exceeding 80% and specificity surpassing 90% across more than 780 functions when utilizing all available features. Focusing on TFs and cofactors identified from ChIP-seq data (encompassing 736 TFs and 87 cofactors), we maintained high accuracy for predicting 650 functions. Employing random forest ML models, we achieved a minimum AUROC of 0.8 for over half of the gene-sets analyzed (5022 out of 9559), indicating strong predictive capability (302). Our findings were independently validated using separate datasets, and we compared GFPredict's performance against other methods employing diverse features such as primary amino acid sequences (DeepGo, NetGo 2.0), gene expression (correlation AnalyzeR), and gene-cofunctionality-based networks (GenetICA-Network) (289; 290; 291; 292). This comparative analysis underscores the effectiveness of our approach, which leverages epigenomic data and TF-binding patterns at promoters, providing a distinct advantage in predicting non-coding RNA functions.

In subsequent analysis, clustering of functions revealed a notable pattern: gene-sets sharing common top predictors, particularly ChIP-seq profiles from identical cell types, tended to align with similar major cellular processes upon manual annotation. This convergence occurred despite initial perceptions of disparate biological roles among these

gene-sets, highlighting their association with overarching processes such as cell cycle regulation and transport. This observation was underpinned by shared epigenomic and TF-binding patterns at gene promoters. Moreover, the identification of common epigenomic and TF-binding features as significant predictors for two distinct gene-sets suggested their involvement in comparable major functions. This analysis transcends traditional gene-set definitions, emphasizing the pivotal role of TFs. For example, in cluster-47, key predictors included PARP-1, MYC, SMARCA4, and CTCF, all recognized for their roles in regulating the cell cycle (303; 304; 305; 306).

While previous studies have touched on the co-regulation of functions through TF and cofactor binding (330), our research stands out by extensively analyzing common top predictive TFs, illustrating their interdependence in molecular and biological processes. This analysis also offers insights into how perturbations in key regulators can potentially affect a wide range of functions. Our study introduces novelty in two main ways: i) decoding the combination of TF-binding patterns at gene promoters to associate them with specific functions, and ii) clustering established gene-sets using top co-predictors to identify shared major cellular processes across clusters. These function groups, encompassing both molecular functions and biological processes, promise to provide deeper insights into CRISPR screens, revealing the roles of coding and non-coding genes in broader cellular contexts such as the cell cycle and immune response (295; 296). The derived clusters from GFPredict, which represent ontological functions tied to major cellular processes, can enhance the understanding of how genes identified in CRISPR screens relate to specific biological processes and molecular functions.

Our downstream analysis highlights the robustness and sensitivity of our models in accurately predicting non-coding RNA functions. Furthermore, through the clustering of functions, we have elucidated the broader roles played by specific non-coding RNAs (Supplementary File 6). For instance, our approach identified non-coding RNA genes such as LINC01137, LINC00441, DLG1-AS1, and UBL7-AS1, associating them with various functions predominantly involved in cell cycle activity. UBL7-AS1 (331) and DLG1-AS1 (332) are known to play roles in proliferation, whereas LINC00441 (333) and LINC01137 (334) are linked to cancer development. These findings provide valuable insights into the roles of non-coding RNAs in major cellular processes, offering guidance to biologists in designing experiments for further validation.

Our approach, which integrates TF and cofactor binding patterns as features for predicting gene functions and clusters functions to uncover the roles of both coding and non-coding genes, distinguishes itself from traditional methods of gene-function prediction. While existing tools like the Cistrome BETA suite (335), which predicts the regulatory behaviors of transcription factors, and Reshef et al.'s method (336) for identifying directional effects of functional annotations on diseases through signed linkage disequilibrium profile regression, utilize transcription factor ChIP-seq profiles in various ways, there is limited research focused on utilizing TF-binding patterns at promoters to predict the functions of both coding and non-coding genes. Moreover, tools such as MAGIC (337) identify TFs and cofactors responsible for gene expression changes across different conditions, but few studies interpret the relationship between combinatorial TF-binding patterns at promoters and clusters of functions. This underscores the novelty and distinctive nature of our approach and its analytical framework.

The resource has been developed to aid biologists in validating their experimental findings and using predictions to guide the design of experiments aimed at elucidating the biological and molecular roles of both coding and non-coding genes. However, it should be noted that the current version of GFPredict may not achieve optimal accuracy across all ontological gene-sets. Specifically, out of 9559 gene-sets, our method achieves a minimum AUROC of 80% for only 5022 gene-sets (52%) using the random forest model. This limitation may stem from several factors: firstly, some functions may lack sufficient positive genes for training the prediction model; secondly, the number of features (such as TFs, DNase-seq, CAGE-tags) may be inadequate; and thirdly, additional types of features beyond epigenome, TF-binding, and CAGE-tags patterns may be necessary for many gene-sets. Incorporating TF-binding signals from enhancer-bound promoters could potentially enhance the accuracy of gene function prediction. As the consensus list of enhancer-promoter interactions improves across different cell types, our method could be adapted to integrate enhancer data, thereby improving its accuracy and utility. Nevertheless, our analysis provides valuable insights into epigenomic and TF-binding patterns at the promoters of ncRNA genes, which are informative for predicting their functions. We expect that advancements in new epigenomic profiling technologies like FloChIP (338) and multi-CUT&Tag (339) will expand the availability of epigenome profiles, enabling GFPredict to achieve higher prediction accuracy across a broader range of functions. Our study emphasizes the importance of

leveraging diverse TF-binding profiles and epigenomic to enhance our understanding of non-coding RNA functions.

The ChIP-seq profiles and other datasets that are used as features, along with their source described in the Methods section, have an inherent drawback in that they cannot capture the complete ground truth of the epigenomic regulation of the cells as intended in our computational pipeline. Despite this, as discussed above, we can get good predictability in terms of sensitivity and specificity in a good fraction of ontological gene sets. This decrease in good predictability for ontological gene sets is because of the lack of ChIP-seq profiles of multiple TFs from individual cell types in the publicly available databases. This leads to skewness in the good predictability towards ontological functions regulated by TFs and epigenomic features of cell types that are present in the dataset. To mitigate this bias, using single-cell ChIP-seq profiles of TFs and epigenomes of different cell types is necessary. With the availability of such features in the near future, it will be possible to get good predictions for all types of functions.

In our current approach, we only utilized the epigenome and TF binding signatures of the promoters of the genes. However, as Chapter 1 of this thesis discusses, enhancers play a major role in regulating gene expression. Therefore, it is necessary to identify the association of enhancers to specific biological processes and molecular functions. This will be the future work of the thesis, by calculating the epigenome and TF binding scores at enhancers and to model their association with different ontological gene-sets

CHAPTER 3

Chapter 3. Explainable models using transcription factor binding and epigenome patterns at promoters reveal disease-associated genes and their regulators in the context of cell-types

Establishing that epigenome and transcription factor (TF) binding patterns are predictive of the ontology-based functions of coding and non-coding genes, further understanding the role of (TFs) and epigenomic marks in disease was sought. This chapter explores the use of transcription factors and epigenomic marks as features to predict their associations with various diseases. This method also enabled the identification of novel associations between coding and non-coding genes and diseases. Validation of these novel predictions was conducted using GWAS data and survival analysis. One remarkable finding of this study is the identification of connections between various immune cells and disease conditions through TF binding pattern, even when these diseases are of different cell type than the linked immune cells. Furthermore, a few case studies indicate how such connections can be exploited for the diagnosis, prognosis, and potentially the therapeutics of diseases.

3.1 Role of epigenome and TFs in diseases and approaches to identify such links

Gene expression regulation within specific biological processes is controlled by a very few unique combination of transcription factors (TFs) and cofactors, among all the regulatory elements. Disruption of genes' expression participating in these processes can result in serious diseases. Researchers have concentrated on exploring the roles of epigenomics and TFs in contributing to different disease conditions. Van Ouwerkerk et al. examine the impact of epigenomic modifications and transcription factors (TFs) on

atrial fibrillation (AF) (340), emphasizing that numerous genomic loci linked to AF are found in non-coding regions that serve as binding sites for heart-specific TFs. These TFs play a pivotal role in modulating gene expression by influencing TF activity and chromatin epigenetic status. Mazzone et al. provide an overview of the effects of DNA methylation, histone modifications, and non-coding RNAs on autoimmune diseases such as lupus, arthritis, and type 1 diabetes (341). Therefore, recognizing epigenetic changes and transcription factor binding patterns that lead to abnormal gene expression and genomic instability is essential. Creating models of transcriptional regulatory processes linked to diseases can uncover potential therapeutic targets and biomarkers for a variety of conditions.

Several computational methods have been used to pinpoint crucial transcription factors in different biological contexts. One example is NetAct, which employs computational modeling to outline a core gene regulatory network encompassing essential transcription factors.This method combines expression data with a TF-target database, examining gene co-expression patterns, cis-motif predictions, and TF-binding motifs (342). Li et al. devised an integrated modeling framework using time-series expression data to investigate the coordinated regulation of transcription factors, long non-coding RNAs (lncRNAs), and miRNAs during cardiac development (343). However, these methodologies may not always capture direct TF binding to target genes, as this is typically assessed through ChIP-seq assays. The presence of a TF transcript does not guarantee the binding of the corresponding TF protein at its target site. Furthermore, research exploring the association between diseases and TF binding patterns at gene promoter regions is limited.

This study aims to achieve two primary objectives. Firstly, the random forest algorithm was applied to model the relationship between various transcription factors, co-factors, and diseases, using them as features across diverse sets of disease-related genes (DisGeNET) (18). Our secondary objective was to reveal new associations between genes, including non-coding genes, and diseases. This approach is based on the regulation of both coding and non-coding genes, which relies on the active binding of multiple transcription factors and co-factors at their respective promoter regions. The binding of transcription factors is influenced by the local chromatin accessibility, characterized by open chromatin conformation, histone modifications and (DNase-seq, ChIP-seq,), as well as the expression levels of these genes. Similarly, non-coding genes

often exert regulatory roles over coding genes through analogous mechanisms.

Transcription factors, co-factors, and epigenetic markers collectively govern the expression of both coding and non-coding genes involved in identical biological processes. Modeling this regulatory mechanism has the potential to reveal new connections between genes (both coding and non-coding) and biological processes (344). The identification of genes associated with diseases remains a significant and difficult task. These genes can either cause or be affected by diseases, offering possibilities for therapeutic or prognostic applications. Choosing the right features that represents the class of genes that are involved in a disease condition in a first crucial step. This task is particularly challenging when connecting non-coding RNAs (ncRNAs) to diseases, given the diverse mechanisms through which ncRNAs regulate coding genes (345; 146).

Various computational approaches employ diverse strategies to establish associations between genes and diseases. Certain studies utilize gene expression data to link genes with specific diseases (346; 347; 348). However, changes in gene expression can sometimes be a consequence rather than a cause of the disease condition (349). Therefore, relying solely on expression changes may not reliably identify causal genes associated with diseases. Other methodologies utilize gene ontology information as features to link more genes to diseases, but this approach has its limitations. Non-coding genes are sparsely annotated in gene ontology (approximately 1200), which reduces the sensitivity of ontology-based models for predicting the functions of non-coding genes. Additionally, protein-protein interaction-based features are not effective in predicting non-coding genes. Some methods combine various data sources, including gene expression, protein-protein interactions, and gene ontology knowledge, to identify novel associations between genes and diseases (350; 351).

Most existing methods primarily focus on predictive capabilities and often lack insights into the regulatory mechanisms underlying genes implicated in diseases or their phenotypes. Moreover, these methods predominantly center on coding genes, which are well-annotated with biological processes and molecular functions, posing a challenge in accurately predicting associations between non-coding genes and diseases to unravel their regulatory roles. To tackle this challenge, we have developed a novel approach that integrates transcription factor binding patterns and epigenomic profiles to link both coding and non-coding genes with diseases. This method stands out by prioritizing the

regulatory roles of transcription factors in disease-associated genes and predicting the involvement of non-coding genes across various disorders.

3.2 Material and methods

3.2.1 Epigenome and TF-binding features score calculation for promoters

Each disease-related gene-set was treated as a distinct class, where known member genes served as positive instances, framing the identification of novel associations as a classification challenge utilizing epigenomic and transcription binding assay data. These features encompassed data from human samples and tissues mapped to the hg19 genome assembly. Histone modification ChIP-seq profiles, DNAse-seq open-chromatin profiles, and transcription factor binding profiles across multiple human cell types were retrieved from the ChIP-atlas database in bigWig format (281). Using the bigWigToBedgraph tool, the bigWig files were converted to bedGraph format. Normalization of scaled read-count scores was performed by averaging within each 200-base pair genomic bin to ensure consistency across datasets. For each ChIP-seq profile, the normalized read-count scores within 1 kilobase pairs (Kbp) of transcription start sites (TSS) were aggregated to derive a composite score for each gene. Similarly, CAGE-tag profiles from various cell types were obtained from the FANTOM project (RIKEN, Japan) in bam file format (283), converted to bedGraph format, and normalized based on mean read-counts in 200-bp bins. Aggregated scores from bins within 1 Kbp of TSS were then calculated to represent the activity levels of genes.

The transcription start sites (TSS) for non-coding genes were sourced from the gencode (V30) and RefSeq gene transcript databases (284; 3). Each gene was considered to have multiple TSS if the distances between them were at least 500 base pairs. In total, 89,747 promoter regions associated with these genes were examined in the analysis.

3.2.2 Prediction method

For each gene-set included in DisGeNET (18), designate annotated genes as positive data points and randomly choose an equal number of non-annotated genes as negative data points. Predicting disease-gene associations was approached as a comprehensible classification task using epigenomic and TF binding data as predictive features. We trained five distinct models for each gene-set: random forest, XGBoost, SVM (support vector machine), lasso regression, and ridge regression (L2-regularized logistic regression). To assess the models' performance across all gene-sets, we employed five-fold cross-validation, evaluating metrics such as error rate, accuracy, balanced accuracy, F1-score, and Matthew's correlation coefficient (MCC).

We implemented logistic regression with L2-regularization (ridge regression) using cv.glmnet with alpha = 0 and family = "binomial" from the glmnet R package. For the Random Forest algorithm, we used the randomForest function from the randomForest R package. For SVM models, we implement using the svm function from the e1071 R package. We used cv.glmnet with alpha = 1 (Lasso) from the glmnet R package for linear regression. To use, the XGBoost algorithm, use the xgb.train function from the xgboost R package.

We used our trained models to predict novel associations of genes with disease gene-sets, and we estimated reliable scores by calculating the maximum precision for each gene-set prediction.

3.2.3 Method for survival plots

Survival analysis was conducted using clinical and genomic data from patients with various cancer types sourced from TCGA (The Cancer Genome Atlas) (352). Diseases related to cancer were identified by comparing predicted disease names with TCGA cancer names using string matching techniques. Survival plots for genes of interest in each cancer were obtained using the 'GEPIA2' server (353). The server's API was utilized to collectively search for predicted genes associated with each cancer type. Subsequently, for each cancer type, survival plots were extracted for the predicted genes, and p(HR) values were calculated. The p(HR) value signifies the p-value of the hazard ratio between high and low gene expression levels within a specific cancer type, serving

as an indicator of the survival difference based on gene expression levels.

The analysis was conducted using Python programming to execute these procedures. For each cancer type, genes that were not associated were randomly sampled to generate a 'null model'. Under the null hypothesis, survival plots and p(HR) values were obtained for these genes as well. The statistical analysis included comparing the p(HR) values between the predicted genes and those derived from the null model for every type of cancer. With significance set at $p < 0.005$, the predictions for each cancer type were assessed using the MannWhitney U test.

3.2.4 Method for GWAS validation

The SNP (Single nucleotide polymorphism)-disease association datasets were obtained from the GWAS catalog database (354). To validate the model's predictions, the disease terms from DisGenNet used in training were compared with the disease terms in the GWAS data. If a disease term matched, the corresponding predicted gene was checked in the matching row of the GWAS data. If both the disease term and gene term matched, a score was assigned to the predicted associations.

3.3 Results

Each gene-set cataloged in DisGeNet (18) was treated as a separate class, and predictive models were constructed using epigenomic data and TF ChIP-seq patterns at gene promoters to predict their linkage with respective disease gene-sets. All available features, including epigenomic data, TF profiles, co-factors, open-chromatin details, and CAGE tag profiles, were utilized to forecast gene-disease associations. Our focus was particularly on leveraging transcription factor binding patterns as a critical feature to gain insights into regulatory mechanisms.

3.3.1 Epigenome and TF binding patterns at promoters are predictive of gene-disease association

Performance of Machine Learning Models on Disease gene-sets

Figure 3.1A displays bar charts illustrating the outcomes of five-fold cross-validation using five distinct machine learning models. These models were employed to predict disease gene-sets based on TF binding and epigenomic patterns at gene promoters. Among the evaluated models (including SVM, XGBoost, linear regression, logistic regression), the random forest model demonstrated the highest performance, achieving an average AUROC of 80% across the five-fold cross-validation. Detailed performance metrics for all models can be accessed in Supplementary File 1.

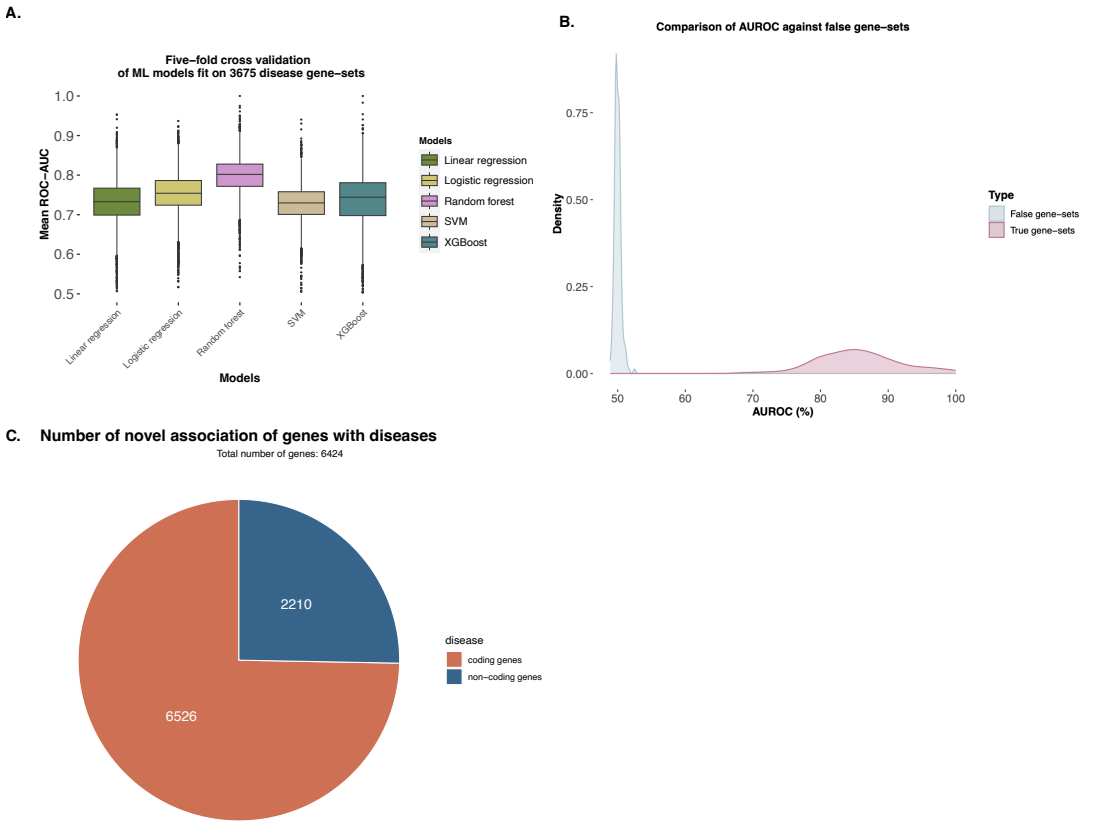

Figure 3.1: This figure presents the outcomes of our prediction model for identifying associations between diseases and genes: A) Gene-set memberships for various diseases were predicted using five distinct machine-learning models. The panel showcases the Operating Characteristic (ROC) values derived from predictions across 3,675 disease gene-sets. B) This section illustrates the distribution of ROC-AUC values, highlighting comparisons with other methodologies. The panel emphasizes the proportion of predictions that reveal novel associations through our approach.
[Figure source: Chandra, O., Pramanik, D., Gautam, S., Sharma, M., Dubey, N., Mahato, B. and Kumar, V., 2024. "Explainable models using transcription factor binding and epigenome patterns at promoters reveal disease-associated genes and their regulators in the context of cell-types." bioRxiv, pp.2024-05]

Specificity of Epigenome and TF Binding as Predictors of Gene-Disease Asso-

ciation

To validate the reliability of our modeling approach, we conducted feature shuffling on gene attributes. Using the top-performing machine learning algorithm, random forest, we applied it to genes with randomized TF binding and epigenome signals at their promoters. The null model demonstrated an average AUROC of 55% (Figure 3.1B), contrasting with an average AUROC of 85% achieved by the same gene-sets with unshuffled features. This outcome underscores the non-random nature and predictive efficacy of TF binding and epigenome features in forecasting gene-disease associations. Figure 3.1C presents the breakdown of coding and non-coding genes linked to diseases.

3.3.2 Independent validations of predicted gene-disease associations

3.3.2.1 Validation using PubMed abstract mining

PubMed abstracts were used to validate the predicted links between non-coding genes and disease gene-sets through an unbiased literature search. Between 1990 and 2022, this analysis assessed the frequency of co-occurrence between predicted non-coding gene terms and disease terms in abstracts. The results indicated a statistically significant increase in co-occurrences compared to random pairings of non-coding gene and disease terms (Figure 3.2A). These findings strengthen the validity of our predictions and increase confidence in the connections between non-coding genes and terms related to disease.

3.3.2.2 Validation using the result of GWAS

Using data from the GWAS database, we investigated the occurrence of mutations within the predicted genes associated with the same disease conditions to further validate our predictions. We observed a substantial overlap between the predicted genes and reported mutations, significantly higher than random gene-disease associations (Figure 3.2B). The dataset used for this analysis is available in the Supplementary File 2.

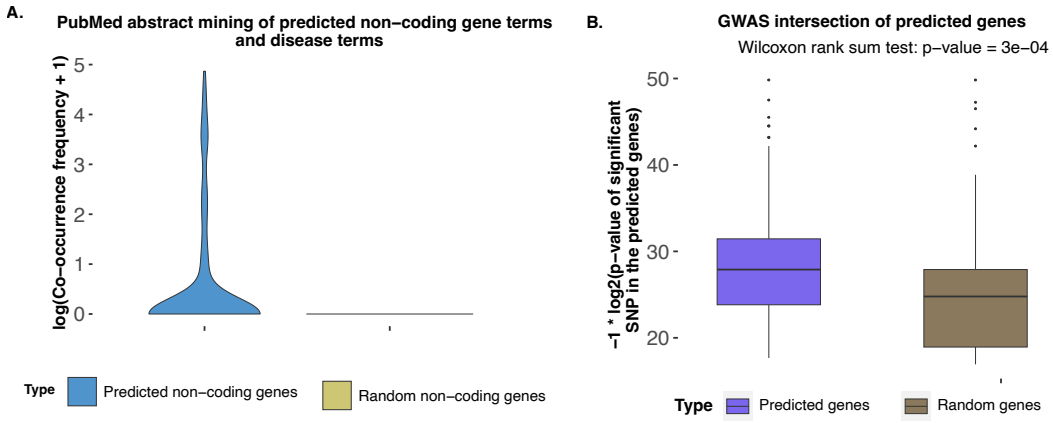

Figure 3.2: Validation using PubMed abstracts mining: The vertical axis of this figure depicts the frequency of disease names co-occurring with non-coding gene names in PubMed abstracts. For comparison, the figure also displays the co-occurrence frequencies of randomly paired disease names and non-coding gene names in PubMed abstracts. B) Validation Using GWAS Data: This section of the figure presents the validation outcomes utilizing GWAS-derived mutations in genes predicted to be associated with specific diseases. Only GWAS mutations corresponding to the predicted diseases were considered. As a control, the number of GWAS mutations associated with the target disease was assessed in randomly selected genes.
[Figure source: Chandra, O., Pramanik, D., Gautam, S., Sharma, M., Dubey, N., Mahato, B. and Kumar, V., 2024. "Explainable models using transcription factor binding and epigenome patterns at promoters reveal disease-associated genes and their regulators in the context of cell-types." bioRxiv, pp.2024-05]

3.3.2.3 Survival analysis of predictive transcription factors and predicted genes for diseases

The most critical features identified during the training and prediction of the machine learning model are TFs and co-factors. These features include ChIP-seq profiles of various TFs used to train and validate the model against known ground truth data.For each disease, the top 20 predictors were selected based on the importance values of features used as predictors after training the ML models. All top predictors for each disease-gene-set can be found in Supplementary File 3.

Hazard ratio results from survival analysis conducted with these top predictors showed significance for some of the cancer types. Figure 3.3 illustrates that for BLCA (Bladder Urothelial Carcinoma), the significance between the prediction and random selection of TFs is < 0.05. Subsequent analyses indicate that certain TFs among the top predictors for BLCA demonstrate significant values in survival analysis.

Survival analysis was further conducted on genes predicted to be associated with cancers based on these top predictors. However, analysis of predicted genes (both coding and non-coding) from these top predictors did not yield significant results in terms of cancer survival when comparing high versus low expression of individual genes.

Survival analysis of top predictors (TFs) for Bladder Cancer

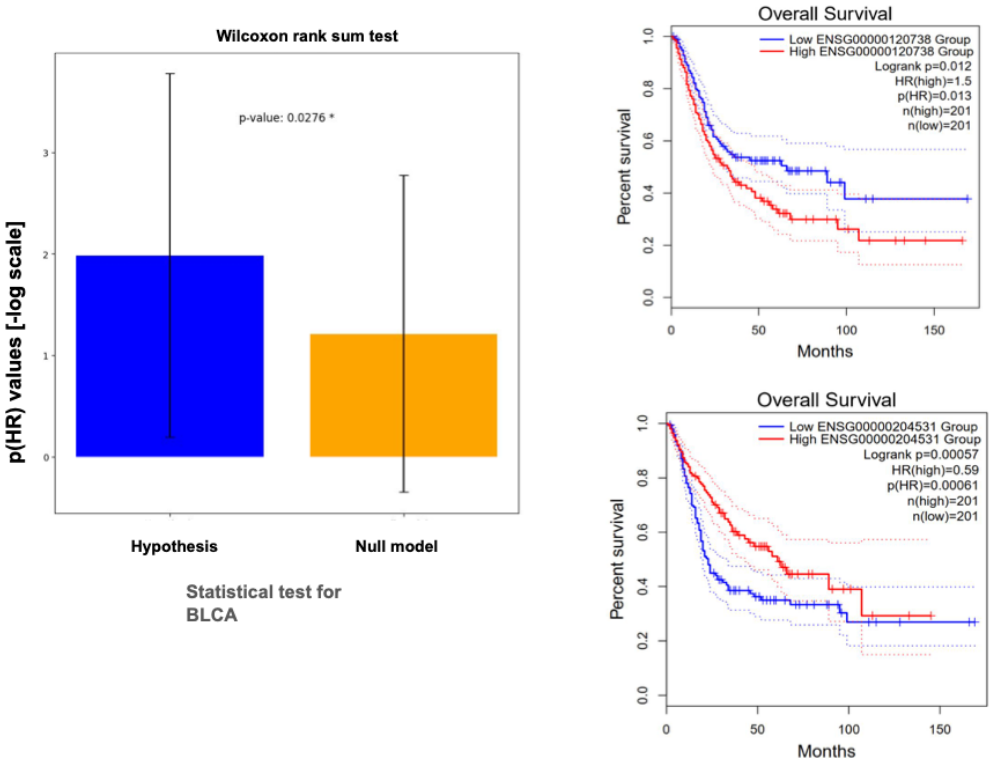

Figure 3.3: Survival analysis of genes associated with bladder urothelial carcinoma (TCGA-BLCA). The box plot on the left illustrates the statistical significance (-log(P-value)) of survival association. On the right, Kaplan-Meier plots depict the survival outcomes for genes predicted to be linked with TCGA-BLCA.
[Figure source: Chandra, O., Pramanik, D., Gautam, S., Sharma, M., Dubey, N., Mahato, B. and Kumar, V., 2024. "Explainable models using transcription factor binding and epigenome patterns at promoters reveal disease-associated genes and their regulators in the context of cell-types." bioRxiv, pp.2024-05]

3.3.3 Regulatory insights using the association between predictive TF and diseases

Exploring disease-Gene associations using TF and co-factors as features, top predictive TFs were identified using outlier criteria among top predictors. Figure 3.4 for detailed model performance metrics using only TFs and co-factors as features. Our analysis

highlighted notable TF binding patterns as influential predictors for diseases, underscoring their potential in identifying novel disease associations (Supplementary File 4). Also, check Table 3.1 for few examples obtained using the cell type-specific predictive TF association with the diseases. Results elucidating these associations and their regulatory implications are presented below:

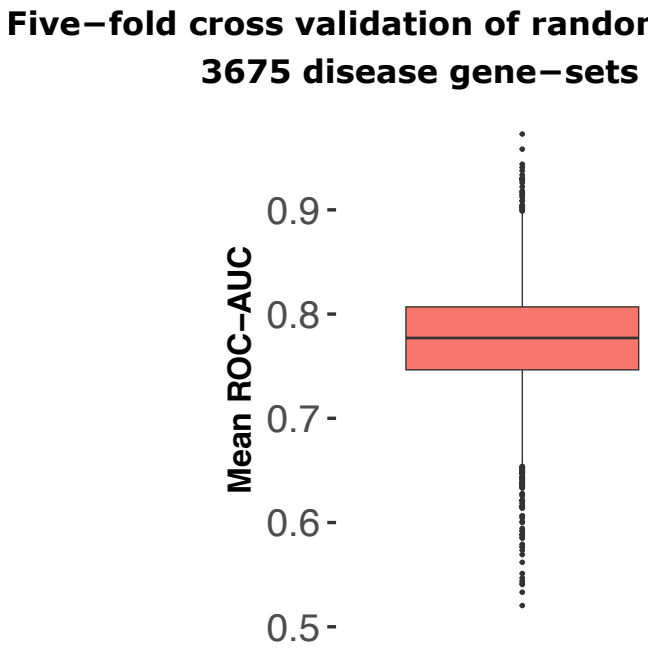

Figure 3.4: Five-fold cross validation result of random forest model train on disease-gene-sets using TFs as features. bioRxiv, pp.2024-05]

Case study of ZNF366

We discovered that the promoter binding profile of ZNF366 in dendritic cells emerged as a prominent outlier predictor across several diseases, including various lymphomas, arthritis, dermatitis, and certain types of leukemia. ZNF366, also referred to as DC-SCRIPT (dendritic cell-specific transcript), is typically expressed exclusively in dendritic cells. Our findings underscore the significant involvement of dendritic cells in diseases where the ZNF366 binding pattern (ChIP-seq in dendritic cells) plays a pivotal predictive role. For example, numerous studies have highlighted the role of dendritic cells in lymphomas such as adult and childhood diffuse large B-cell lymphoma, as well as classical Hodgkins lymphoma (355; 356). Additionally, defective dendritic cells have been implicated in leukemia (357). Apart from leukemia and lymphoma, ZNF366 was also identified as a leading predictor for genes associated with dermatitis, pancreatitis,

and encephalitis.

Case study of LDB1

The binding profile of the transcription factor LDB1 (Lim Domain Binding Protein 1) in erythroblast cells emerged as a significant predictor for genes associated with two specific types of anemia: hemolytic anemia and iron-refractory iron deficiency anemia. Despite incorporating binding profiles of numerous transcription factors across various cell types as features, the prominence of the LDB1 binding pattern in erythroblast cells as the foremost predictor for genes related to anemia highlights the specificity of our approach. LDB1 plays a critical role as a master regulator in erythroid differentiation, and targeted deletion experiments in adult mice have demonstrated its essential role in preventing severe anemia (358).

Interestingly, the binding pattern of LDB1 in erythroblast cells also emerged as a notable predictor for apnea. Further investigation through literature analysis suggested a potential link between erythrocyte count and characteristics and apnea. Additionally, our study identified an association between the LDB1 binding pattern in erythrocytes and Parkinson's Disease, Familial, Type 1 (PDFT1). Previous research partially supports this connection, highlighting elevated levels of oligomeric -synuclein in erythrocytes even during the early motor stages of Parkinson's Disease (359; 360).

In the case of PDFT1, we also observed that the promoter binding profiles of STAT5A in T lymphocytes served as predictive markers for associated genes. Since neuroinflammation plays a crucial role in the progression of Parkinson's disease and is mediated through T cells, STAT5A in T cells likely plays a significant role in regulating genes linked to PDFT1 (361; 362). Therefore, these key predictors for disease-related genesets not only identify relevant transcription factors but also emphasize specific cell types that may contribute to the disease's pathogenesis or symptoms.

Table 3.1: Cell type-specific association of TFs with diseases

Disease	**TF and Cell Type**	**Reference**
Acute Erythroblastic Leukemia	BPTF_Erythroid Cells	(363)
Continued on next page		

Table 3.1 – continued from previous page

Disease	TF/Gene-Cell Type	Reference
Childhood Diffuse Large B-Cell Lymphoma	ZNF366_Dendritic Cells	(364)
Arthritis, Adjuvant-Induced	ZNF366_Dendritic Cells_P	(365)
Arthritis, Psoriatic	ZNF366_Dendritic Cells	(366)
Acute pancreatitis	ZNF366_Dendritic Cells	(367)
Chlamydia Infections	ZNF366_Dendritic Cells	(368)
Malaria, Falciparum	TAL1_Erythroid Precursor Cells	(369)
Hepatitis	SPI1_Dendritic Cells	(370)
Anemia, Hemolytic	LDB1_Erythroblasts	(371)
Skin Diseases, Genetic	MAFB_Keratinocytes	(372)
Abnormal vision	E2F4_HRPEpiC	(373)
Central Serous Chorioretinopathy	LHX2_HRPEpiC	(374)
Hypotrichosis	KLF4_Keratinocytes	(375)
Skin Diseases, Genetic	MAFB_Keratinocytes	(376)
Increased hepatocellular carcinoma risk	RNF2_293	(377)

3.3.4 Association of non-coding genes

During transcription, regulatory factors governing coding genes also exert influence on non-coding genes involved in similar biological processes. By integrating epigenomic data and transcription factor binding patterns into our models for disease-gene associations, we successfully predicted connections between numerous non-coding genes and various diseases (Supplementary File 3).

Our model identified an association between MIR137HG and the 'Ductal Breast Carcinoma' disease gene-set. Located on chromosome 1 at position p21.3 (Figure 3.5A), MIR137HG was previously unlinked to breast cancer gene-sets. Lee et al. explored its role alongside the DEL1 gene in triple-negative breast cancer (TNBC) patient

samples. Their luciferase reporter assay confirmed direct binding of MIR137HG to the 3-UTR of DEL1, modulating tumor growth and apoptosis inhibition through the p53 pathway alteration (378; 379). They noted decreased MIR137HG expression and increased DEL1 expression in TNBC compared to normal breast tissue. Overexpression of MIR137HG reduced TNBC cell proliferation, invasion, and migration by suppressing DEL1 expression (380).

Additionally, the long non-coding RNA LINC00877 was associated with the "Meningococcal Infections" gene-set, located on chromosome 3 at band p13 (Figure 3.5B). Genome-wide association studies (GWAS) supported this prediction, linking multiple variants and risk alleles of LINC00877 to monocyte count (rs11708187-A, rs9809116-G, rs55890339-T) and platelet counts (rs9809116, rs11708187, rs12497693). These findings suggest LINC00877's involvement in the prothrombotic trait associated with meningococcal infections (381). Monocyte-derived macrophages play a pivotal role in fighting infections, and platelet count can indicate infection severity (382).

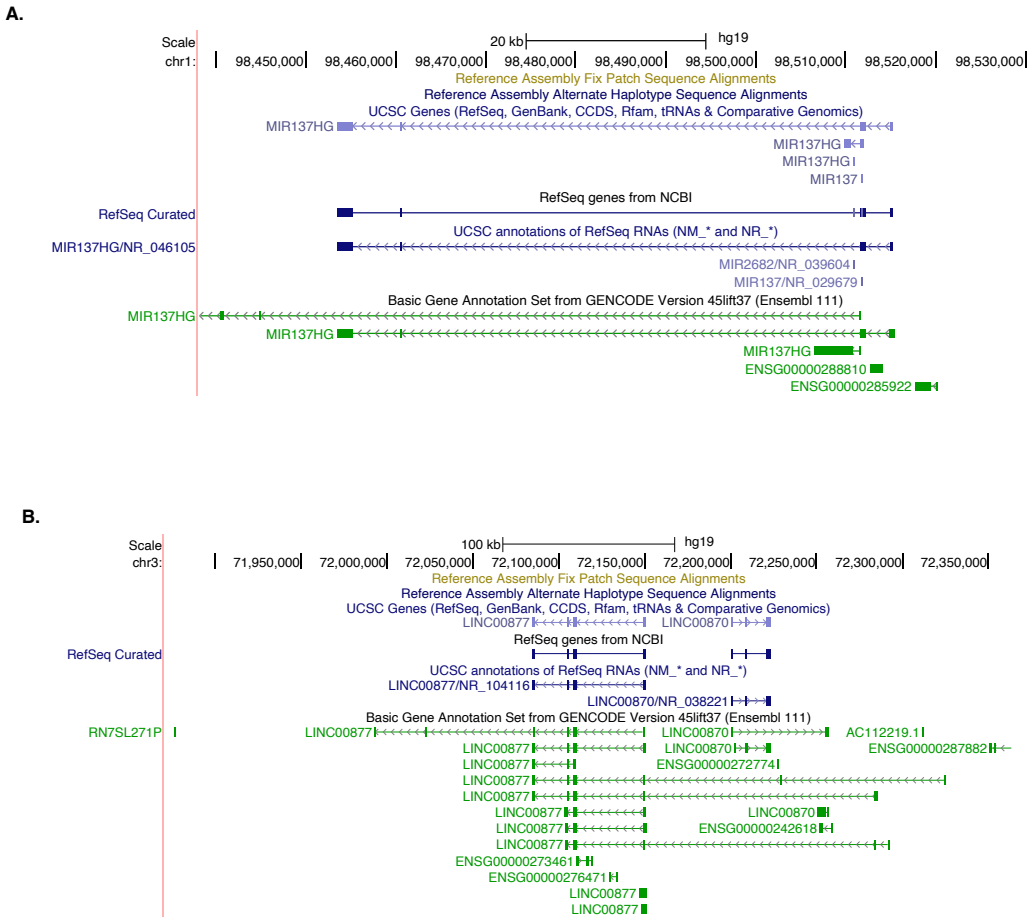

Figure 3.5: Screenshot from the UCSC Browser. The screenshot from the UCSC Browser illustrates the genomic positions of two non-coding genes whose functions were predicted using our methodology: A) Genomic location of MIR137HG gene. B) Genomic location of LINC00877 gene.
[Figure source: Chandra, O., Pramanik, D., Gautam, S., Sharma, M., Dubey, N., Mahato, B. and Kumar, V., 2024. Explainable models using transcription factor binding and epigenome patterns at promoters reveal disease-associated genes and their regulators in the context of cell-types." bioRxiv, pp.2024-05]

3.4 Discussion

Numerous approaches have been proposed by various research teams to identify connections between genes and diseases. However, most of these methods prioritize prediction accuracy without thoroughly investigating the key regulatory factors influencing disease-related genes. Few studies have leveraged epigenomic data and transcription factor binding patterns to predict associations between diseases and genes, thereby uncovering potential links involving non-coding RNAs. Moreover, studies focused on regulatory inference rarely incorporate interpretable prediction frameworks, unlike our

approach. Therefore, the novelty of our method lies in identifying transcription factors as top predictors, providing valuable evidence-based insights.

In addition to transcription factors, the specific cell types used for their ChIP-seq profiles have also offered significant insights into the roles of diverse immune cell types in disease processes. For example, our analysis predicted an association between the ChIP-seq profile of the transcription factor ZNF366 in dendritic cells and psoriatic arthritis. Catapano et al., using blood samples from generalized pustular psoriasis (GPP) patients, demonstrated that interleukin-36 production by immune cells directly impacts plasmacytoid dendritic cells, enhancing toll-like receptor (TLR)-9 activation and IFN- production (383). Elevated IFN- levels trigger inflammation, contributing to increased extracutaneous comorbidities in psoriatic individuals, who are at higher risk of developing psoriatic arthritis (384). Furthermore, Catapano et al. observed elevated expression of the ZNF366 gene in the blood of GPP patients, highlighting its potential role in disease pathology (383).

Our analysis identified significant associations involving ZNF366 ChIP-seq data from dendritic cells and 'acute pancreatitis.' Research by Bedrosian et al. underscored the critical role of dendritic cells in pancreas health and their potential protective effects following pancreatitis episodes in murine models (367). Their findings highlighted elevated production of activating cytokines like IFN- by pancreatic dendritic cells. ZNF366, encoding the transcription factor DC-SCRIPT, regulates cytokine levels such as IFN- and IL-12p40 in dendritic cells, suggesting its involvement in disease processes (367; 385). Thus, our approach facilitates insights into cell type-specific contributions to diseases, aiding in the formulation of new research hypotheses.

A key outcome of our study is the prediction of connections between non-coding genes and diseases (Supplementary File 3). For example: a) CDKN2B-AS1 and Pancreatic Ductal Carcinoma: Giaccherini et al. found that the rs1412832 polymorphism in the CDKN2B-AS1/ANRIL region significantly increases the risk of pancreatic ductal adenocarcinoma (386). b) ATXN8OS and drug response: Luo et al. demonstrated in glioma xenograft models that ATXN8OS enhances sensitivity to temozolomide (TMZ), a chemotherapy drug (387). Furthermore, our model predictions align with substantial evidence from the literature (Supplementary File 3).

As our study aimed to emphasize the role of transcription factors (TFs) and their

binding profiles in various cell types using an explainable model, we did not benchmark for accuracy comparison in disease-gene association prediction. Our analysis highlighted numerous genes and TFs potentially associated with various cancer types. For example: a) TF RELB from ChIP-seq data in GM12878 cells was linked to primary peritoneal carcinoma (Supplementary File 5). Haro et al. demonstrated that B cells in the peritoneal cavity produce IgM, which protects against tumor growth (388), while RELB is crucial for B cell development and maturation (389). b) TF CREB1 from ChIP-seq data in GM12878 cells was associated with the ontology class "Tumor Promotion" (Supplementary File 5). Yang et al. demonstrated that B cells promote tumor progression through angiogenesis in melanoma and lung carcinoma, while Frissora et al. showed that CREB-1 binding induces B cell proliferation (390; 391).

Our downstream analysis not only connects immune cell types with cancer but also generates numerous potential hypotheses. We validated many of these hypotheses regarding the association between top predictive TFs and cancer using survival analysis with clinical datasets from TCGA.

An important observation from our analysis of predictive TFs and diseases is the identification of distinct and overlapping roles of different cell types in disease contexts. Genome-wide association studies (GWAS) often do not capture the specific relevance of cell types. However, our approach integrates cell type context with TF binding profiles at promoters, offering significant potential for validating findings in relevant cell types. Identifying specific cell types can also greatly influence diagnostic, prognostic, and therapeutic strategies.

The ChIP-seq profiles and other datasets that are used as features, along with their source described in the Methods section, have an inherent drawback in that they cannot capture the complete ground truth of the epigenomic regulation of the cells as intended in our computational pipeline. Despite this, as discussed above, we can get good predictability in terms of sensitivity and specificity in a good fraction of ontological gene sets. This lower in good predictability for ontological gene sets is because of the lack of ChIP-seq profiles of multiple TFs from individual cell types in the publicly available databases. This leads to skewness in the good predictability towards ontological functions regulated by TFs and epigenomic features of cell types that are present in the dataset. To mitigate this bias, using single-cell ChIP-seq profiles of TFs and epigenomes

of different cell types is necessary. With the availability of such features in the near future, it will be possible to get good predictions for all types of functions.

Our study revealed that genes linked with apnea and targeted by top predictive TFs (Supplementary File 3) exhibit active expression and functional roles in erythroblast cells. The human body responds to low oxygen levels associated with apnea by increasing erythrocyte production, impacting sleep apnea prognosis. However, the causal relationship between gene activity and apnea remains unclear.

Furthermore, our findings highlight the significant role of diverse immune cell types in various disease contexts. Current gene annotations often encompass genes identified from bulk RNA expression analyses, including both tissue-specific cells and immune cells. Thus, a key insight from this analysis is on the nature of the current disease gene-sets, the annotation of the divergent cell type genes in the gene-sets indicates the presence of consequential genes in addition to causal genes for disease conditions.

Several examples suggest that the current gene-sets associated with diseases may feature genes active in immune cells, either dysregulated due to disease (effect) or actively involved as causal factors. Our analysis provides insights into both perspectives, underscoring broader challenges with current disease-associated gene annotations.

While genes impacted by a disorder can offer insights into diagnosis and prognosis, there is a critical need to refine annotations and disentangle the causative roles and effects among gene-set members currently linked to diseases.

CHAPTER 4

Conclusion

The advent of sequencing technology has given us a lot of information to understand the complexity of the human genome. In the thesis work, the novel computational framework that has been developed has contributed to the understanding of the functional elements of the human genome utilizing epigenomic and transcription factors as features in the context of biological functions and disease conditions. The entire work contributes to the ontological understand of the biology as the predictions from this computational work can be validated to expand the current gene-sets of the existing gene ontology.

4.1 Summary of contribution

4.1.1 Chapter 2. Epigenome and TF binding patterns are predictive of ontology-based functions of coding and non-coding genes

1. Proof-of-concept that the combination of epigenomic and transcription factor binding pattern at promoters of genes are predictive of gene function

From the work described in Chapter 2 of the thesis, it is established that epigenome and TF binding patterns at the promoters of the genes are predictive of the functions of the genes, both coding as well as non-coding genes. As explained in Chapter 2, the past gene-function association prediction methods suffer from not being able to accumulate the features of the non-coding genes for their reliable prediction. The novel approach developed in this thesis work uses epigenome and TF binding patterns and open-chromatin profiles to accommodate the features of the non-coding genes at the promoter level for their predictions into ontological gene-sets.

2. Derivation of latent clusters of biological and molecular functions

The downstream analysis using the top predictors of individual ontological gene sets revealed latent groups of biological processes or molecular functions that are coupled together temporally to play their role in a larger cellular process like cell cycle, immune response, etc. The classes of functions that are found could represent latent gene-sets, and can serve as a proxy for understanding the major cellular processes. This derived analysis is a conceptual advancement for identifying new gene-sets and for understanding the underlying genomic regulatory mechanisms of biological processes.

3. Utility of the computational framework

The entire framework is available as an R package, GFPredict, that takes in user-defined biologically related gene-sets as inputs to predict novel biologically similar genes. The R package, as demonstrated in Chapter 2, can be utilized to expand the existing gene-sets of CRISPR for various phenotypes.

We have created a repository of predictions for biologists to use the predictions to corroborate their research findings and develop novel hypotheses in their functional genomics research work.

A novel computational framework has been developed for the interpretable prediction of functions for coding and non-coding genes and their involvement in larger cellular processes.

4.1.2 Chapter 3. Explainable models using transcription factor binding and epigenome patterns at promoters reveal disease-associated genes and their regulators in the context of cell-types

1. Systematic genome-wide association of regulatory elements in various disease conditions

Many of the disease conditions arise because of the dysregulated expression of the genes. Many groups have developed computational methods to understand the role of epigenome and transcription factors in disease conditions in anecdotal studies as discussed in the chapter 3. In this thesis, a systematic analysis of genome-wide association of transcription factors and epigenome in disease conditions using disease gene-sets has been done.

2. Linking TFs to diseases in the context of cell-types and its implication to the current disease gene classification system

Through this analysis, we can link immune cells to disease conditions of origin from unrelated cell types. Such insights can help develop a diagnosis, prognosis, and potential therapeutic strategies for diseases, as discussed in Chapter 3.

One of the key inferences that can be drawn from linking divergent (immune) cell types to diseases is that the current disease ontological gene-sets contain causal and consequential genes. There is a need for systematic reclassification of disease gene-sets based on the type of association of genes.

Overall, the genome-wide association of epigenome and TFs in disease conditions and the identification of causal and consequential genes can help better understand the molecular regulatory mechanisms of the disease pathology by developing effective drug targets against causal genes themselves or their products. The consequential genes can be utilized constructively to track the prognosis of the patients condition.

4.2 Future work

The ongoing concluding work of this thesis is the experimental validation of the predicted disease-gene associations. The predicted genes in different cancer conditions will be validated to further increase the reliability in the predictions of the computational framework.

The results of this thesis work advocates of the development of hypothesis of utilizing the binding signal of TF and epigenomic marks at the enhancer regions to link them to different biological processes and molecular functions.

Also, the clustering of ontological gene-sets and their overall involvement in larger cellular processes proposes for a study of temporal convergence of biological processes and molecular functions in different cellular contexts.

REFERENCES

[1] **Consortium** (2001). Initial sequencing and analysis of the human genome. *Nature*, **409**(6822), 860–921.

[2] **Consortium** (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**(7414), 57–74.

[3] **Frankish, A.**, **M. Diekhans**, **I. Jungreis**, **J. Lagarde**, **J. E. Loveland**, **J. M. Mudge**, **C. Sisu**, **J. C. Wright**, **J. Armstrong**, **I. Barnes**, **A. Berry**, **A. Bignell**, **C. Boix**, **S. S. Carbonell**, **F. Cunningham**, **T. Di Domenico**, **S. Donaldson**, **I. T. Fiddes**, **G. C. García**, **J. M. Gonzalez**, **T. Grego**, **M. Hardy**, **T. Hourlier**, **K. L. Howe**, **T. Hunt**, **O. G. Izuogu**, **R. Johnson**, **F. J. Martin**, **L. Martínez**, **S. Mohanan**, **P. Muir**, **F. C. P. Navarro**, **A. Parker**, **B. Pei**, **F. Pozo**, **F. C. Riera**, **M. Ruffier**, **B. M. Schmitt**, **E. Stapleton**, **M. M. Suner**, **I. Sycheva**, **B. Uszczynska-Ratajczak**, **M. Y. Wolf**, **J. Xu**, **Y. T. Yang**, **A. Yates**, **D. Zerbino**, **Y. Zhang**, **J. S. Choudhary**, **M. Gerstein**, **R. Guigó**, **T. J. P. Hubbard**, **M. Kellis**, **B. Paten**, **M. L. Tress**, and **P. Flicek** (2021). GENCODE 2021. *Nucleic Acids Res.*, **49**(D1).

[4] **Tutar, Y.** (2012). Pseudogenes. *Comp. Funct. Genomics*, **2012**.

[5] **The Gene Ontology Consortium**, **M. Ashburner**, **C. A. Ball**, **J. A. Blake**, **D. Botstein**, **H. Butler**, **J. Michael Cherry**, **A. P. Davis**, **K. Dolinski**, **S. S. Dwight**, **J. T. Eppig**, **M. A. Harris**, **D. P. Hill**, **L. Issel-Tarver**, **A. Kasarskis**, **S. Lewis**, **J. C. Matese**, **J. E. Richardson**, **M. Ringwald**, **G. M. Rubin**, and **G. Sherlock** (2000). Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**(1), 25.

[6] **Balakrishnan, R.**, **M. A. Harris**, **R. Huntley**, **K. Van Auken**, and **J. Michael Cherry** (2013). A guide to best practices for gene ontology (GO) manual annotation. *Database*, **2013**.

[7] **Zhang, Y.**, **D. Bu**, **P. Huo**, **Z. Wang**, **H. Rong**, **Y. Li**, **J. Liu**, **M. Ye**, **Y. Wu**, **Z. Jiang**, **Q. Liao**, and **Y. Zhao** (2021). ncFANs v2.0: an integrative platform for functional annotation of non-coding RNAs. *Nucleic Acids Res.*, **49**(W1), W459–W468.

[8] **Gratten, J.** and **P. M. Visscher** (2016). Genetic pleiotropy in complex traits and diseases: implications for genomic medicine. *Genome Med.*, **8**(1), 1–3.

[9] **Lily, O.** (2004). Chronic autoimmune disease caused by somatic mutation to t-lymphocyte regulatory receptors. *Med. Hypotheses*, **62**(4), 582–586.

[10] **Greenman, C.**, **P. Stephens**, **R. Smith**, **G. L. Dalgliesh**, **C. Hunter**, **G. Bignell**, **H. Davies**, **J. Teague**, **A. Butler**, **C. Stevens**, **S. Edkins**, **S. O'Meara**, **I. Vastrik**, **E. E. Schmidt**, **T. Avis**, **S. Barthorpe**, **G. Bhamra**, **G. Buck**, **B. Choudhury**, **J. Clements**, **J. Cole**, **E. Dicks**, **S. Forbes**, **K. Gray**, **K. Halliday**, **R. Harrison**, **K. Hills**, **J. Hinton**, **A. Jenkinson**, **D. Jones**, **A. Menzies**, **T. Mironenko**,

J. Perry, **K. Raine**, **D. Richardson**, **R. Shepherd**, **A. Small**, **C. Tofts**, **J. Varian**, **T. Webb**, **S. West**, **S. Widaa**, **A. Yates**, **D. P. Cahill**, **D. N. Louis**, **P. Goldstraw**, **A. G. Nicholson**, **F. Brasseur**, **L. Looijenga**, **B. L. Weber**, **Y.-E. Chiew**, **A. deFazio**, **M. F. Greaves**, **A. R. Green**, **P. Campbell**, **E. Birney**, **D. F. Easton**, **G. Chenevix-Trench**, **M.-H. Tan**, **S. K. Khoo**, **B. T. Teh**, **S. T. Yuen**, **S. Y. Leung**, **R. Wooster**, **P. Andrew Futreal**, and **M. R. Stratton** (2007). Patterns of somatic mutation in human cancer genomes. *Nature*, **446**(7132), 153.

[11] **Markitantova, Y.** and **V. Simirskii** (2020). Inherited eye diseases with retinal manifestations through the eyes of homeobox genes. *Int. J. Mol. Sci.*, **21**(5).

[12] **Ali, O.** (2013). Genetics of type 2 diabetes. *World J. Diabetes*, **4**(4), 114.

[13] **Hu, H.-F.**, **Z. Ye**, **Y. Qin**, **X.-W. Xu**, **X.-J. Yu**, **Q.-F. Zhuo**, and **S.-R. Ji** (2021). Mutations in key driver genes of pancreatic cancer: molecularly targeted therapies and other clinical implications. *Acta Pharmacol. Sin.*, **42**(11), 1725–1741.

[14] **Venkat, S.**, **A. A. Alahmari**, and **M. E. Feigin** (2021). Drivers of gene expression dysregulation in pancreatic cancer. *Trends Cancer Res.*, **7**(7), 594.

[15] **Tonyan, Z. N.**, **Y. A. Nasykhova**, **M. M. Danilova**, **Y. A. Barbitoff**, **A. I. Changalidi**, **A. A. Mikhailova**, and **A. S. Glotov** (2022). Overview of transcriptomic research on type 2 diabetes: Challenges and perspectives. *Genes*, **13**(7), 1176.

[16] **Kanehisa, M.**, **M. Furumichi**, **M. Tanabe**, **Y. Sato**, and **K. Morishima** (2016). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.*, **45**(D1), D353–D361.

[17] **Grissa, D.**, **A. Junge**, **T. I. Oprea**, and **L. J. Jensen** (2022). Diseases 2.0: a weekly updated database of disease–gene associations from text mining and data integration. *Database*, **2022**.

[18] **Piñero, J.**, **N. Queralt-Rosinach**, **À. Bravo**, **J. Deu-Pons**, **A. Bauer-Mehren**, **M. Baron**, **F. Sanz**, and **L. I. Furlong** (2015). DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database*, **2015**.

[19] **Portin, P.** and **A. Wilkins** (2017). The evolving definition of the term "gene". *Genetics*, **205**(4), 1353.

[20] **Wright, B. W.**, **M. P. Molloy**, and **P. R. Jaschke** (2021). Overlapping genes in natural and engineered genomes. *Nat. Rev. Genet.*, **23**(3), 154–168.

[21] **Reyes, A.** and **W. Huber** (2017). Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues. *Nucleic Acids Res.*, **46**(2), 582–592.

[22] **Berk, A. J.** (2016). Discovery of RNA splicing and genes in pieces. *Proc. Natl. Acad. Sci. U. S. A.*, **113**(4), 801.

[23] **Arrey, G.**, **S. T. Keating**, and **B. Regenberg** (2022). A unifying model for extrachromosomal circular DNA load in eukaryotic cells. *Semin. Cell Dev. Biol.*, **128**.

[24] **Mattick, J. S.** (2005). The functional genomics of noncoding RNA. *Science*, **309**(5740).

[25] **Parra, G.**, **A. Reymond**, **N. Dabbouseh**, **E. T. Dermitzakis**, **R. Castelo**, **T. M. Thomson**, **S. E. Antonarakis**, and **R. Guigó** (2006). Tandem chimerism as a means to increase protein complexity in the human genome. *Genome Res.*, **16**(1).

[26] **Policastro, R. A.** and **G. E. Zentner** (2021). Global approaches for profiling transcription initiation. *Cell Reports Methods*, **1**(5).

[27] **Vacik, T.** and **I. Raska** (2017). Alternative intronic promoters in development and disease. *Protoplasma*, **254**(3), 1201–1206.

[28] **Maruyama, K.** and **S. Sugano** (1994). Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides. *Gene*, **138**(1-2).

[29] **Hashimoto, S.-I.**, **Y. Suzuki**, **Y. Kasai**, **K. Morohoshi**, **T. Yamada**, **J. Sese**, **S. Morishita**, **S. Sugano**, and **K. Matsushima** (2004). 5-end SAGE for the analysis of transcriptional start sites. *Nat. Biotechnol.*, **22**(9), 1146–1149.

[30] **Yamashita, R.**, **N. P. Sathira**, **A. Kanai**, **K. Tanimoto**, **T. Arauchi**, **Y. Tanaka**, **S. Hashimoto**, **S. Sugano**, **K. Nakai**, and **Y. Suzuki** (2011). Genome-wide characterization of transcriptional start sites in humans by integrative transcriptome analysis. *Genome Res.*, **21**(5).

[31] **Ni, T.**, **D. L. Corcoran**, **E. A. Rach**, **S. Song**, **E. P. Spana**, **Y. Gao**, **U. Ohler**, and **J. Zhu** (2010). A paired-end sequencing strategy to map the complex landscape of transcription initiation. *Nat. Methods*, **7**(7).

[32] **Fuchs, R. T.**, **Z. Sun**, **F. Zhuang**, and **G. B. Robb** (2015). Bias in ligation-based small RNA sequencing library construction is determined by adaptor and RNA structure. *PLoS One*, **10**(5).

[33] **Carninci, P.**, **C. Kvam**, **A. Kitamura**, **T. Ohsumi**, **Y. Okazaki**, **M. Itoh**, **M. Kamiya**, **K. Shibata**, **N. Sasaki**, **M. Izawa**, **M. Muramatsu**, **Y. Hayashizaki**, and **C. Schneider** (1996). High-efficiency full-length cDNA cloning by biotinylated CAP trapper. *Genomics*, **37**(3).

[34] **Shiraki, T.**, **S. Kondo**, **S. Katayama**, **K. Waki**, **T. Kasukawa**, **H. Kawaji**, **R. Kodzius**, **A. Watahiki**, **M. Nakamura**, **T. Arakawa**, **S. Fukuda**, **D. Sasaki**, **A. Podhajska**, **M. Harbers**, **J. Kawai**, **P. Carninci**, and **Y. Hayashizaki** (2003). Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci. U. S. A.*, **100**(26), 15776.

[35] **Gariglio, P.**, **M. Bellard**, and **P. Chambon** (1981). Clustering of RNA polymerase B molecules in the 5' moiety of the adult beta-globin gene of hen erythrocytes. *Nucleic Acids Res.*, **9**(11), 2589.

[36] **García-Martínez, J.** (2004). Genomic Run-On evaluates transcription rates for all yeast genes and identifies gene regulatory mechanisms. *Mol. Cell*, **15**(2), 303–313.

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

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

[39] **Smale, S. T.** and **J. T. Kadonaga** (2003). The RNA polymerase II core promoter. *Annu. Rev. Biochem.*, **72**, 449–479.

[40] **Lagrange, T.**, **A. N. Kapanidis**, **H. Tang**, **D. Reinberg**, and **R. H. Ebright** (1998). New core promoter element in RNA polymerase II-dependent transcription: sequence-specific DNA binding by transcription factor IIB. *Genes Dev.*, **12**(1), 34–44.

[41] **Lim, C. Y.**, **B. Santoso**, **T. Boulay**, **E. Dong**, **U. Ohler**, and **J. T. Kadonaga** (2004). The MTE, a new core promoter element for transcription by RNA polymerase II. *Genes Dev.*, **18**(13), 1606–1617.

[42] **Deng, W.** and **S. G. Roberts** (2005). A core promoter element downstream of the TATA box that is recognized by TFIIB. *Genes Dev.*, **19**(20).

[43] **Juven-Gershon, T.** and **J. T. Kadonaga** (2010). Regulation of gene expression via the core promoter and the basal transcriptional machinery. *Dev. Biol.*, **339**(2).

[44] **Kadonaga, J. T.** (2012). Perspectives on the RNA polymerase II core promoter. *Wiley Interdiscip. Rev. Dev. Biol.*, **1**(1).

[45] **Zabidi, M. A.**, **C. D. Arnold**, **K. Schernhuber**, **M. Pagani**, **M. Rath**, **O. Frank**, and **A. Stark** (2015). Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature*, **518**(7540), 556–559.

[46] **Lenhard, B.**, **A. Sandelin**, and **P. Carninci** (2012). Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nat. Rev. Genet.*, **13**(4).

[47] **Vo Ngoc L**, **C. J. Cassidy**, **C. Y. Huang**, **S. H. Duttke**, and **J. T. Kadonaga** (2017). The human initiator is a distinct and abundant element that is precisely positioned in focused core promoters. *Genes Dev.*, **31**(1).

[48] **Carninci, P.**, **A. Sandelin**, **B. Lenhard**, **S. Katayama**, **K. Shimokawa**, **J. Ponjavic**, **C. A. Semple**, **M. S. Taylor**, **P. G. Engström**, **M. C. Frith**, **A. R. Forrest**, **W. B. Alkema**, **S. L. Tan**, **C. Plessy**, **R. Kodzius**, **T. Ravasi**, **T. Kasukawa**, **S. Fukuda**, **M. Kanamori-Katayama**, **Y. Kitazume**, **H. Kawaji**, **C. Kai**, **M. Nakamura**, **H. Konno**, **K. Nakano**, **S. Mottagui-Tabar**, **P. Arner**, **A. Chesi**, **S. Gustincich**, **F. Persichetti**, **H. Suzuki**, **S. M. Grimmond**, **C. A. Wells**, **V. Orlando**, **C. Wahlestedt**, **E. T. Liu**, **M. Harbers**, **J. Kawai**, **V. B. Bajic**, **D. A. Hume**, and **Y. Hayashizaki** (2006). Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.*, **38**(6).

[49] **Forrest, A. R.**, **H. Kawaji**, **M. Rehli**, **J. K. Baillie**, **M. J. de Hoon**, **V. Haberle**, **T. Lassmann**, **I. V. Kulakovskiy**, **M. Lizio**, **M. Itoh**, **R. Andersson**, **C. J.**

Mungall, **T. F. Meehan**, **S. Schmeier**, **N. Bertin**, **M. Jørgensen**, **E. Dimont**, **E. Arner**, **C. Schmidl**, **U. Schaefer**, **Y. A. Medvedeva**, **C. Plessy**, **M. Vitezic**, **J. Severin**, **C. Semple**, **Y. Ishizu**, **R. S. Young**, **M. Francescatto**, **I. Alam**, **D. Albanese**, **G. M. Altschuler**, **T. Arakawa**, **J. A. Archer**, **P. Arner**, **M. Babina**, **S. Rennie**, **P. J. Balwierz**, **A. G. Beckhouse**, **S. Pradhan-Bhatt**, **J. A. Blake**, **A. Blumenthal**, **B. Bodega**, **A. Bonetti**, **J. Briggs**, **F. Brombacher**, **A. M. Burroughs**, **A. Califano**, **C. V. Cannistraci**, **D. Carbajo**, **Y. Chen**, **M. Chierici**, **Y. Ciani**, **H. C. Clevers**, **E. Dalla**, **C. A. Davis**, **M. Detmar**, **A. D. Diehl**, **T. Dohi**, **F. Drabløs**, **A. S. Edge**, **M. Edinger**, **K. Ekwall**, **M. Endoh**, **H. Enomoto**, **M. Fagiolini**, **L. Fairbairn**, **H. Fang**, **M. C. Farach-Carson**, **G. J. Faulkner**, **A. V. Favorov**, **M. E. Fisher**, **M. C. Frith**, **R. Fujita**, **S. Fukuda**, **C. Furlanello**, **M. Furino**, **J. Furusawa**, **T. B. Geijtenbeek**, **A. P. Gibson**, **T. Gingeras**, **D. Goldowitz**, **J. Gough**, **S. Guhl**, **R. Guler**, **S. Gustincich**, **T. J. Ha**, **M. Hamaguchi**, **M. Hara**, **M. Harbers**, **J. Harshbarger**, **A. Hasegawa**, **Y. Hasegawa**, **T. Hashimoto**, **M. Herlyn**, **K. J. Hitchens**, **H. S. Sj**, **O. M. Hofmann**, **I. Hoof**, **F. Hori**, **L. Huminiecki**, **K. Iida**, **T. Ikawa**, **B. R. Jankovic**, **H. Jia**, **A. Joshi**, **G. Jurman**, **B. Kaczkowski**, **C. Kai**, **K. Kaida**, **A. Kaiho**, **K. Kajiyama**, **M. Kanamori-Katayama**, **A. S. Kasianov**, **T. Kasukawa**, **S. Katayama**, **S. Kato**, **S. Kawaguchi**, **H. Kawamoto**, **Y. I. Kawamura**, **T. Kawashima**, **J. S. Kempfle**, **T. J. Kenna**, **J. Kere**, **L. M. Khachigian**, **T. Kitamura**, **S. P. Klinken**, **A. J. Knox**, **M. Kojima**, **S. Kojima**, **N. Kondo**, **H. Koseki**, **S. Koyasu**, **S. Krampitz**, **A. Kubosaki**, **A. T. Kwon**, **J. F. Laros**, **W. Lee**, **A. Lennartsson**, **K. Li**, **B. Lilje**, **L. Lipovich**, **A. Mackay-Sim**, **R. Manabe**, **J. C. Mar**, **B. Marchand**, **A. Mathelier**, **N. Mejhert**, **A. Meynert**, **Y. Mizuno**, **D. A. de Lima Morais**, **H. Morikawa**, **M. Morimoto**, **K. Moro**, **E. Motakis**, **H. Motohashi**, **C. L. Mummery**, **M. Murata**, **S. Nagao-Sato**, **Y. Nakachi**, **F. Nakahara**, **T. Nakamura**, **Y. Nakamura**, **K. Nakazato**, **E. van Nimwegen**, **N. Ninomiya**, **H. Nishiyori**, **S. Noma**, **S. Noma**, **T. Noazaki**, **S. Ogishima**, **N. Ohkura**, **H. Ohimiya**, **H. Ohno**, **M. Ohshima**, **M. Okada-Hatakeyama**, **Y. Okazaki**, **V. Orlando**, **D. A. Ovchinnikov**, **A. Pain**, **R. Passier**, **M. Patrikakis**, **H. Persson**, **S. Piazza**, **J. G. Prendergast**, **O. J. Rackham**, **J. A. Ramilowski**, **M. Rashid**, **T. Ravasi**, **P. Rizzu**, **M. Roncador**, **S. Roy**, **M. B. Rye**, **E. Saijyo**, **A. Sajantila**, **A. Saka**, **S. Sakaguchi**, **M. Sakai**, **H. Sato**, **S. Savvi**, **A. Saxena**, **C. Schneider**, **E. A. Schultes**, **G. G. Schulze-Tanzil**, **A. Schwegmann**, **T. Sengstag**, **G. Sheng**, **H. Shimoji**, **Y. Shimoni**, **J. W. Shin**, **C. Simon**, **D. Sugiyama**, **T. Sugiyama**, **M. Suzuki**, **N. Suzuki**, **R. K. Swoboda**, **t. H. Pa**, **M. Tagami**, **N. Takahashi**, **J. Takai**, **H. Tanaka**, **H. Tatsukawa**, **Z. Tatum**, **M. Thompson**, **H. Toyodo**, **T. Toyoda**, **E. Valen**, **M. van de Wetering**, **L. M. van den Berg**, **R. Verado**, **D. Vijayan**, **I. E. Vorontsov**, **W. W. Wasserman**, **S. Watanabe**, **C. A. Wells**, **L. N. Winteringham**, **E. Wolvetang**, **E. J. Wood**, **Y. Yamaguchi**, **M. Yamamoto**, **M. Yoneda**, **Y. Yonekura**, **S. Yoshida**, **S. E. Zabierowski**, **P. G. Zhang**, **X. Zhao**, **S. Zucchelli**, **K. M. Summers**, **H. Suzuki**, **C. O. Daub**, **J. Kawai**, **P. Heutink**, **W. Hide**, **T. C. Freeman**, **B. Lenhard**, **V. B. Bajic**, **M. S. Taylor**, **V. J. Makeev**, **A. Sandelin**, **D. A. Hume**, **P. Carninci**, and **Y. Hayashizaki** (2014). A promoter-level mammalian expression atlas. *Nature*, **507**(7493).

[50] **Bernstein, B. E.**, **T. S. Mikkelsen**, **X. Xie**, **M. Kamal**, **D. J. Huebert**, **J. Cuff**, **B. Fry**, **A. Meissner**, **M. Wernig**, **K. Plath**, **R. Jaenisch**, **A. Wagschal**, **R. Feil**,

S. L. Schreiber, and **E. S. Lander** (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell*, **125**(2).

[51] **Saxonov, S.**, **P. Berg**, and **D. L. Brutlag** (2006). A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc. Natl. Acad. Sci. U. S. A.*, **103**(5), 1412.

[52] **Hughes, A. L.**, **J. R. Kelley**, and **R. J. Klose** (2020). Understanding the interplay between CpG island-associated gene promoters and H3K4 methylation. *Biochim. Biophys. Acta Gene Regul. Mech.*, **1863**(8).

[53] **Lettice, L. A.**, **S. J. Heaney**, **L. A. Purdie**, **L. Li**, **P. de Beer**, **B. A. Oostra**, **D. Goode**, **G. Elgar**, **R. E. Hill**, and **E. de Graaff** (2003). A long-range shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum. Mol. Genet.*, **12**(14).

[54] **Banerji, J.**, **S. Rusconi**, and **W. Schaffner** (1981). Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell*, **27**(2 Pt 1).

[55] **Shlyueva, D.**, **G. Stampfel**, and **A. Stark** (2014). Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.*, **15**(4).

[56] **Koch, F.**, **R. Fenouil**, **M. Gut**, **P. Cauchy**, **T. K. Albert**, **J. Zacarias-Cabeza**, **S. Spicuglia**, **A. L. de la Chapelle**, **M. Heidemann**, **C. Hintermair**, **D. Eick**, **I. Gut**, **P. Ferrier**, and **J. C. Andrau** (2011). Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters. *Nat. Struct. Mol. Biol.*, **18**(8).

[57] **De Santa, F.**, **I. Barozzi**, **F. Mietton**, **S. Ghisletti**, **S. Polletti**, **B. K. Tusi**, **H. Muller**, **J. Ragoussis**, **C.-L. Wei**, and **G. Natoli** (2010). A large fraction of extragenic RNA pol II transcription sites overlap enhancers. *PLoS Biol.*, **8**(5).

[58] **Kim, T.-K.**, **M. Hemberg**, **J. M. Gray**, **A. M. Costa**, **D. M. Bear**, **J. Wu**, **D. A. Harmin**, **M. Laptewicz**, **K. Barbara-Haley**, **S. Kuersten**, **E. Markenscoff-Papadimitriou**, **D. Kuhl**, **H. Bito**, **P. F. Worley**, **G. Kreiman**, and **M. E. Greenberg** (2010). Widespread transcription at neuronal activity-regulated enhancers. *Nature*, **465**(7295), 182.

[59] **Li, W.**, **D. Notani**, and **M. G. Rosenfeld** (2016). Enhancers as non-coding RNA transcription units: recent insights and future perspectives. *Nat. Rev. Genet.*, **17**(4), 207–223.

[60] **Li, W.**, **D. Notani**, **Q. Ma**, **B. Tanasa**, **E. Nunez**, **A. Y. Chen**, **D. Merkurjev**, **J. Zhang**, **K. Ohgi**, **X. Song**, **S. Oh**, **H.-S. Kim**, **C. K. Glass**, and **M. G. Rosenfeld** (2013). Functional importance of eRNAs for estrogen-dependent transcriptional activation events. *Nature*, **498**(7455), 516.

[61] **Ørom, U. A.**, **T. Derrien**, **M. Beringer**, **K. Gumireddy**, **A. Gardini**, **G. Bussotti**, **F. Lai**, **M. Zytnicki**, **C. Notredame**, **Q. Huang**, **R. Guigo**, and **R. Shiekhattar** (2010). Long noncoding RNAs with enhancer-like function in human cells. *Cell*, **143**(1).

[62] **Schaukowitch, K.**, **J. Y. Joo**, **X. Liu**, **J. K. Watts**, **C. Martinez**, and **T. K. Kim** (2014). Enhancer RNA facilitates NELF release from immediate early genes. *Mol. Cell*, **56**(1).

[63] **Nolis, I. K.**, **D. J. McKay**, **E. Mantouvalou**, **S. Lomvardas**, **M. Merika**, and **D. Thanos** (2009). Transcription factors mediate long-range enhancer–promoter interactions. *Proceedings of the National Academy of Sciences*, **106**(48), 20222–20227.

[64] **Deng, W.**, **J. Lee**, **H. Wang**, **J. Miller**, **A. Reik**, **P. D. Gregory**, **A. Dean**, and **G. A. Blobel** (2012). Controlling long range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell*, **149**(6), 1233.

[65] **Alvarez-Dominguez, J. R.**, **M. Knoll**, **A. A. Gromatzky**, and **H. F. Lodish** (2017). The Super-Enhancer-Derived alncRNA-EC7/ bloodlinc potentiates red blood cell development in trans. *Cell Rep.*, **19**(12), 2503.

[66] **Arnold, P. R.**, **A. D. Wells**, and **X. C. Li** (2019). Diversity and emerging roles of enhancer RNA in regulation of gene expression and cell fate. *Frontiers in Cell and Developmental Biology*, **7**.

[67] **Tang, F.**, **Z. Yang**, **Y. Tan**, and **Y. Li** (2020). Super-enhancer function and its application in cancer targeted therapy. *npj Precision Oncology*, **4**(1), 1–7.

[68] **Dean, A.**, **D. R. Larson**, and **V. Sartorelli** (2021). Enhancers, gene regulation, and genome organization. *Genes Dev.*, **35**(7-8), 427–432.

[69] **Visel, A.**, **M. J. Blow**, **Z. Li**, **T. Zhang**, **J. A. Akiyama**, **A. Holt**, **I. Plajzer-Frick**, **M. Shoukry**, **C. Wright**, **F. Chen**, **V. Afzal**, **B. Ren**, **E. M. Rubin**, and **L. A. Pennacchio** (2009). ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, **457**(7231), 854.

[70] **Hartley, K. O.**, **D. Gell**, **G. C. Smith**, **H. Zhang**, **N. Divecha**, **M. A. Connelly**, **A. Admon**, **S. P. Lees-Miller**, **C. W. Anderson**, and **S. P. Jackson** (1995). DNA-dependent protein kinase catalytic subunit: a relative of phosphatidylinositol 3-kinase and the ataxia telangiectasia gene product. *Cell*, **82**(5).

[71] **Chepelev, I.**, **G. Wei**, **D. Wangsa**, **Q. Tang**, and **K. Zhao** (2012). Characterization of genome-wide enhancer-promoter interactions reveals co-expression of interacting genes and modes of higher order chromatin organization. *Cell Res.*, **22**(3), 490.

[72] **Djebali, S.**, **C. A. Davis**, **A. Merkel**, **A. Dobin**, **T. Lassmann**, **A. M. Mortazavi**, **A. Tanzer**, **J. Lagarde**, **W. Lin**, **F. Schlesinger**, **C. Xue**, **G. K. Marinov**, **J. Khatun**, **B. A. Williams**, **C. Zaleski**, **J. Rozowsky**, **M. Röder**, **F. Kokocinski**, **R. F. Abdelhamid**, **T. Alioto**, **I. Antoshechkin**, **M. T. Baer**, **N. S. Bar**, **P. Batut**, **K. Bell**, **I. Bell**, **S. Chakrabortty**, **X. Chen**, **J. Chrast**, **J. Curado**, **T. Derrien**, **J. Drenkow**, **E. Dumais**, **J. Dumais**, **R. Duttagupta**, **E. Falconnet**, **M. Fastuca**, **K. Fejes-Toth**, **P. Ferreira**, **S. Foissac**, **M. J. Fullwood**, **H. Gao**, **D. Gonzalez**, **A. Gordon**, **H. Gunawardena**, **C. Howald**, **S. Jha**, **R. Johnson**, **P. Kapranov**, **B. King**, **C. Kingswood**, **O. J. Luo**, **E. Park**, **K. Persaud**, **J. B. Preall**, **P. Ribeca**, **B. Risk**, **D. Robyr**, **M. Sammeth**, **L. Schaffer**, **L.-H. See**, **A. Shahab**, **J. Skancke**, **A. M. Suzuki**, **H. Takahashi**, **H. Tilgner**, **D. Trout**,

N. Walters, **H. Wang**, **J. Wrobel**, **Y. Yu**, **X. Ruan**, **Y. Hayashizaki**, **J. Harrow**, **M. Gerstein**, **T. Hubbard**, **A. Reymond**, **S. E. Antonarakis**, **G. Hannon**, **M. C. Giddings**, **Y. Ruan**, **B. Wold**, **P. Carninci**, **R. Guigó**, and **T. R. Gingeras** (2012). Landscape of transcription in human cells. *Nature*, **489**(7414), 101.

[73] **Kleinjan, D. A.** and **V. van Heyningen** (2005). Long-range control of gene expression: emerging mechanisms and disruption in disease. *Am. J. Hum. Genet.*, **76**(1), 8–32.

[74] **Peng, Y.** and **Y. Zhang** (2018). Enhancer and super-enhancer: Positive regulators in gene transcription. *Animal models and experimental medicine*, **1**(3).

[75] **Fulco, C. P.**, **M. Munschauer**, **R. Anyoha**, **G. Munson**, **S. R. Grossman**, **E. M. Perez**, **M. Kane**, **B. Cleary**, **E. S. Lander**, and **J. M. Engreitz** (2016). Systematic mapping of functional enhancer–promoter connections with CRISPR interference. *Science*.

[76] **Belton, J. M.**, **R. P. McCord**, **J. H. Gibcus**, **N. Naumova**, **Y. Zhan**, and **J. Dekker** (2012). Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods*, **58**(3).

[77] **et al., P. M.** (2023). Integrative approaches to study enhancer–promoter communication. *Curr. Opin. Genet. Dev.*, **80**, 102052.

[78] **Ogbourne, S.** and **T. M. Antalis** (1998). Transcriptional control and the role of silencers in transcriptional regulation in eukaryotes. *Biochem. J.*, **331 (Pt 1)**(Pt 1).

[79] **Qi, H.**, **M. Liu**, **D. W. Emery**, and **G. Stamatoyannopoulos** (2015). Functional validation of a constitutive autonomous silencer element. *PLoS One*, **10**(4), e0124588.

[80] **Ellmeier, W.**, **S. Sawada**, and **D. R. Littman** (1999). The regulation of CD4 and CD8 coreceptor gene expression during T cell development. *Annu. Rev. Immunol.*, **17**.

[81] **Huang, D.**, **H. M. Petrykowska**, **B. F. Miller**, **L. Elnitski**, and **I. Ovcharenko** (2019). Identification of human silencers by correlating cross-tissue epigenetic profiles and gene expression. *Genome Res.*, **29**(4).

[82] **Pang, B.** and **M. P. Snyder** (2020). Systematic identification of silencers in human cells. *Nat. Genet.*, **52**(3).

[83] **Nishioka, K.**, **J. C. Rice**, **K. Sarma**, **H. Erdjument-Bromage**, **J. Werner**, **Y. Wang**, **S. Chuikov**, **P. Valenzuela**, **P. Tempst**, **R. Steward**, **J. T. Lis**, **C. D. Allis**, and **D. Reinberg** (2002). PR-Set7 is a nucleosome-specific methyltransferase that modifies lysine 20 of histone H4 and is associated with silent chromatin. *Mol. Cell*, **9**(6).

[84] **Ngan, C. Y.**, **C. H. Wong**, **H. Tjong**, **W. Wang**, **R. L. Goldfeder**, **C. Choi**, **H. He**, **L. Gong**, **J. Lin**, **B. Urban**, **J. Chow**, **M. Li**, **J. Lim**, **V. Philip**, **S. A. Murray**, **H. Wang**, and **C. L. Wei** (2020). Chromatin interaction analyses elucidate the roles of PRC2-bound silencers in mouse development. *Nat. Genet.*, **52**(3).

[85] **Doni, J. N.**, **A. Jajodia**, **A. Mishra**, and **R. D. Hawkins** (2020). Candidate silencer elements for the human and mouse genomes. *Nat. Commun.*, **11**(1).

[86] **Brand, A. H.**, **L. Breeden**, **J. Abraham**, **R. Sternglanz**, and **K. Nasmyth** (1985). Characterization of a "silencer" in yeast: a DNA sequence with properties opposite to those of a transcriptional enhancer. *Cell*, **41**(1).

[87] **Brand, A. H.**, **G. Micklem**, and **K. Nasmyth** (1987). A yeast silencer contains sequences that can promote autonomous plasmid replication and transcriptional activation. *Cell*, **51**(5).

[88] **Gisselbrecht, S. S.**, **A. Palagi**, **J. V. Kurland**, **J. M. Rogers**, **H. Ozadam**, **Y. Zhan**, **J. Dekker**, and **M. L. Bulyk** (2020). Transcriptional silencers in drosophila serve a dual role as transcriptional enhancers in alternate cellular contexts. *Mol. Cell*, **77**(2).

[89] **He, J.**, **X. Huo**, **G. Pei**, **Z. Jia**, **Y. Yan**, **J. Yu**, **H. Qu**, **Y. Xie**, **J. Yuan**, **Y. Zheng**, **Y. Hu**, **M. Shi**, **K. You**, **T. Li**, **T. Ma**, **M. Q. Zhang**, **S. Ding**, **P. Li**, and **Y. Li** (2024). Dual-role transcription factors stabilize intermediate expression levels. *Cell*.

[90] **Segert, J. A.**, **S. S. Gisselbrecht**, and **M. L. Bulyk** (2021). Transcriptional silencers: Driving gene expression with the brakes on. *Trends Genet.*, **37**(6), 514–527.

[91] **Cusanovich, D. A.**, **A. J. Hill**, **D. Aghamirzaie**, **R. M. Daza**, **H. A. Pliner**, **J. B. Berletch**, **G. N. Filippova**, **X. Huang**, **L. Christiansen**, **W. S. DeWitt**, **C. Lee**, **S. G. Regalado**, **D. F. Read**, **F. J. Steemers**, **C. M. Disteche**, **C. Trapnell**, and **J. Shendure** (2018). A Single-Cell atlas of in vivo mammalian chromatin accessibility. *Cell*, **174**(5).

[92] **Lobanenkov, V. V.**, **R. H. Nicolas**, **V. V. Adler**, **H. Paterson**, **E. M. Klenova**, **A. V. Polotskaja**, and **G. H. Goodwin** (1990). A novel sequence-specific DNA binding protein which interacts with three regularly spaced direct repeats of the CCCTC-motif in the 5'-flanking sequence of the chicken c-myc gene. *Oncogene*, **5**(12), 1743–1753.

[93] **Kellum, R.** and **P. Schedl** (1992). A group of scs elements function as domain boundaries in an enhancer-blocking assay. *Mol. Cell. Biol.*, **12**(5).

[94] **Huang, S.**, **X. Li**, **T. M. Yusufzai**, **Y. Qiu**, and **G. Felsenfeld** (2007). USF1 recruits histone modification complexes and is critical for maintenance of a chromatin barrier. *Mol. Cell. Biol.*, **27**(22).

[95] **Dhillon, N.**, **J. Raab**, **J. Guzzo**, **S. J. Szyjka**, **S. Gangadharan**, **O. M. Aparicio**, **B. Andrews**, and **R. T. Kamakaka** (2009). DNA polymerase epsilon, acetylases and remodellers cooperate to form a specialized chromatin structure at a tRNA insulator. *EMBO J.*, **28**(17).

[96] **Chung, J. H.**, **M. Whiteley**, and **G. Felsenfeld** (1993). A 5' element of the chicken beta-globin domain serves as an insulator in human erythroid cells and protects against position effect in drosophila. *Cell*, **74**(3).

[97] **Bell, A. C.**, **A. G. West**, and **G. Felsenfeld** (1999). The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. *Cell*, **98**(3).

[98] **Parelho, V.**, **S. Hadjur**, **M. Spivakov**, **M. Leleu**, **S. Sauer**, **H. C. Gregson**, **A. Jarmuz**, **C. Canzonetta**, **Z. Webster**, **T. Nesterova**, **B. S. Cobb**, **K. Yokomori**, **N. Dillon**, **L. Aragon**, **A. G. Fisher**, and **M. Merkenschlager** (2008). Cohesins functionally associate with CTCF on mammalian chromosome arms. *Cell*, **132**(3).

[99] **Rubio, E. D.**, **D. J. Reiss**, **P. L. Welcsh**, **C. M. Disteche**, **G. N. Filippova**, **N. S. Baliga**, **R. Aebersold**, **J. A. Ranish**, and **A. Krumm** (2008). CTCF physically links cohesin to chromatin. *Proc. Natl. Acad. Sci. U. S. A.*, **105**(24).

[100] **Mishiro, T.**, **K. Ishihara**, **S. Hino**, **S. Tsutsumi**, **H. Aburatani**, **K. Shirahige**, **Y. Kinoshita**, and **M. Nakao** (2009). Architectural roles of multiple chromatin insulators at the human apolipoprotein gene cluster. *EMBO J.*, **28**(9).

[101] **Hou, C.**, **R. Dale**, and **A. Dean** (2010). Cell type specificity of chromatin organization mediated by CTCF and cohesin. *Proc. Natl. Acad. Sci. U. S. A.*, **107**(8).

[102] **Ishihara, K.**, **M. Oshimura**, and **M. Nakao** (2006). CTCF-dependent chromatin insulator is linked to epigenetic remodeling. *Mol. Cell*, **23**(5).

[103] **Chernukhin, I. V.**, **S. Shamsuddin**, **A. F. Robinson**, **A. F. Carne**, **A. Paul**, **A. I. El-Kady**, **V. V. Lobanenkov**, and **E. M. Klenova** (2000). Physical and functional interaction between two pluripotent proteins, the y-box DNA/RNA-binding factor, YB-1, and the multivalent zinc finger factor, CTCF. *J. Biol. Chem.*, **275**(38).

[104] **Kim, J.**, **A. Kollhoff**, **A. Bergmann**, and **L. Stubbs** (2003). Methylation-sensitive binding of transcription factor YY1 to an insulator sequence within the paternally expressed imprinted gene, peg3. *Hum. Mol. Genet.*, **12**(3).

[105] **Defossez, P. A.**, **K. F. Kelly**, **G. J. Filion**, **R. Pérez-Torrado**, **F. Magdinier**, **H. Menoni**, **C. L. Nordgaard**, **J. M. Daniel**, and **E. Gilson** (2005). The human enhancer blocker CTC-binding factor interacts with the transcription factor kaiso. *J. Biol. Chem.*, **280**(52).

[106] **Donohoe, M. E.**, **S. S. Silva**, **S. F. Pinter**, **N. Xu**, and **J. T. Lee** (2009). The pluripotency factor oct4 interacts with ctcf and also controls x-chromosome pairing and counting. *Nature*, **460**(7251).

[107] **Yusufzai, T. M.**, **H. Tagami**, **Y. Nakatani**, and **G. Felsenfeld** (2004). CTCF tethers an insulator to subnuclear sites, suggesting shared insulator mechanisms across species. *Mol. Cell*, **13**(2).

[108] **Guelen, L.**, **L. Pagie**, **E. Brasset**, **W. Meuleman**, **M. B. Faza**, **W. Talhout**, **B. H. Eussen**, **A. de Klein**, **L. Wessels**, **W. de Laat**, and **B. van Steensel** (2008). Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature*, **453**(7197).

[109] **Filippova, G. N.**, **M. K. Cheng**, **J. M. Moore**, **J. P. Truong**, **Y. J. Hu**, **D. K. Nguyen**, **K. D. Tsuchiya**, and **C. M. Disteche** (2005). Boundaries between chromosomal domains of X inactivation and escape bind CTCF and lack CpG methylation during early development. *Dev. Cell*, **8**(1).

[110] **Barski, A.**, **S. Cuddapah**, **K. Cui**, **T.-Y. Roh**, **D. E. Schones**, **Z. Wang**, **G. Wei**, **I. Chepelev**, and **K. Zhao** (2007). High-resolution profiling of histone methylations in the human genome. *Cell*, **129**(4), 823–837.

[111] **Xie, X.**, **T. S. Mikkelsen**, **A. Gnirke**, **K. Lindblad-Toh**, **M. Kellis**, and **E. S. Lander** (2007). Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. *Proceedings of the National Academy of Sciences*, **104**(17), 7145–7150.

[112] **Cuddapah, S.**, **R. Jothi**, **D. E. Schones**, **T.-Y. Roh**, **K. Cui**, and **K. Zhao** (2009). Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Res.*, **19**(1), 24–32.

[113] **Ferraiuolo, M. A.**, **M. Rousseau**, **C. Miyamoto**, **S. Shenker**, **X. Q. Wang**, **M. Nadler**, **M. Blanchette**, and **J. Dostie** (2010). The three-dimensional architecture of hox cluster silencing. *Nucleic Acids Res.*, **38**(21).

[114] **Bell, A. C.** and **G. Felsenfeld** (2000). Methylation of a CTCF-dependent boundary controls imprinted expression of the igf2 gene. *Nature*, **405**(6785).

[115] **Yang, J.** and **V. G. Corces** (2011). Chromatin insulators: a role in nuclear organization and gene expression. *Adv. Cancer Res.*, **110**, 43–76.

[116] **Ribeiro-Dos-Santos, A. M.**, **M. S. Hogan**, **R. D. Luther**, **R. Brosh**, and **M. T. Maurano** (2022). Genomic context sensitivity of insulator function. *Genome Res.*, **32**(3), 425–436.

[117] **Damante, G.**, **D. Fabbro**, **L. Pellizzari**, **D. Civitareale**, **S. Guazzi**, **M. Polycarpou-Schwartz**, **S. Cauci**, **F. Quadrifoglio**, **S. Formisano**, and **R. Di Lauro** (1994). Sequence-specific DNA recognition by the thyroid transcription factor-1 homeodomain. *Nucleic Acids Res.*, **22**(15), 3075–3083.

[118] **Pratt, H. E.**, **G. R. Andrews**, **N. Phalke**, **M. J. Purcaro**, **A. van der Velde**, **J. E. Moore**, and **Z. Weng** (2022). Factorbook: an updated catalog of transcription factor motifs and candidate regulatory motif sites. *Nucleic Acids Res.*, **50**(D1).

[119] **Wingender, E.** (1988). Compilation of transcription regulating proteins. *Nucleic Acids Res.*, **16**(5), 1879–1902.

[120] **Varga-Weisz, P. D.** and **P. B. Becker** (1995). Transcription factor-mediated chromatin remodelling: mechanisms and models. *FEBS Lett.*, **369**(1), 118–121.

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

[122] **Wunderlich, Z.** and **L. A. Mirny** (2009). Different gene regulation strategies revealed by analysis of binding motifs. *Trends Genet.*, **25**(10), 434–440.

[123] **Sönmezer, C.**, **R. Kleinendorst**, **D. Imanci**, **G. Barzaghi**, **L. Villacorta**, **D. Schübeler**, **V. Benes**, **N. Molina**, and **A. R. Krebs** (2021). Molecular co-occupancy identifies transcription factor binding cooperativity in vivo. *Mol. Cell*, **81**(2).

[124] **Jones, S.** (2004). An overview of the basic helix-loop-helix proteins. *Genome Biol.*, **5**(6), 226.

[125] **Severne, Y.**, **S. Wieland**, **W. Schaffner**, and **S. Rusconi** (1988). Metal binding 'finger' structures in the glucocorticoid receptor defined by site-directed mutagenesis. *EMBO J.*, **7**(8), 2503–2508.

[126] **Wingender, E.**, **T. Schoeps**, **M. Haubrock**, **M. Krull**, and **J. Dönitz** (2018). TFClass: expanding the classification of human transcription factors to their mammalian orthologs. *Nucleic Acids Res.*, **46**(D1), D343–D347.

[127] **Reece-Hoyes, J. S.** and **A. J. Marian Walhout** (2012). Yeast one-hybrid assays: a historical and technical perspective. *Methods*, **57**(4), 441–447.

[128] **Tacheny, A.**, **M. Dieu**, **T. Arnould**, and **P. Renard** (2013). Mass spectrometry-based identification of proteins interacting with nucleic acids. *J. Proteomics*, **94**, 89–109.

[129] **Letunic, I.**, **T. Doerks**, and **P. Bork** (2015). SMART: recent updates, new developments and status in 2015. *Nucleic Acids Res.*, **43**(Database issue), D257–60.

[130] **Finn, R. D.**, **P. Coggill**, **R. Y. Eberhardt**, **S. R. Eddy**, **J. Mistry**, **A. L. Mitchell**, **S. C. Potter**, **M. Punta**, **M. Qureshi**, **A. Sangrador-Vegas**, **G. A. Salazar**, **J. Tate**, and **A. Bateman** (2016). The pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.*, **44**(D1), D279–85.

[131] **Finn, R. D.**, **T. K. Attwood**, **P. C. Babbitt**, **A. Bateman**, **P. Bork**, **A. J. Bridge**, **H.-Y. Chang**, **Z. Dosztányi**, **S. El-Gebali**, **M. Fraser**, **J. Gough**, **D. Haft**, **G. L. Holliday**, **H. Huang**, **X. Huang**, **I. Letunic**, **R. Lopez**, **S. Lu**, **A. Marchler-Bauer**, **H. Mi**, **J. Mistry**, **D. A. Natale**, **M. Necci**, **G. Nuka**, **C. A. Orengo**, **Y. Park**, **S. Pesseat**, **D. Piovesan**, **S. C. Potter**, **N. D. Rawlings**, **N. Redaschi**, **L. Richardson**, **C. Rivoire**, **A. Sangrador-Vegas**, **C. Sigrist**, **I. Sillitoe**, **B. Smithers**, **S. Squizzato**, **G. Sutton**, **N. Thanki**, **P. D. Thomas**, **S. C. E. Tosatto**, **C. H. Wu**, **I. Xenarios**, **L.-S. Yeh**, **S.-Y. Young**, and **A. L. Mitchell** (2017). InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Res.*, **45**(D1), D190–D199.

[132] **Johnson, D. S.**, **A. Mortazavi**, **R. M. Myers**, and **B. Wold** (2007). Genome-Wide mapping of in vivo Protein-DNA interactions. *Science*.

[133] **Stormo, G. D.** and **Y. Zhao** (2010). Determining the specificity of protein–DNA interactions. *Nat. Rev. Genet.*, **11**(11), 751–760.

[134] **Rockel, S.**, **M. Geertz**, and **S. J. Maerkl** (2012). MITOMI: a microfluidic platform for in vitro characterization of transcription factor-DNA interaction. *Methods Mol. Biol.*, **786**.

[135] **Stallcup, M. R.** and **C. Poulard** (2020). Gene-Specific actions of transcriptional coregulators facilitate physiological plasticity: Evidence for a physiological coregulator code. *Trends Biochem. Sci.*, **45**(6), 497–510.

[136] **Rosenfeld, M. G.**, **V. V. Lunyak**, and **C. K. Glass** (2006). Sensors and signals: a coactivator/corepressor/epigenetic code for integrating signal-dependent programs of transcriptional response. *Genes Dev.*, **20**(11), 1405–1428.

[137] **Eberharter, A.** and **P. B. Becker** (2002). Histone acetylation: a switch between repressive and permissive chromatin. second in review series on chromatin dynamics. *EMBO Rep.*, **3**(3), 224–229.

[138] **Nagy, Z.** and **L. Tora** (2007). Distinct GCN5/PCAF-containing complexes function as co-activators and are involved in transcription factor and global histone acetylation. *Oncogene*, **26**(37), 5341–5357.

[139] **Delcuve, G. P.**, **D. H. Khan**, and **J. R. Davie** (2012). Roles of histone deacetylases in epigenetic regulation: emerging paradigms from studies with inhibitors. *Clin. Epigenetics*, **4**(1), 5.

[140] **Hervouet, E.**, **P. Peixoto**, **R. Delage-Mourroux**, **M. Boyer-Guittaut**, and **P.-F. Cartron** (2018). Specific or not specific recruitment of DNMTs for DNA methylation, an epigenetic dilemma. *Clin. Epigenetics*, **10**(1), 1–18.

[141] **Talukdar, P. D.** and **U. Chatterji** (2023). Transcriptional co-activators: emerging roles in signaling pathways and potential therapeutic targets for diseases. *Signal Transduction and Targeted Therapy*, **8**(1), 1–41.

[142] **Jolma, A.**, **Y. Yin**, **K. R. Nitta**, **K. Dave**, **A. Popov**, **M. Taipale**, **M. Enge**, **T. Kivioja**, **E. Morgunova**, and **J. Taipale** (2015). DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature*, **527**(7578), 384–388.

[143] **Reiter, F.**, **S. Wienerroither**, and **A. Stark** (2017). Combinatorial function of transcription factors and cofactors. *Curr. Opin. Genet. Dev.*, **43**.

[144] **Jaenisch, R.** and **A. Bird** (2003). Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat. Genet.*, **33 Suppl**.

[145] **Takai, D.** and **P. A. Jones** (2002). Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proceedings of the National Academy of Sciences*, **99**(6), 3740–3745.

[146] **Patil, V. S.**, **R. Zhou**, and **T. M. Rana** (2014). Gene regulation by non-coding RNAs. *Crit. Rev. Biochem. Mol. Biol.*, **49**(1), 16–32.

[147] **Lee, J.-H.**, **Y. Saito**, **S.-J. Park**, and **K. Nakai** (2020). Existence and possible roles of independent non-CpG methylation in the mammalian brain. *DNA Res.*, **27**(4), dsaa020.

[148] **Busslinger, M.**, **J. Hurst**, and **R. A. Flavell** (1983). DNA methylation and the regulation of globin gene expression. *Cell*, **34**(1), 197–206.

[149] **Bird, A.** (2002). DNA methylation patterns and epigenetic memory. *Genes Dev.*, **16**(1), 6–21.

[150] **Schulz, W. A.**, **C. Steinhoff**, and **A. R. Florl** (2006). Methylation of endogenous human retroelements in health and disease. *Curr. Top. Microbiol. Immunol.*, **310**, 211–250.

[151] **Silva, S. S.**, **R. K. Rowntree**, **S. Mekhoubad**, and **J. T. Lee** (2008). X-chromosome inactivation and epigenetic fluidity in human embryonic stem cells. *Proceedings of the National Academy of Sciences*, **105**(12), 4820–4825.

[152] **Pervjakova, N.**, **S. Kasela**, **A. P. Morris**, **M. Kals**, **A. Metspalu**, **C. M. Lindgren**, **A. Salumets**, and **R. Mägi** (2016). Imprinted genes and imprinting control regions show predominant intermediate methylation in adult somatic tissues. *Epigenomics*, **8**(6).

[153] **Ehrlich, M.**, **M. A. Gama-Sosa**, **L. H. Huang**, **R. M. Midgett**, **K. C. Kuo**, **R. A. McCune**, and **C. Gehrke** (1982). Amount and distribution of 5-methylcytosine in human DNA from different types of tissues of cells. *Nucleic Acids Res.*, **10**(8), 2709–2721.

[154] **Boyes, J.** and **A. Bird** (1991). DNA methylation inhibits transcription indirectly via a methyl-CpG binding protein. *Cell*, **64**(6).

[155] **Nan, X.**, **R. R. Meehan**, and **A. Bird** (1993). Dissection of the methyl-CpG binding domain from the chromosomal protein MeCP2. *Nucleic Acids Res.*, **21**(21), 4886–4892.

[156] **Bostick, M.**, **J. K. Kim**, **P. O. Estève**, **A. Clark**, **S. Pradhan**, and **S. E. Jacobsen** (2007). UHRF1 plays a role in maintaining DNA methylation in mammalian cells. *Science*, **317**(5845).

[157] **Hodges, A. J.**, **N. O. Hudson**, and **B. A. Buck-Koehntop** (2020). Cys2His2 zinc finger Methyl-CpG binding proteins: Getting a handle on methylated DNA. *J. Mol. Biol.*, **432**(6).

[158] **Nan, X.**, **H. H. Ng**, **C. A. Johnson**, **C. D. Laherty**, **B. M. Turner**, **R. N. Eisenman**, and **A. Bird** (1998). Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. *Nature*, **393**(6683), 386–389.

[159] **Fuks, F.**, **W. A. Burgers**, **A. Brehm**, **L. Hughes-Davies**, and **T. Kouzarides** (2000). DNA methyltransferase dnmt1 associates with histone deacetylase activity. *Nat. Genet.*, **24**(1).

[160] **Benetti, R.**, **S. Gonzalo**, **I. Jaco**, **P. Muñoz**, **S. Gonzalez**, **S. Schoeftner**, **E. Murchison**, **T. Andl**, **T. Chen**, **P. Klatt**, **E. Li**, **M. Serrano**, **S. Millar**, **G. Hannon**, and **M. A. Blasco** (2008). A mammalian microRNA cluster controls DNA methylation and telomere recombination via rbl2-dependent regulation of DNA methyltransferases. *Nat. Struct. Mol. Biol.*, **15**(9), 998.

[161] **Zhu, H.**, **G. Wang**, and **J. Qian** (2016). Transcription factors as readers and effectors of DNA methylation. *Nat. Rev. Genet.*, **17**(9), 551–565.

[162] **Moore, L. D.**, **T. Le**, and **G. Fan** (2012). DNA methylation and its basic function. *Neuropsychopharmacology*, **38**(1), 23–38.

[163] **Luger, K.**, **A. W. Mäder**, **R. K. Richmond**, **D. F. Sargent**, and **T. J. Richmond** (1997). Crystal structure of the nucleosome core particle at 2.8 a resolution. *Nature*, **389**(6648), 251–260.

[164] **Allfrey, V. G.**, **R. Faulkner**, and **A. E. Mirsky** (1964). ACETYLATION AND METHYLATION OF HISTONES AND THEIR POSSIBLE ROLE IN THE REGULATION OF RNA SYNTHESIS. *Proc. Natl. Acad. Sci. U. S. A.*, **51**(5), 786.

[165] **Dillon, S. C.**, **X. Zhang**, **R. C. Trievel**, and **X. Cheng** (2005). The SET-domain protein superfamily: protein lysine methyltransferases. *Genome Biol.*, **6**(8).

[166] **Herz, H.-M.**, **A. Garruss**, and **A. Shilatifard** (2013). SET for life: biochemical activities and biological functions of SET domain-containing proteins. *Trends Biochem. Sci.*, **38**(12), 621–639.

[167] **Musselman, C. A.**, **M.-E. Lalonde**, **J. Côté**, and **T. G. Kutateladze** (2012). Perceiving the epigenetic landscape through histone readers. *Nat. Struct. Mol. Biol.*, **19**(12), 1218–1227.

[168] **Maurer-Stroh, S.**, **N. J. Dickens**, **L. Hughes-Davies**, **T. Kouzarides**, **F. Eisenhaber**, and **C. P. Ponting** (2003). The tudor domain 'royal family': Tudor, plant agenet, chromo, PWWP and MBT domains. *Trends Biochem. Sci.*, **28**(2), 69–74.

[169] **Kouzarides, T.** (2007). Chromatin modifications and their function. *Cell*, **128**(4), 693–705.

[170] **Strahl, B. D.** and **C. D. Allis** (2000). The language of covalent histone modifications. *Nature*, **403**(6765), 41–45.

[171] **Shogren-Knaak, M.**, **H. Ishii**, **J.-M. Sun**, **M. J. Pazin**, **J. R. Davie**, and **C. L. Peterson** (2006). Histone H4-K16 acetylation controls chromatin structure and protein interactions. *Science*, **311**(5762), 844–847.

[172] **Dai, Z.**, **V. Ramesh**, and **J. W. Locasale** (2020). Author correction: The evolving metabolic landscape of chromatin biology and epigenetics. *Nat. Rev. Genet.*, **21**(12), 782.

[173] **Wiese, M.** and **A. J. Bannister** (2020). Two genomes, one cell: Mitochondrial-nuclear coordination via epigenetic pathways. *Molecular metabolism*, **38**.

[174] **Mishra, S.**, **N. Pandey**, **S. Chawla**, **M. Sharma**, **O. Chandra**, **I. P. Jha**, **D. SenGupta**, **K. N. Natarajan**, and **V. Kumar** (2023). Matching queried single-cell open-chromatin profiles to large pools of single-cell transcriptomes and epigenomes for reference supported analysis. *Genome Research*, **33**(2), 218–231.

[175] **Chen, C. C.**, **J. J. Carson**, **J. Feser**, **B. Tamburini**, **S. Zabaronick**, **J. Linger**, and **J. K. Tyler** (2008). Acetylated lysine 56 on histone H3 drives chromatin assembly after repair and signals for the completion of repair. *Cell*, **134**(2).

[176] **Hurd, P. J.**, **A. J. Bannister**, **K. Halls**, **M. A. Dawson**, **M. Vermeulen**, **J. V. Olsen**, **H. Ismail**, **J. Somers**, **M. Mann**, **T. Owen-Hughes**, **I. Gout**, and **T. Kouzarides** (2009). Phosphorylation of histone H3 thr-45 is linked to apoptosis. *J. Biol. Chem.*, **284**(24).

[177] **Smeenk, G.** and **N. Mailand** (2016). Writers, readers, and erasers of histone ubiquitylation in DNA Double-Strand break repair. *Front. Genet.*, **7**.

[178] **Hyun, K.**, **J. Jeon**, **K. Park**, and **J. Kim** (2017). Writing, erasing and reading histone lysine methylations. *Exp. Mol. Med.*, **49**(4), e324–e324.

[179] **Skene, P. J.** and **S. Henikoff** (2017). An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *Elife*, **6**.

[180] **Chen, X.**, **H. Xu**, **X. Shu**, and **C.-X. Song** (2023). Mapping epigenetic modifications by sequencing technologies. *Cell Death Differ.*, 1–10.

[181] **Shilatifard, A.** (2012). The COMPASS family of histone H3K4 methylases: mechanisms of regulation in development and disease pathogenesis. *Annu. Rev. Biochem.*, **81**, 65–95.

[182] **Cano-Rodriguez, D.**, **R. A. Gjaltema**, **L. J. Jilderda**, **P. Jellema**, **J. Dokter-Fokkens**, **M. H. Ruiters**, and **M. G. Rots** (2016). Writing of H3K4Me3 overcomes epigenetic silencing in a sustained but context-dependent manner. *Nat. Commun.*, **7**.

[183] **Howe, F. S.**, **H. Fischl**, **S. C. Murray**, and **J. Mellor** (2017). Is H3K4me3 instructive for transcription activation? *Bioessays*, **39**(1), 1–12.

[184] **Bledau, A. S.**, **K. Schmidt**, **K. Neumann**, **U. Hill**, **G. Ciotta**, **A. Gupta**, **D. C. Torres**, **J. Fu**, **A. Kranz**, **A. F. Stewart**, and **K. Anastassiadis** (2014). The H3K4 methyltransferase setd1a is first required at the epiblast stage, whereas setd1b becomes essential after gastrulation. *Development*, **141**(5), 1022–1035.

[185] **Benayoun, B. A.**, **E. A. Pollina**, **D. Ucar**, **S. Mahmoudi**, **K. Karra**, **E. D. Wong**, **K. Devarajan**, **A. C. Daugherty**, **A. B. Kundaje**, **E. Mancini**, **B. C. Hitz**, **R. Gupta**, **T. A. Rando**, **J. C. Baker**, **M. P. Snyder**, **J. M. Cherry**, and **A. Brunet** (2015). H3K4me3 breadth is linked to cell identity and transcriptional consistency. *Cell*, **163**(5), 1281–1286.

[186] **Siklenka, K.**, **S. Erkek**, **M. Godmann**, **R. Lambrot**, **S. McGraw**, **C. Lafleur**, **T. Cohen**, **J. Xia**, **M. Suderman**, **M. Hallett**, **J. Trasler**, **A. H. F. M. Peters**, and **S. Kimmins** (2015). Disruption of histone methylation in developing sperm impairs offspring health transgenerationally. *Science*, **350**(6261), aab2006.

[187] **Lismer, A.**, **V. Dumeaux**, **C. Lafleur**, **R. Lambrot**, **J. Brind'Amour**, **M. C. Lorincz**, and **S. Kimmins** (2021). Histone H3 lysine 4 trimethylation in sperm is transmitted to the embryo and associated with diet-induced phenotypes in the offspring. *Dev. Cell*, **56**(5).

[188] **Ooi, S. K. T.**, **C. Qiu**, **E. Bernstein**, **K. Li**, **D. Jia**, **Z. Yang**, **H. Erdjument-Bromage**, **P. Tempst**, **S.-P. Lin**, **C. D. Allis**, **X. Cheng**, and **T. H. Bestor** (2007). DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. *Nature*, **448**(7154), 714–717.

[189] **Murphy, P. J.**, **S. F. Wu**, **C. R. James**, **C. L. Wike**, and **B. R. Cairns** (2018). Placeholder nucleosomes underlie Germline-to-Embryo DNA methylation reprogramming. *Cell*, **172**(5), 993–1006.e13.

[190] **Bannister, A. J.**, **R. Schneider**, **F. A. Myers**, **A. W. Thorne**, **C. Crane-Robinson**, and **T. Kouzarides** (2005). Spatial distribution of di- and tri-methyl lysine 36 of histone H3 at active genes. *J. Biol. Chem.*, **280**(18), 17732–17736.

[191] **Carrozza, M. J.**, **B. Li**, **L. Florens**, **T. Suganuma**, **S. K. Swanson**, **K. K. Lee**, **W.-J. Shia**, **S. Anderson**, **J. Yates**, **M. P. Washburn**, and **J. L. Workman** (2005). Histone H3 methylation by set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription. *Cell*, **123**(4), 581–592.

[192] **Weinberg, D. N.**, **S. Papillon-Cavanagh**, **H. Chen**, **Y. Yue**, **X. Chen**, **K. N. Rajagopalan**, **C. Horth**, **J. T. McGuire**, **X. Xu**, **H. Nikbakht**, **A. E. Lemiesz**, **D. M. Marchione**, **M. R. Marunde**, **M. J. Meiners**, **M. A. Cheek**, **M.-C. Keogh**, **E. Bareke**, **A. Djedid**, **A. S. Harutyunyan**, **N. Jabado**, **B. A. Garcia**, **H. Li**, **C. D. Allis**, **J. Majewski**, and **C. Lu** (2019). The histone mark H3K36me2 recruits DNMT3A and shapes the intergenic DNA methylation landscape. *Nature*, **573**(7773), 281–286.

[193] **Lam, U. T. F.**, **B. K. Y. Tan**, **J. J. X. Poh**, and **E. S. Chen** (2022). Structural and functional specificity of H3K36 methylation. *Epigenetics Chromatin*, **15**(1), 1–20.

[194] **Dawson, M. A.**, **S. D. Foster**, **A. J. Bannister**, **S. C. Robson**, **R. Hannah**, **X. Wang**, **B. Xhemalce**, **A. D. Wood**, **A. R. Green**, **B. Göttgens**, and **T. Kouzarides** (2012). Three distinct patterns of histone H3Y41 phosphorylation mark active genes. *Cell Rep.*, **2**(3), 470–477.

[195] **Brehove, M.**, **T. Wang**, **J. North**, **Y. Luo**, **S. J. Dreher**, **J. C. Shimko**, **J. J. Ottesen**, **K. Luger**, and **M. G. Poirier** (2015). Histone core phosphorylation regulates DNA accessibility. *J. Biol. Chem.*, **290**(37), 22612–22621.

[196] **Lo, W. S.**, **R. C. Trievel**, **J. R. Rojas**, **L. Duggan**, **J. Y. Hsu**, **C. D. Allis**, **R. Marmorstein**, and **S. L. Berger** (2000). Phosphorylation of serine 10 in histone H3 is functionally linked in vitro and in vivo to gcn5-mediated acetylation at lysine 14. *Mol. Cell*, **5**(6), 917–926.

[197] **Zippo, A.**, **R. Serafini**, **M. Rocchigiani**, **S. Pennacchini**, **A. Krepelova**, and **S. Oliviero** (2009). Histone crosstalk between H3S10ph and H4K16ac generates a histone code that mediates transcription elongation. *Cell*, **138**(6), 1122–1136.

[198] **Fischle, W.**, **B. S. Tseng**, **H. L. Dormann**, **B. M. Ueberheide**, **B. A. Garcia**, **J. Shabanowitz**, **D. F. Hunt**, **H. Funabiki**, and **C. D. Allis** (2005). Regulation of HP1-chromatin binding by histone H3 methylation and phosphorylation. *Nature*, **438**(7071), 1116–1122.

[199] **Gehani, S. S.**, **S. Agrawal-Singh**, **N. Dietrich**, **N. S. Christophersen**, **K. Helin**, and **K. Hansen** (2010). Polycomb group protein displacement and gene activation through MSK-dependent H3K27me3S28 phosphorylation. *Mol. Cell*, **39**(6), 886–900.

[200] **Schuettengruber, B.** and **G. Cavalli** (2009). Recruitment of polycomb group complexes and their role in the dynamic regulation of cell fate choice. *Development*, **136**(21), 3531–3542.

[201] **Aranda, S.**, **G. Mas**, and **L. Di Croce** (2015). Regulation of gene transcription by polycomb proteins. *Science advances*, **1**(11).

[202] **Riising, E. M.**, **I. Comet**, **B. Leblanc**, **X. Wu**, **J. V. Johansen**, and **K. Helin** (2014). Gene silencing triggers polycomb repressive complex 2 recruitment to CpG islands genome wide. *Mol. Cell*, **55**(3), 347–360.

[203] **Allshire, R. C.** and **H. D. Madhani** (2018). Ten principles of heterochromatin formation and function. *Nat. Rev. Mol. Cell Biol.*, **19**(4), 229–244.

[204] **Nicetto, D.** and **K. S. Zaret** (2019). Role of H3K9me3 heterochromatin in cell identity establishment and maintenance. *Curr. Opin. Genet. Dev.*, **55**, 1–10.

[205] **Schotta, G.**, **M. Lachner**, **K. Sarma**, **A. Ebert**, **R. Sengupta**, **G. Reuter**, **D. Reinberg**, and **T. Jenuwein** (2004). A silencing pathway to induce H3-K9 and H4-K20 trimethylation at constitutive heterochromatin. *Genes Dev.*, **18**(11), 1251–1262.

[206] **Jack, A. P.**, **S. Bussemer**, **M. Hahn**, **S. Pünzeler**, **M. Snyder**, **M. Wells**, **G. Csankovszki**, **I. Solovei**, **G. Schotta**, and **S. B. Hake** (2013). H3K56me3 is a novel, conserved heterochromatic mark that largely but not completely overlaps with H3K9me3 in both regulation and localization. *PLoS One*, **8**(2).

[207] **Burton, A.**, **V. Brochard**, **C. Galan**, **E. R. Ruiz-Morales**, **Q. Rovira**, **D. Rodriguez-Terrones**, **K. Kruse**, **S. Le Gras**, **V. S. Udayakumar**, **H. G. Chin**, **A. Eid**, **X. Liu**, **C. Wang**, **S. Gao**, **S. Pradhan**, **J. M. Vaquerizas**, **N. Beaujean**, **T. Jenuwein**, and **M.-E. Torres-Padilla** (2020). Heterochromatin establishment during early mammalian development is regulated by pericentromeric RNA and characterized by non-repressive H3K9me3. *Nat. Cell Biol.*, **22**(7), 767–778.

[208] **Bannister, A. J.** and **T. Kouzarides** (2011). Regulation of chromatin by histone modifications. *Cell Res.*, **21**(3), 381–395.

[209] **Workman, J. L.** and **R. E. Kingston** (1998). Alteration of nucleosome structure as a mechanism of transcriptional regulation. *Annu. Rev. Biochem.*, **67**.

[210] **et al., T. K.** (2020). Transcription through the nucleosome. *Curr. Opin. Struct. Biol.*, **61**, 42–49.

[211] **Pandey, N.**, **O. Chandra**, **S. Mishra**, and **V. Kumar** (2021). Improving chromatin-interaction prediction using single-cell open-chromatin profiles and making insight into the cis-regulatory landscape of the human brain. *Frontiers in Genetics*, **12**, 738194.

[212] **Klemm, S. L.**, **Z. Shipony**, and **W. J. Greenleaf** (2019). Chromatin accessibility and the regulatory epigenome. *Nat. Rev. Genet.*, **20**(4).

[213] **Grewal, S. I.** and **D. Moazed** (2003). Heterochromatin and epigenetic control of gene expression. *Science*, **301**(5634).

[214] **Pandey, N.**, **M. Sharma**, **A. Mathur**, **C. G. Anene-Nzel**, **M. Hakimullah**, **I. P. Jha**, **O. Chandra**, **S. Mishra**, **A. Sharma**, **R. Foo**, *et al.* (2023). Deciphering the phenotypic heterogeneity and drug response in cancer cells using genome-wide activity and interaction of chromatin domains. *bioRxiv*.

[215] **Wang, Y.-M.**, **P. Zhou**, **L.-Y. Wang**, **Z.-H. Li**, **Y.-N. Zhang**, and **Y.-X. Zhang** (2012). Correlation between DNase I hypersensitive site distribution and gene expression in HeLa S3 cells. *PLoS One*, **7**(8).

[216] **Flavahan, W. A.**, **Y. Drier**, **B. B. Liau**, **S. M. Gillespie**, **A. S. Venteicher**, **A. O. Stemmer-Rachamimov**, **M. L. Suvà**, and **B. E. Bernstein** (2016). Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature*, **529**(7584), 110–114.

[217] **Cho, S. W.**, **J. Xu**, **R. Sun**, **M. R. Mumbach**, **A. C. Carter**, **Y. G. Chen**, **K. E. Yost**, **J. Kim**, **J. He**, **S. A. Nevins**, **S.-F. Chin**, **C. Caldas**, **S. J. Liu**, **M. A. Horlbeck**, **D. A. Lim**, **J. S. Weissman**, **C. Curtis**, and **H. Y. Chang** (2018). Promoter of lncRNA gene PVT1 is a Tumor-Suppressor DNA boundary element. *Cell*, **173**(6), 1398–1412.e22.

[218] **Fuda, N. J.**, **M. Behfar Ardehali**, and **J. T. Lis** (2009). Defining mechanisms that regulate RNA polymerase II transcription in vivo. *Nature*, **461**(7261), 186.

[219] **Voss, T. C.** and **G. L. Hager** (2013). Dynamic regulation of transcriptional states by chromatin and transcription factors. *Nat. Rev. Genet.*, **15**(2), 69–81.

[220] **Hafner, A.** and **A. Boettiger** (2022). The spatial organization of transcriptional control. *Nat. Rev. Genet.*, **24**(1), 53–68.

[221] **Kim, S.** and **J. Wysocka** (2023). Deciphering the multi-scale, quantitative cis-regulatory code. *Mol. Cell*, **83**(3).

[222] **Isoda, T.**, **A. J. Moore**, **Z. He**, **V. Chandra**, **M. Aida**, **M. Denholtz**, **P. van Hamburg J**, **K. M. Fisch**, **A. N. Chang**, **S. P. Fahl**, **D. L. Wiest**, and **C. Murre** (2017). Non-coding transcription instructs chromatin folding and compartmentalization to dictate Enhancer-Promoter communication and T cell fate. *Cell*, **171**(1).

[223] **Mumbach, M. R.**, **J. M. Granja**, **R. A. Flynn**, **C. M. Roake**, **A. T. Satpathy**, **A. J. Rubin**, **Y. Qi**, **Z. Jiang**, **S. Shams**, **B. H. Louie**, **J. K. Guo**, **D. G. Gennert**, **M. R. Corces**, **P. A. Khavari**, **M. K. Atianand**, **S. E. Artandi**, **K. A. Fitzgerald**, **W. J. Greenleaf**, and **H. Y. Chang** (2019). HiChIRP reveals RNA-associated chromosome conformation. *Nat. Methods*, **16**(6), 489–492.

[224] **Yang, F.**, **X. Deng**, **W. Ma**, **J. B. Berletch**, **N. Rabaia**, **G. Wei**, **J. M. Moore**, **G. N. Filippova**, **J. Xu**, **Y. Liu**, **W. S. Noble**, **J. Shendure**, and **C. M. Disteche** (2015). The lncRNA firre anchors the inactive X chromosome to the nucleolus by binding CTCF and maintains H3K27me3 methylation. *Genome Biol.*, **16**(1).

[225] **Saldaña-Meyer, R.**, **J. Rodriguez-Hernaez**, **T. Escobar**, **M. Nishana**, **K. Jácome-López**, **E. P. Nora**, **B. G. Bruneau**, **A. Tsirigos**, **M. Furlan-Magaril**, **J. Skok**, and **D. Reinberg** (2019). RNA interactions are essential for CTCF-Mediated genome organization. *Mol. Cell*, **76**(3), 412–422.e5.

[226] **Sigova, A. A.**, **B. J. Abraham**, **X. Ji**, **B. Molinie**, **N. M. Hannett**, **Y. E. Guo**, **M. Jangi**, **C. C. Giallourakis**, **P. A. Sharp**, and **R. A. Young** (2015). Transcription factor trapping by RNA in gene regulatory elements. *Science*, **350**(6263).

[227] **Liu, Y.**, **S. Chen**, **S. Wang**, **F. Soares**, **M. Fischer**, **F. Meng**, **Z. Du**, **C. Lin**, **C. Meyer**, **J. A. DeCaprio**, **M. Brown**, **X. S. Liu**, and **H. H. He** (2017). Transcriptional landscape of the human cell cycle. *Proceedings of the National Academy of Sciences*, **114**(13), 3473–3478.

[228] **Consortium** (1998). Identification of regulatory regions which confer muscle-specific gene expression. *Journal of Molecular Biology*, **278**(1), 167–181.

[229] **Geeven, G.**, **R. E. van Kesteren**, **A. B. Smit**, and **M. C. M. de Gunst** (2011). Identification of context-specific gene regulatory networks with GEMULA—gene expression modeling using LAsso. *Bioinformatics*, **28**(2), 214–221.

[230] **Erdem, C.**, **S. M. Gross**, **L. M. Heiser**, and **M. R. Birtwistle** (2023). MOBILE pipeline enables identification of context-specific networks and regulatory mechanisms. *Nat. Commun.*, **14**(1), 1–16.

[231] **Chawla, S.**, **S. Samydurai**, **S. L. Kong**, **Z. Wu**, **Z. Wang**, **W. L. Tam**, **D. Sengupta**, and **V. Kumar** (2021). Unipath: a uniform approach for pathway and gene-set based analysis of heterogeneity in single-cell epigenome and transcriptome profiles. *Nucleic acids research*, **49**(3), e13–e13.

[232] **Matharu, N.** and **N. Ahituv** (2020). Modulating gene regulation to treat genetic disorders. *Nat. Rev. Drug Discov.*, **19**(11), 757–775.

[233] **Amaral, P. P.**, **T. Leonardi**, **N. Han**, **E. Viré**, **D. K. Gascoigne**, **R. Arias-Carrasco**, **M. Büscher**, **L. Pandolfini**, **A. Zhang**, **S. Pluchino**, **V. Maracaja-Coutinho**, **H. I. Nakaya**, **M. Hemberg**, **R. Shiekhattar**, **A. J. Enright**, and **T. Kouzarides** (2018). Genomic positional conservation identifies topological anchor point RNAs linked to developmental loci. *Genome Biol.*, **19**(1), 1–21.

[234] **Lodde, V.**, **G. Murgia**, **E. R. Simula**, **M. Steri**, **M. Floris**, and **M. L. Idda** (2020). Long noncoding RNAs and circular RNAs in autoimmune diseases. *Biomolecules*, **10**(7).

[235] **Maurano, M. T.**, **R. Humbert**, **E. Rynes**, **R. E. Thurman**, **E. Haugen**, **H. Wang**, **A. P. Reynolds**, **R. Sandstrom**, **H. Qu**, **J. Brody**, **A. Shafer**, **F. Neri**, **K. Lee**, **T. Kutyavin**, **S. Stehling-Sun**, **A. K. Johnson**, **T. K. Canfield**, **E. Giste**, **M. Diegel**, **D. Bates**, **R. Scott Hansen**, **S. Neph**, **P. J. Sabo**, **S. Heimfeld**, **A. Raubitschek**, **S. Ziegler**, **C. Cotsapas**, **N. Sotoodehnia**, **I. Glass**, **S. R. Sunyaev**, **R. Kaul**, and **J. A. Stamatoyannopoulos** (2012). Systematic localization of common Disease-Associated variation in regulatory DNA. *Science*, **337**(6099), 1190.

[236] **Lagos-Quintana, M.**, **R. Rauhut**, **W. Lendeckel**, and **T. Tuschl** (2001). Identification of novel genes coding for small expressed RNAs. *Science*, **294**(5543).

[237] **Tomari, Y.** and **P. D. Zamore** (2005). Perspective: machines for RNAi. *Genes Dev.*, **19**(5), 517–529.

[238] **Dana, H.**, **G. M. Chalbatani**, **H. Mahmoodzadeh**, **R. Karimloo**, **O. Rezaiean**, **A. Moradzadeh**, **N. Mehmandoost**, **F. Moazzen**, **A. Mazraeh**, **V. Marmari**, **M. Ebrahimi**, **M. M. Rashno**, **S. J. Abadi**, and **E. Gharagouzlo** (2017). Molecular mechanisms and biological functions of siRNA. *Int. J. Biomed. Sci.*, **13**(2), 48.

[239] **Wang, X.**, **A. Ramat**, **M. Simonelig**, and **M. F. Liu** (2023). Emerging roles and functional mechanisms of PIWI-interacting RNAs. *Nat. Rev. Mol. Cell Biol.*, **24**(2).

[240] **Ma, L.**, **V. B. Bajic**, and **Z. Zhang** (2013). On the classification of long noncoding RNAs. *RNA Biol.*, **10**(6), 924.

[241] **Liang, Y.**, **N. Liu**, **L. Yang**, **J. Tang**, **Y. Wang**, and **M. Mei** (2021). A brief review of circRNA biogenesis, detection, and function. *Curr. Genomics*, **22**(7), 485.

[242] **Stein, A. J.**, **G. Fuchs**, **C. Fu**, **S. L. Wolin**, and **K. M. Reinisch** (2005). Structural insights into RNA quality control: the ro autoantigen binds misfolded RNAs via its central cavity. *Cell*, **121**(4).

[243] **Christov, C. P.**, **T. J. Gardiner**, **D. Szüts**, and **T. Krude** (2006). Functional requirement of noncoding Y RNAs for human chromosomal DNA replication. *Mol. Cell. Biol.*, **26**(18).

[244] **Sun, Y.-M.** and **Y.-Q. Chen** (2020). Principles and innovative technologies for decrypting noncoding RNAs: from discovery and functional prediction to clinical application. *J. Hematol. Oncol.*, **13**(1), 1–27.

[245] **Stephens, Z. D.**, **S. Y. Lee**, **F. Faghri**, **R. H. Campbell**, **C. Zhai**, **M. J. Efron**, **R. Iyer**, **M. C. Schatz**, **S. Sinha**, and **G. E. Robinson** (2015). Big data: Astronomical or genomical? *PLoS Biol.*, **13**(7).

[246] **Subramanian, I.**, **S. Verma**, **S. Kumar**, **A. Jere**, and **K. Anamika** (2020). Multi-omics data integration, interpretation, and its application. *Bioinformatics and biology insights*, **14**, 1177932219899051.

[247] **Bersanelli, M.**, **E. Mosca**, **D. Remondini**, **E. Giampieri**, **C. Sala**, **G. Castellani**, and **L. Milanesi** (2016). Methods for the integration of multi-omics data: mathematical aspects. *BMC bioinformatics*, **17**, 167–177.

[248] **Jung, G. T.**, **K.-P. Kim**, and **K. Kim** (2020). How to interpret and integrate multi-omics data at systems level. *Animal cells and systems*, **24**(1), 1–7.

[249] **Sharma, M.**, **I. P. Jha**, **S. Chawla**, **N. Pandey**, **O. Chandra**, **S. Mishra**, and **V. Kumar** (2022). Associating pathways with diseases using single-cell expression profiles and making inferences about potential drugs. *Briefings in Bioinformatics*, **23**(4), bbac241.

[250] **Routledge, M.** and **A. Conway Morris** (2024). all models are wrong, some are useful: George box.

[251] **Baysoy, A.**, **Z. Bai**, **R. Satija**, and **R. Fan** (2023). The technological landscape and applications of single-cell multi-omics. *Nature Reviews Molecular Cell Biology*, **24**(10), 695–713.

[252] **Fischer, M.**, **A. E. Schade**, **T. B. Branigan**, **G. A. Müller**, and **J. A. DeCaprio** (2022). Coordinating gene expression during the cell cycle. *Trends Biochem. Sci.*, **47**(12), 1009–1022.

[253] **Kouno, T.**, **M. de Hoon**, **J. C. Mar**, **Y. Tomaru**, **M. Kawano**, **P. Carninci**, **H. Suzuki**, **Y. Hayashizaki**, and **J. W. Shin** (2013). Temporal dynamics and transcriptional control using single-cell gene expression analysis. *Genome Biol.*, **14**(10), 1–12.

[254] **Wang, R.**, **Y. Wang**, **X. Zhang**, **Y. Zhang**, **X. Du**, **Y. Fang**, and **G. Li** (2019). Hierarchical cooperation of transcription factors from integration analysis of DNA sequences, ChIP-Seq and ChIA-PET data. *BMC Genomics*, **20**(3), 1–13.

[255] **et. al., A. R.** (2022). Transcriptional and epigenetic regulation of temporal patterning in neural progenitors. *Dev. Biol.*, **481**, 116–128.

[256] **Rinn, J. L.** and **H. Y. Chang** (2012). Genome regulation by long noncoding RNAs. *Annu. Rev. Biochem.*, **81**, 145–166.

[257] **Wang, K. C.** and **H. Y. Chang** (2011). Molecular mechanisms of long noncoding RNAs. *Mol. Cell*, **43**(6), 904.

[258] **Zhang, X.**, **W. Wang**, **W. Zhu**, **J. Dong**, **Y. Cheng**, **Z. Yin**, and **F. Shen** (2019). Mechanisms and functions of long Non-Coding RNAs at multiple regulatory levels. *Int. J. Mol. Sci.*, **20**(22).

[259] **Noviello, T. M. R.**, **A. Di Liddo**, **G. M. Ventola**, **A. Spagnuolo**, **S. D'Aniello**, **M. Ceccarelli**, and **L. Cerulo** (2018). Detection of long non-coding RNA homology, a comparative study on alignment and alignment-free metrics. *BMC Bioinformatics*, **19**(1), 407.

[260] **Zhao, Y.**, **J. Wang**, **J. Chen**, **X. Zhang**, **M. Guo**, and **G. Yu** (2020). A literature review of gene function prediction by modeling gene ontology. *Front. Genet.*, **11**, 400.

[261] **Zhang, H.**, **C.-L. Hung**, **M. Liu**, **X. Hu**, and **Y.-Y. Lin** (2019). NCNet: Deep learning network models for predicting function of non-coding DNA. *Front. Genet.*, **10**, 432.

[262] **Zhou, N.**, **Y. Jiang**, **T. R. Bergquist**, **A. J. Lee**, **B. Z. Kacsoh**, **A. W. Crocker**, **K. A. Lewis**, **G. Georghiou**, **H. N. Nguyen**, **M. N. Hamid**, **L. Davis**, **T. Dogan**, **V. Atalay**, **A. S. Rifaioglu**, **A. Dalkıran**, **R. Cetin Atalay**, **C. Zhang**, **R. L. Hurto**, **P. L. Freddolino**, **Y. Zhang**, **P. Bhat**, **F. Supek**, **J. M. Fernández**, **B. Gemovic**, **V. R. Perovic**, **R. S. Davidović**, **N. Sumonja**, **N. Veljkovic**, **E. Asgari**, **M. R. K. Mofrad**, **G. Profiti**, **C. Savojardo**, **P. L. Martelli**, **R. Casadio**, **F. Boecker**, **H. Schoof**, **I. Kahanda**, **N. Thurlby**, **A. C. McHardy**, **A. Renaux**, **R. Saidi**, **J. Gough**, **A. A. Freitas**, **M. Antczak**, **F. Fabris**, **M. N. Wass**, **J. Hou**, **J. Cheng**, **Z. Wang**, **A. E. Romero**, **A. Paccanaro**, **H. Yang**, **T. Goldberg**, **C. Zhao**, **L. Holm**, **P. Törönen**, **A. J. Medlar**, **E. Zosa**, **I. Borukhov**, **I. Novikov**, **A. Wilkins**, **O. Lichtarge**, **P.-H. Chi**, **W.-C. Tseng**, **M. Linial**, **P. W. Rose**, **C. Dessimoz**, **V. Vidulin**, **S. Dzeroski**, **I. Sillitoe**, **S. Das**, **J. G. Lees**, **D. T. Jones**, **C. Wan**, **D. Cozzetto**, **R. Fa**, **M. Torres**, **A. Warwick Vesztrocy**, **J. M. Rodriguez**, **M. L. Tress**, **M. Frasca**, **M. Notaro**, **G. Grossi**, **A. Petrini**, **M. Re**, **G. Valentini**, **M. Mesiti**, **D. B. Roche**, **J. Reeb**, **D. W. Ritchie**, **S. Aridhi**, **S. Z. Alborzi**, **M.-D. Devignes**, **D. C. E. Koo**, **R. Bonneau**, **V. Gligorijević**, **M. Barot**, **H. Fang**, **S. Toppo**, **E. Lavezzo**, **M. Falda**, **M. Berselli**, **S. C. E. Tosatto**, **M. Carraro**, **D. Piovesan**, **H. Ur Rehman**, **Q. Mao**, **S. Zhang**,

S. Vucetic, **G. S. Black**, **D. Jo**, **E. Suh**, **J. B. Dayton**, **D. J. Larsen**, **A. R. Omdahl**, **L. J. McGuffin**, **D. A. Brackenridge**, **P. C. Babbitt**, **J. M. Yunes**, **P. Fontana**, **F. Zhang**, **S. Zhu**, **R. You**, **Z. Zhang**, **S. Dai**, **S. Yao**, **W. Tian**, **R. Cao**, **C. Chandler**, **M. Amezola**, **D. Johnson**, **J.-M. Chang**, **W.-H. Liao**, **Y.-W. Liu**, **S. Pascarelli**, **Y. Frank**, **R. Hoehndorf**, **M. Kulmanov**, **I. Boudellioua**, **G. Politano**, **S. Di Carlo**, **A. Benso**, **K. Hakala**, **F. Ginter**, **F. Mehryary**, **S. Kaewphan**, **J. Björne**, **H. Moen**, **M. E. E. Tolvanen**, **T. Salakoski**, **D. Kihara**, **A. Jain**, **T. Šmuc**, **A. Altenhoff**, **A. Ben-Hur**, **B. Rost**, **S. E. Brenner**, **C. A. Orengo**, **C. J. Jeffery**, **G. Bosco**, **D. A. Hogan**, **M. J. Martin**, **C. O'Donovan**, **S. D. Mooney**, **C. S. Greene**, **P. Radivojac**, and **I. Friedberg** (2019). The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome Biol.*, **20**(1), 244.

[263] **Kulmanov, M.** and **R. Hoehndorf** (2021). DeepGOPlus: improved protein function prediction from sequence. *Bioinformatics*, **37**(8), 1187.

[264] **Yang, X.**, **J. Coulombe-Huntington**, **S. Kang**, **G. M. Sheynkman**, **T. Hao**, **A. Richardson**, **S. Sun**, **F. Yang**, **Y. A. Shen**, **R. R. Murray**, **K. Spirohn**, **B. E. Begg**, **M. Duran-Frigola**, **A. MacWilliams**, **S. J. Pevzner**, **Q. Zhong**, **S. A. Wanamaker**, **S. Tam**, **L. Ghamsari**, **N. Sahni**, **S. Yi**, **M. D. Rodriguez**, **D. Balcha**, **G. Tan**, **M. Costanzo**, **B. Andrews**, **C. Boone**, **X. J. Zhou**, **K. Salehi-Ashtiani**, **B. Charloteaux**, **A. A. Chen**, **M. A. Calderwood**, **P. Aloy**, **F. P. Roth**, **D. E. Hill**, **L. M. Iakoucheva**, **Y. Xia**, and **M. Vidal** (2016). Widespread expansion of protein interaction capabilities by alternative splicing. *Cell*, **164**(4), 805–817.

[265] **Yang, P.**, **X.-L. Li**, **J.-P. Mei**, **C.-K. Kwoh**, and **S.-K. Ng** (2012). Positive-unlabeled learning for disease gene identification. *Bioinformatics*, **28**(20), 2640–2647.

[266] **Liao, Q.**, **C. Liu**, **X. Yuan**, **S. Kang**, **R. Miao**, **H. Xiao**, **G. Zhao**, **H. Luo**, **D. Bu**, **H. Zhao**, *et al.* (2011). Large-scale prediction of long non-coding rna functions in a coding–non-coding gene co-expression network. *Nucleic acids research*, **39**(9), 3864–3878.

[267] **Liao, Q.**, **C. Liu**, **X. Yuan**, **S. Kang**, **R. Miao**, **H. Xiao**, **G. Zhao**, **H. Luo**, **D. Bu**, **H. Zhao**, **G. Skogerbø**, **Z. Wu**, and **Y. Zhao** (2011). Large-scale prediction of long non-coding RNA functions in a coding-non-coding gene co-expression network. *Nucleic Acids Res.*, **39**(9).

[268] **Uygun, S.**, **C. Peng**, **M. D. Lehti-Shiu**, **R. L. Last**, and **S.-H. Shiu** (2016). Utility and limitations of using gene expression data to identify functional associations. *PLoS Comput. Biol.*, **12**(12).

[269] **Sun, X.** and **D. Wong** (2016). Long non-coding RNA-mediated regulation of glucose homeostasis and diabetes. *Am. J. Cardiovasc. Dis.*, **6**(2), 17–25.

[270] **Zhang, J.**, **S. Zou**, and **L. Deng** (2018). Gene ontology-based function prediction of long non-coding RNAs using bi-random walk. *BMC Med. Genomics*, **11**(5), 1–10.

[271] **Pyfrom, S. C.**, **H. Luo**, and **J. E. Payton** (2019). PLAIDOH: a novel method for functional prediction of long non-coding RNAs identifies cancer-specific LncRNA activities. *BMC Genomics*, **20**(1), 1–24.

[272] **Venters, B. J.** and **B. F. Pugh** (2013). Genomic organization of human transcription initiation complexes. *Nature*, **502**(7469).

[273] **Yan, J.**, **Y. Qiu**, **R. Dos Santos AM**, **Y. Yin**, **Y. E. Li**, **N. Vinckier**, **N. Nariai**, **P. Benaglio**, **A. Raman**, **X. Li**, **S. Fan**, **J. Chiou**, **F. Chen**, **K. A. Frazer**, **K. J. Gaulton**, **M. Sander**, **J. Taipale**, and **B. Ren** (2021). Systematic analysis of binding of transcription factors to noncoding variants. *Nature*, **591**(7848).

[274] **Li, B.**, **M. Carey**, and **J. L. Workman** (2007). The role of chromatin during transcription. *Cell*, **128**(4).

[275] **Kumar, V.**, **M. Muratani**, **N. A. Rayan**, **P. Kraus**, **T. Lufkin**, **H. H. Ng**, and **S. Prabhakar** (2013). Uniform, optimal signal processing of mapped deep-sequencing data. *Nat. Biotechnol.*, **31**(7), 615–622.

[276] **Tak, Y. G.** and **P. J. Farnham** (2015). Making sense of GWAS: using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome. *Epigenetics Chromatin*, **8**, 57.

[277] **Roider, H. G.**, **T. Manke**, **S. O'Keeffe**, **M. Vingron**, and **S. A. Haas** (2009). PASTAA: identifying transcription factors associated with sets of co-regulated genes. *Bioinformatics*, **25**(4), 435–442.

[278] **Xu, W.**, **X. Zhao**, **X. Wang**, **H. Feng**, **M. Gou**, **W. Jin**, **X. Wang**, **X. Liu**, and **C. Dong** (2019). The transcription factor tox2 drives T follicular helper cell development via regulating chromatin accessibility. *Immunity*, **51**(5), 826–839.e5.

[279] **Ahmed, M.**, **D. S. Min**, and **D. R. Kim** (2020). Integrating binding and expression data to predict transcription factors combined function. *BMC Genomics*, **21**(1), 610.

[280] **Venkatesh, I.**, **V. Mehra**, **Z. Wang**, **M. T. Simpson**, **E. Eastwood**, **A. Chakraborty**, **Z. Beine**, **D. Gross**, **M. Cabahug**, **G. Olson**, and **M. G. Blackmore** (2021). Co-occupancy identifies transcription factor co-operation for axon growth. *Nat. Commun.*, **12**(1), 2555.

[281] **Oki, S.**, **T. Ohta**, **G. Shioi**, **H. Hatanaka**, **O. Ogasawara**, **Y. Okuda**, **H. Kawaji**, **R. Nakaki**, **J. Sese**, and **C. Meno** (2018). ChIP-Atlas: a data-mining suite powered by full integration of public ChIP-seq data. *EMBO Rep.*, **19**(12).

[282] **Pohl, A.** and **M. Beato** (2014). bwtool: a tool for bigwig files. *Bioinformatics*, **30**(11), 1618–1619.

[283] **Noguchi, S.**, **T. Arakawa**, **S. Fukuda**, **M. Furuno**, **A. Hasegawa**, **F. Hori**, **S. Ishikawa-Kato**, **K. Kaida**, **A. Kaiho**, **M. Kanamori-Katayama**, **T. Kawashima**, **M. Kojima**, **A. Kubosaki**, **R.-I. Manabe**, **M. Murata**, **S. Nagao-Sato**, **K. Nakazato**, **N. Ninomiya**, **H. Nishiyori-Sueki**, **S. Noma**, **E. Saijyo**, **A. Saka**, **M. Sakai**, **C. Simon**, **N. Suzuki**, **M. Tagami**, **S. Watanabe**, **S. Yoshida**, **P. Arner**, **R. A. Axton**, **M. Babina**, **J. K. Baillie**, **T. C. Barnett**,

A. G. Beckhouse, **A. Blumenthal**, **B. Bodega**, **A. Bonetti**, **J. Briggs**, **F. Brombacher**, **A. J. Carlisle**, **H. C. Clevers**, **C. A. Davis**, **M. Detmar**, **T. Dohi**, **A. S. B. Edge**, **M. Edinger**, **A. Ehrlund**, **K. Ekwall**, **M. Endoh**, **H. Enomoto**, **A. Eslami**, **M. Fagiolini**, **L. Fairbairn**, **M. C. Farach-Carson**, **G. J. Faulkner**, **C. Ferrai**, **M. E. Fisher**, **L. M. Forrester**, **R. Fujita**, **J.-I. Furusawa**, **T. B. Geijtenbeek**, **T. Gingeras**, **D. Goldowitz**, **S. Guhl**, **R. Guler**, **S. Gustincich**, **T. J. Ha**, **M. Hamaguchi**, **M. Hara**, **Y. Hasegawa**, **M. Herlyn**, **P. Heutink**, **K. J. Hitchens**, **D. A. Hume**, **T. Ikawa**, **Y. Ishizu**, **C. Kai**, **H. Kawamoto**, **Y. I. Kawamura**, **J. S. Kempfle**, **T. J. Kenna**, **J. Kere**, **L. M. Khachigian**, **T. Kitamura**, **S. Klein**, **S. P. Klinken**, **A. J. Knox**, **S. Kojima**, **H. Koseki**, **S. Koyasu**, **W. Lee**, **A. Lennartsson**, **A. Mackay-sim**, **N. Mejhert**, **Y. Mizuno**, **H. Morikawa**, **M. Morimoto**, **K. Moro**, **K. J. Morris**, **H. Motohashi**, **C. L. Mummery**, **Y. Nakachi**, **F. Nakahara**, **T. Nakamura**, **Y. Nakamura**, **T. Nozaki**, **S. Ogishima**, **N. Ohkura**, **H. Ohno**, **M. Ohshima**, **M. Okada-Hatakeyama**, **Y. Okazaki**, **V. Orlando**, **D. A. Ovchinnikov**, **R. Passier**, **M. Patrikakis**, **A. Pombo**, **S. Pradhan-Bhatt**, **X.-Y. Qin**, **M. Rehli**, **P. Rizzu**, **S. Roy**, **A. Sajantila**, **S. Sakaguchi**, **H. Sato**, **H. Satoh**, **S. Savvi**, **A. Saxena**, **C. Schmidl**, **C. Schneider**, **G. G. Schulze-Tanzil**, **A. Schwegmann**, **G. Sheng**, **J. W. Shin**, **D. Sugiyama**, **T. Sugiyama**, **K. M. Summers**, **N. Takahashi**, **J. Takai**, **H. Tanaka**, **H. Tatsukawa**, **A. Tomoiu**, **H. Toyoda**, **M. van de Wetering**, **L. M. van den Berg**, **R. Verardo**, **D. Vijayan**, **C. A. Wells**, **L. N. Winteringham**, **E. Wolvetang**, **Y. Yamaguchi**, **M. Yamamoto**, **C. Yanagi-Mizuochi**, **M. Yoneda**, **Y. Yonekura**, **P. G. Zhang**, **S. Zucchelli**, **I. Abugessaisa**, **E. Arner**, **J. Harshbarger**, **A. Kondo**, **T. Lassmann**, **M. Lizio**, **S. Sahin**, **T. Sengstag**, **J. Severin**, **H. Shimoji**, **M. Suzuki**, **H. Suzuki**, **J. Kawai**, **N. Kondo**, **M. Itoh**, **C. O. Daub**, **T. Kasukawa**, **H. Kawaji**, **P. Carninci**, **A. R. R. Forrest**, and **Y. Hayashizaki** (2017). FANTOM5 CAGE profiles of human and mouse samples. *Scientific Data*, **4**(1), 1–10.

[284] **O'Leary, N. A.**, **M. W. Wright**, **J. R. Brister**, **S. Ciufo**, **D. Haddad**, **R. McVeigh**, **B. Rajput**, **B. Robbertse**, **B. Smith-White**, **D. Ako-Adjei**, **A. Astashyn**, **A. Badretdin**, **Y. Bao**, **O. Blinkova**, **V. Brover**, **V. Chetvernin**, **J. Choi**, **E. Cox**, **O. Ermolaeva**, **C. M. Farrell**, **T. Goldfarb**, **T. Gupta**, **D. Haft**, **E. Hatcher**, **W. Hlavina**, **V. S. Joardar**, **V. K. Kodali**, **W. Li**, **D. Maglott**, **P. Masterson**, **K. M. McGarvey**, **M. R. Murphy**, **K. O'Neill**, **S. Pujar**, **S. H. Rangwala**, **D. Rausch**, **L. D. Riddick**, **C. Schoch**, **A. Shkeda**, **S. S. Storz**, **H. Sun**, **F. Thibaud-Nissen**, **I. Tolstoy**, **R. E. Tully**, **A. R. Vatsan**, **C. Wallin**, **D. Webb**, **W. Wu**, **M. J. Landrum**, **A. Kimchi**, **T. Tatusova**, **M. DiCuccio**, **P. Kitts**, **T. D. Murphy**, and **K. D. Pruitt** (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**(D1), D733–45.

[285] **Brenning, A.**, Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: The r package sperrorest. *In 2012 IEEE international geoscience and remote sensing symposium*. IEEE, 2012.

[286] **Huynh-Thu, V. A.**, **A. Irrthum**, **L. Wehenkel**, and **P. Geurts** (2010). Inferring regulatory networks from expression data using tree-based methods. *PLoS One*, **5**(9).

[287] **Marbach, D.**, **J. C. Costello**, **R. Küffner**, **N. M. Vega**, **R. J. Prill**, **D. M. Camacho**, **K. R. Allison**, **DREAM5 Consortium**, **M. Kellis**, **J. J. Collins**, and **G. Stolovitzky** (2012). Wisdom of crowds for robust gene network inference. *Nat. Methods*, **9**(8), 796–804.

[288] **Aibar, S.**, **C. B. González-Blas**, **T. Moerman**, **V. A. Huynh-Thu**, **H. Imrichova**, **G. Hulselmans**, **F. Rambow**, **J.-C. Marine**, **P. Geurts**, **J. Aerts**, **J. van den Oord**, **Z. K. Atak**, **J. Wouters**, and **S. Aerts** (2017). SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods*, **14**(11), 1083–1086.

[289] **Kulmanov, M.**, **M. A. Khan**, **R. Hoehndorf**, and **J. Wren** (2018). DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics*, **34**(4), 660–668.

[290] **Miller, H. E.** and **A. J. R. Bishop** (2021). Correlation AnalyzeR: functional predictions from gene co-expression correlations. *BMC Bioinformatics*, **22**(1), 206.

[291] **Urzúa-Traslaviña, C. G.**, **V. C. Leeuwenburgh**, **A. Bhattacharya**, **S. Loipfinger**, **M. A. T. M. van Vugt**, **E. G. E. de Vries**, and **R. S. N. Fehrmann** (2021). Improving gene function predictions using independent transcriptional components. *Nat. Commun.*, **12**(1), 1464.

[292] **Yao, S.**, **R. You**, **S. Wang**, **Y. Xiong**, **X. Huang**, and **S. Zhu** (2021). NetGO 2.0: improving large-scale protein function prediction with massive sequence, text, domain, family and network information. *Nucleic Acids Res.*, **49**(W1), W469–W475.

[293] **The UniProt Consortium** (2017). UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **45**(D1), D158–D169.

[294] **Haney, M. S.**, **C. J. Bohlen**, **D. W. Morgens**, **J. A. Ousey**, **A. A. Barkal**, **C. K. Tsui**, **B. K. Ego**, **R. Levin**, **R. A. Kamber**, **H. Collins**, **A. Tucker**, **A. Li**, **D. Vorselen**, **L. Labitigan**, **E. Crane**, **E. Boyle**, **L. Jiang**, **J. Chan**, **E. Rincón**, **W. J. Greenleaf**, **B. Li**, **M. P. Snyder**, **I. L. Weissman**, **J. A. Theriot**, **S. R. Collins**, **B. A. Barres**, and **M. C. Bassik** (2018). Identification of phagocytosis regulators using magnetic genome-wide CRISPR screens. *Nat. Genet.*, **50**(12), 1716–1727.

[295] **Yilmaz, A.**, **M. Peretz**, **A. Aharony**, **I. Sagi**, and **N. Benvenisty** (2018). Defining essential genes for human pluripotent stem cells by CRISPR-Cas9 screening in haploid cells. *Nat. Cell Biol.*, **20**(5), 610–619.

[296] **Jeng, E. E.**, **V. Bhadkamkar**, **N. U. Ibe**, **H. Gause**, **L. Jiang**, **J. Chan**, **R. Jian**, **D. Jimenez-Morales**, **E. Stevenson**, **N. J. Krogan**, **D. L. Swaney**, **M. P. Snyder**, **S. Mukherjee**, and **M. C. Bassik** (2019). Systematic identification of host cell regulators of legionella pneumophila pathogenesis using a genome-wide CRISPR screen. *Cell Host Microbe*, **26**(4), 551–563.e6.

[297] **Leto, D. E.**, **D. W. Morgens**, **L. Zhang**, **C. P. Walczak**, **J. E. Elias**, **M. C. Bassik**, and **R. R. Kopito** (2019). Genome-wide CRISPR analysis identifies Substrate-Specific conjugation modules in ER-Associated degradation. *Mol. Cell*, **73**(2), 377–389.e11.

[298] **Liu, J.**, **S. Srinivasan**, **C.-Y. Li**, **I.-L. Ho**, **J. Rose**, **M. Shaheen**, **G. Wang**, **W. Yao**, **A. Deem**, **C. Bristow**, **T. Hart**, and **G. Draetta** (2019). Pooled library screening with multiplexed cpf1 library. *Nat. Commun.*, **10**(1), 3144.

[299] **Schinzel, R. T.**, **R. Higuchi-Sanabria**, **O. Shalem**, **E. A. Moehle**, **B. M. Webster**, **L. Joe**, **R. Bar-Ziv**, **P. A. Frankino**, **J. Durieux**, **C. Pender**, **N. Kelet**, **S. S. Kumar**, **N. Savalia**, **H. Chi**, **M. Simic**, **N.-T. Nguyen**, and **A. Dillin** (2019). The hyaluronidase, TMEM2, promotes ER homeostasis and longevity independent of the UPR. *Cell*, **179**(6), 1306–1318.e18.

[300] **Liu, Y.**, **Z. Cao**, **Y. Wang**, **Y. Guo**, **P. Xu**, **P. Yuan**, **Z. Liu**, **Y. He**, and **W. Wei** (2018). Genome-wide screening for functional long noncoding RNAs in human cells by cas9 targeting of splice sites. *Nat. Biotechnol.*.

[301] **Liberzon, A.**, **A. Subramanian**, **R. Pinchback**, **H. Thorvaldsdóttir**, **P. Tamayo**, and **J. P. Mesirov** (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, **27**(12).

[302] **de Hond, A. A. H.**, **E. W. Steyerberg**, and **B. van Calster** (2022). Interpreting area under the receiver operating characteristic curve. *The Lancet Digital Health*, **4**(12), e853–e855.

[303] **Hendricks, K. B.**, **F. Shanahan**, and **E. Lees** (2004). Role for BRG1 in cell cycle control and tumor suppression. *Mol. Cell. Biol.*, **24**(1), 362–376.

[304] **Yang, L.**, **K. Huang**, **X. Li**, **M. Du**, **X. Kang**, **X. Luo**, **L. Gao**, **C. Wang**, **Y. Zhang**, **C. Zhang**, **Q. Tong**, **K. Huang**, **F. Zhang**, and **D. Huang** (2013). Identification of poly(ADP-ribose) polymerase-1 as a cell cycle regulator through modulating sp1 mediated transcription in human hepatoma cells. *PLoS One*, **8**(12), e82872.

[305] **Hyle, J.**, **Y. Zhang**, **S. Wright**, **B. Xu**, **Y. Shao**, **J. Easton**, **L. Tian**, **R. Feng**, **P. Xu**, and **C. Li** (2019). Acute depletion of CTCF directly affects MYC regulation through loss of enhancer-promoter looping. *Nucleic Acids Res.*, **47**(13), 6699–6713.

[306] **Zhang, H.**, **J. Lam**, **D. Zhang**, **Y. Lan**, **M. W. Vermunt**, **C. A. Keller**, **B. Giardine**, **R. C. Hardison**, and **G. A. Blobel** (2021). CTCF and transcription influence chromatin structure re-configuration after mitosis. *Nat. Commun.*, **12**(1), 5157.

[307] **Bakhmet, E. I.** and **A. N. Tomilin** (2021). Key features of the POU transcription factor oct4 from an evolutionary perspective. *Cell. Mol. Life Sci.*, **78**(23), 7339–7353.

[308] **Kenny, C.**, **E. O'Meara**, **M. Ulaş**, **K. Hokamp**, and **M. J. O'Sullivan** (2021). Global chromatin changes resulting from Single-Gene Inactivation-The role of SMARCB1 in malignant rhabdoid tumor. *Cancers*, **13**(11).

[309] **Meurer, L.**, **L. Ferdman**, **B. Belcher**, and **T. Camarata** (2021). The SIX family of transcription factors: Common themes integrating developmental and cancer biology. *Front Cell Dev Biol*, **9**, 707854.

[310] **Blake, J. A.**, **R. Baldarelli**, **J. A. Kadin**, **J. E. Richardson**, **C. L. Smith**, and **C. J. Bult** (2021). Mouse genome database (MGD): Knowledgebase for mouse-human comparative biology. *Nucleic Acids Res.*, **49**(D1).

[311] **Krall, E. B.**, **B. Wang**, **D. M. Munoz**, **N. Ilic**, **S. Raghavan**, **M. J. Niederst**, **K. Yu**, **D. A. Ruddy**, **A. J. Aguirre**, **J. W. Kim**, **A. J. Redig**, **J. F. Gainor**, **J. A. Williams**, **J. M. Asara**, **J. G. Doench**, **P. A. Janne**, **A. T. Shaw**, **R. E. McDonald, Iii**, **J. A. Engelman**, **F. Stegmeier**, **M. R. Schlabach**, and **W. C. Hahn** (2017). KEAP1 loss modulates sensitivity to kinase targeted therapy in lung cancer. *Elife*, **6**.

[312] **Alimov, I.**, **S. Menon**, **N. Cochran**, **R. Maher**, **Q. Wang**, **J. Alford**, **J. B. Concannon**, **Z. Yang**, **E. Harrington**, **L. Llamas**, **A. Lindeman**, **G. Hoffman**, **T. Schuhmann**, **C. Russ**, **J. Reece-Hoyes**, **S. M. Canham**, and **X. Cai** (2019). Bile acid analogues are activators of pyrin inflammasome. *J. Biol. Chem.*, **294**(10), 3359–3366.

[313] **Rosales, C.** and **E. Uribe-Querol** (2017). Phagocytosis: A fundamental process in immunity. *Biomed Res. Int.*, **2017**, 9042851.

[314] **Chesmore, K. N.**, **J. Bartlett**, **C. Cheng**, and **S. M. Williams** (2016). Complex patterns of association between pleiotropy and transcription factor evolution. *Genome Biol. Evol.*, **8**(10), 3159–3170.

[315] **Breiman, L.** (2001). Random forests. *Mach. Learn.*, **45**(1), 5–32.

[316] **Wang, Z.**, **B.-Y. Liao**, and **J. Zhang** (2010). Genomic patterns of pleiotropy and the evolution of complexity. *Proc. Natl. Acad. Sci. U. S. A.*, **107**(42), 18034–18039.

[317] **Kim, S.**, **N.-K. Yu**, and **B.-K. Kaang** (2015). CTCF as a multifunctional protein in genome regulation and gene expression. *Exp. Mol. Med.*, **47**(6), e166.

[318] **Zhu, L. J.**, **C. Gazin**, **N. D. Lawson**, **H. Pagès**, **S. M. Lin**, **D. S. Lapointe**, and **M. R. Green** (2010). Chippeakanno: a bioconductor package to annotate chip-seq and chip-chip data. *BMC bioinformatics*, **11**, 1–10.

[319] **Chen, Y.-X.**, **J. Ding**, **W.-E. Zhou**, **X. Zhang**, **X.-T. Sun**, **X.-Y. Wang**, **C. Zhang**, **N. Li**, **G.-F. Shao**, **S.-J. Hu**, and **J. Yang** (2021). Identification and functional prediction of long Non-Coding RNAs in dilated cardiomyopathy by bioinformatics analysis. *Front. Genet.*, **12**, 648111.

[320] **Raulerson, C. K.**, **A. Ko**, **J. C. Kidd**, **K. W. Currin**, **S. M. Brotman**, **M. E. Cannon**, **Y. Wu**, **C. N. Spracklen**, **A. U. Jackson**, **H. M. Stringham**, **R. P. Welch**, **C. Fuchsberger**, **A. E. Locke**, **N. Narisu**, **A. J. Lusis**, **M. Civelek**, **T. S. Furey**, **J. Kuusisto**, **F. S. Collins**, **M. Boehnke**, **L. J. Scott**, **D.-Y. Lin**, **M. I. Love**, **M. Laakso**, **P. Pajukanta**, and **K. L. Mohlke** (2019). Adipose tissue gene expression associations reveal hundreds of candidate genes for cardiometabolic traits. *Am. J. Hum. Genet.*, **105**(4), 773–787.

[321] **Elaine Hardman, W.**, **D. A. Primerano**, **M. T. Legenza**, **J. Morgan**, **J. Fan**, and **J. Denvir** (2019). mRNA expression data in breast cancers before and after consumption of walnut by women. *Data Brief*, **25**, 104050.

[322] **Donato, L.**, **C. Scimone**, **S. Alibrandi**, **C. Rinaldi**, **A. Sidoti**, and **R. D'Angelo** (2020). Transcriptome analyses of lncRNAs in A2E-Stressed retinal epithelial cells unveil advanced links between metabolic impairments related to oxidative stress and retinitis pigmentosa. *Antioxidants (Basel)*, **9**(4).

[323] **Li, Z.**, **S. Cai**, **H. Li**, **J. Gu**, **Y. Tian**, **J. Cao**, **D. Yu**, and **Z. Tang** (2021). Developing a lncRNA signature to predict the radiotherapy response of Lower-Grade gliomas using co-expression and ceRNA network analysis. *Front. Oncol.*, **11**, 622880.

[324] **Zhu, P.**, **J. Pan**, **Q. Q. Cai**, **F. Zhang**, **M. Peng**, **X. L. Fan**, **H. Ji**, **Y. W. Dong**, **X. Z. Wu**, and **L. H. Wu** (2022). MicroRNA profile as potential molecular signature for attention deficit hyperactivity disorder in children. *Biomarkers*, 1–10.

[325] **Sánchez-Jiménez, C.**, **I. Carrascoso**, **J. Barrero**, and **J. M. Izquierdo** (2013). Identification of a set of miRNAs differentially expressed in transiently TIA-depleted HeLa cells by genome-wide profiling. *BMC Mol. Biol.*, **14**, 4.

[326] **Sage, A. P.**, **K. W. Ng**, **E. A. Marshall**, **G. L. Stewart**, **B. C. Minatel**, **K. S. S. Enfield**, **S. D. Martin**, **C. J. Brown**, **N. Abraham**, and **W. L. Lam** (2020). Assessment of long non-coding RNA expression reveals novel mediators of the lung tumour immune response. *Sci. Rep.*, **10**(1), 16945.

[327] **Whittington, C. M.**, **D. O'Meally**, **M. K. Laird**, **K. Belov**, **M. B. Thompson**, and **B. M. McAllan** (2018). Transcriptomic changes in the pre-implantation uterus highlight histotrophic nutrition of the developing marsupial embryo. *Sci. Rep.*, **8**(1), 2412.

[328] **Chen, J.**, **Y. Wang**, **C. Wang**, **J.-F. Hu**, and **W. Li** (2020). LncRNA functions as a new emerging epigenetic factor in determining the fate of stem cells. *Front. Genet.*, **11**, 277.

[329] **Khurana, E.**, **Y. Fu**, **D. Chakravarty**, **F. Demichelis**, **M. A. Rubin**, and **M. Gerstein** (2016). Role of non-coding sequence variants in cancer. *Nat. Rev. Genet.*, **17**(2), 93–108.

[330] **Wu, W.-S.** and **F.-J. Lai** (2016). Detecting cooperativity between transcription factors based on functional coherence and similarity of their target gene sets. *PLoS One*, **11**(9), e0162931.

[331] **Cao, M.**, **R. Ma**, **H. Li**, **J. Cui**, **C. Zhang**, and **J. Zhao** (2022). Therapy-resistant and -sensitive lncRNAs, SNHG1 and UBL7-AS1 promote glioblastoma cell proliferation. *Oxid. Med. Cell. Longev.*, **2022**, 2623599.

[332] **Rui, X.**, **Y. Xu**, **Y. Huang**, **L. Ji**, and **X. Jiang** (2018). lncRNA DLG1-AS1 promotes cell proliferation by competitively binding with mir-107 and Up-Regulating ZHX1 expression in cervical cancer. *Cell. Physiol. Biochem.*, **49**(5), 1792–1803.

[333] **Zhou, J.**, **J. Shi**, **X. Fu**, **B. Mao**, **W. Wang**, **W. Li**, **G. Li**, and **S. Zhou** (2018). Linc00441 interacts with DNMT1 to regulate RB1 gene methylation and expression in gastric cancer. *Oncotarget*, **9**(101), 37471–37479.

[334] **Du, Y.**, **H. Yang**, **Y. Li**, **W. Guo**, **Y. Zhang**, **H. Shen**, **L. Xing**, **Y. Li**, **W. Wu**, and **X. Zhang** (2021). Long non-coding RNA LINC01137 contributes to oral squamous cell carcinoma development and is negatively regulated by mir-22-3p. *Cell. Oncol.*, **44**(3), 595–609.

[335] **Wang, S.**, **H. Sun**, **J. Ma**, **C. Zang**, **C. Wang**, **J. Wang**, **Q. Tang**, **C. A. Meyer**, **Y. Zhang**, and **X. S. Liu** (2013). Target analysis by integration of transcriptome and ChIP-seq data with BETA. *Nat. Protoc.*, **8**(12), 2502–2515.

[336] **Reshef, Y. A.**, **H. K. Finucane**, **D. R. Kelley**, **A. Gusev**, **D. Kotliar**, **J. C. Ulirsch**, **F. Hormozdiari**, **J. Nasser**, **L. O'Connor**, **B. van de Geijn**, **P.-R. Loh**, **S. R. Grossman**, **G. Bhatia**, **S. Gazal**, **P. F. Palamara**, **L. Pinello**, **N. Patterson**, **R. P. Adams**, and **A. L. Price** (2018). Detecting genome-wide directional effects of transcription factor binding on polygenic disease risk. *Nat. Genet.*, **50**(10), 1483–1493.

[337] **Roopra, A.** (2020). MAGIC: A tool for predicting transcription factors and cofactors driving gene sets using ENCODE data. *PLoS Comput. Biol.*, **16**(4), e1007800.

[338] **Dainese, R.**, **V. Gardeux**, **G. Llimos**, **D. Alpern**, **J. Y. Jiang**, **A. C. A. Meireles-Filho**, and **B. Deplancke** (2020). A parallelized, automated platform enabling individual or sequential ChIP of histone marks and transcription factors. *Proc. Natl. Acad. Sci. U. S. A.*, **117**(24), 13828–13838.

[339] **Gopalan, S.** and **T. G. Fazzio** (2022). Multi-CUT&Tag to simultaneously profile multiple chromatin factors. *STAR Protoc*, **3**(1), 101100.

[340] **van Ouwerkerk, A. F.**, **A. W. Hall**, **Z. A. Kadow**, **S. Lazarevic**, **J. S. Reyat**, **N. R. Tucker**, **R. D. Nadadur**, **F. M. Bosada**, **V. Bianchi**, **P. T. Ellinor**, **L. Fabritz**, **J. F. Martin**, **W. de Laat**, **P. Kirchhof**, **I. P. Moskowitz**, and **V. M. Christoffels** (2020). Epigenetic and transcriptional networks underlying atrial fibrillation. *Circ. Res.*, **127**(1), 34–50.

[341] **Mazzone, R.**, **C. Zwergel**, **M. Artico**, **S. Taurone**, **M. Ralli**, **A. Greco**, and **A. Mai** (2019). The emerging role of epigenetics in human autoimmune disorders. *Clin. Epigenetics*, **11**(1), 34.

[342] **Su, K.**, **A. Katebi**, **V. Kohar**, **B. Clauss**, **D. Gordin**, **Z. S. Qin**, **R. K. M. Karuturi**, **S. Li**, and **M. Lu** (2022). NetAct: a computational platform to construct core transcription factor regulatory networks using gene activity. *Genome Biol.*, **23**(1), 270.

[343] **Li, S.**, **B. Yan**, **B. Wu**, **J. Su**, **J. Lu**, **T.-W. Lam**, **K. R. Boheler**, **E. N.-Y. Poon**, and **R. Luo** (2023). Integrated modeling framework reveals co-regulation of transcription factors, miRNAs and lncRNAs on cardiac developmental dynamics. *Stem Cell Res. Ther.*, **14**(1), 247.

[344] **Chandra, O.**, **M. Sharma**, **N. Pandey**, **I. P. Jha**, **S. Mishra**, **S. L. Kong**, and **V. Kumar** (2023). Patterns of transcription factor binding and epigenome at promoters allow interpretable predictability of multiple functions of non-coding and coding genes. *Comput. Struct. Biotechnol. J.*, **21**, 3590–3603.

[345] **Kaikkonen, M. U.**, **M. T. Y. Lam**, and **C. K. Glass** (2011). Non-coding RNAs as regulators of gene expression and epigenetics. *Cardiovasc. Res.*, **90**(3), 430–440.

[346] **Nam, Y.**, **J. H. Jhee**, **J. Cho**, **J.-H. Lee**, and **H. Shin** (2019). Disease gene identification based on generic and disease-specific genome networks. *Bioinformatics*, **35**(11), 1923–1930.

[347] **Shah, S. D.** and **R. Braun** (2019). GeneSurrounder: network-based identification of disease genes in expression data. *BMC Bioinformatics*, **20**(1), 229.

[348] **Qumsiyeh, E.**, **L. Showe**, and **M. Yousef** (2022). GediNET for discovering gene associations across diseases using knowledge based machine learning approach. *Sci. Rep.*, **12**(1), 19955.

[349] **Porcu, E.**, **M. C. Sadler**, **K. Lepik**, **C. Auwerx**, **A. R. Wood**, **A. Weihs**, **M. S. B. Sleiman**, **D. M. Ribeiro**, **S. Bandinelli**, **T. Tanaka**, **M. Nauck**, **U. Völker**, **O. Delaneau**, **A. Metspalu**, **A. Teumer**, **T. Frayling**, **F. A. Santoni**, **A. Reymond**, and **Z. Kutalik** (2021). Differentially expressed genes reflect disease-induced rather than disease-causing changes in the transcriptome. *Nat. Commun.*, **12**(1), 5647.

[350] **Chen, B.**, **J. Wang**, **M. Li**, and **F.-X. Wu** (2014). Identifying disease genes by integrating multiple data sources. *BMC Med. Genomics*, **7 Suppl 2**(Suppl 2), S2.

[351] **Asif, M.**, **H. F. M. C. M. Martiniano**, **A. M. Vicente**, and **F. M. Couto** (2018). Identifying disease genes using machine learning and gene functional similarities, assessed through gene ontology. *PLoS One*, **13**(12), e0208626.

[352] **Cancer Genome Atlas Research Network**, **J. N. Weinstein**, **E. A. Collisson**, **G. B. Mills**, **K. R. M. Shaw**, **B. A. Ozenberger**, **K. Ellrott**, **I. Shmulevich**, **C. Sander**, and **J. M. Stuart** (2013). The cancer genome atlas Pan-Cancer analysis project. *Nat. Genet.*, **45**(10), 1113–1120.

[353] **Tang, Z.**, **B. Kang**, **C. Li**, **T. Chen**, and **Z. Zhang** (2019). GEPIA2: an enhanced web server for large-scale expression profiling and interactive analysis. *Nucleic Acids Res.*, **47**(W1), W556–W560.

[354] **Sollis, E.**, **A. Mosaku**, **A. Abid**, **A. Buniello**, **M. Cerezo**, **L. Gil**, **T. Groza**, **O. Güneş**, **P. Hall**, **J. Hayhurst**, **A. Ibrahim**, **Y. Ji**, **S. John**, **E. Lewis**, **J. A. L. MacArthur**, **A. McMahon**, **D. Osumi-Sutherland**, **K. Panoutsopoulou**, **Z. Pendlington**, **S. Ramachandran**, **R. Stefancsik**, **J. Stewart**, **P. Whetzel**, **R. Wilson**, **L. Hindorff**, **F. Cunningham**, **S. A. Lambert**, **M. Inouye**, **H. Parkinson**, and **L. W. Harris** (2023). The NHGRI-EBI GWAS catalog: knowledgebase and deposition resource. *Nucleic Acids Res.*, **51**(D1), D977–D985.

[355] **Kim, S. H.**, **J.-Y. Choe**, **Y. Jeon**, **J. Huh**, **H. R. Jung**, **Y.-D. Choi**, **H.-J. Kim**, **H. J. Cha**, **W. S. Park**, and **J. E. Kim** (2013). Frequent expression of follicular dendritic cell markers in hodgkin lymphoma and anaplastic large cell lymphoma. *J. Clin. Pathol.*, **66**(7), 589–596.

[356] **Bhardwaj, N.** and **J. D. Brody** (2015). Dendritic cells and lymphoma cells: come together right now. *Blood*, **125**(1), 5–7.

[357] **Mohty, M.**, **D. Jarrossay**, **M. Lafage-Pochitaloff**, **C. Zandotti**, **F. Brière**, **X. N. de Lamballeri**, **D. Isnardon**, **D. Sainty**, **D. Olive**, and **B. Gaugler** (2001). Circulating blood dendritic cells from myeloid leukemia patients display quantitative and cytogenetic abnormalities as well as functional impairment. *Blood*, **98**(13), 3750–3756.

[358] **Hattangadi, S. M.**, **P. Wong**, **L. Zhang**, **J. Flygare**, and **H. F. Lodish** (2011). From stem cell to red cell: regulation of erythropoiesis at multiple levels by multiple proteins, RNAs, and chromatin modifications. *Blood*, **118**(24), 6258–6268.

[359] **Peled, N.**, **M. Kassirer**, **M. R. Kramer**, **O. Rogowski**, **D. Shlomi**, **B. Fox**, **A. S. Berliner**, and **D. Shitrit** (2008). Increased erythrocyte adhesiveness and aggregation in obstructive sleep apnea syndrome. *Thromb. Res.*, **121**(5), 631–636.

[360] **Narváez, P. A.**, **C. J. Mohrenberger**, **E. M. Baena**, **C. G. Rivera**, **R. M. Villalona**, **P. L. Meneses**, **D. B. Barreto**, **E. P. Rodriguez**, and **J. G. de Leaniz** (2014). Erythrocytosis in patients with obstructive sleep apnea. *Eur. Respir. J.*, **44**(Suppl 58).

[361] **Chen, Z.**, **S. Chen**, and **J. Liu** (2018). The role of T cells in the pathogenesis of parkinson's disease. *Prog. Neurobiol.*, **169**, 1–23.

[362] **Contaldi, E.**, **L. Magistrelli**, and **C. Comi** (2022). T lymphocytes in parkinson's disease. *J. Parkinsons. Dis.*, **12**(s1), S65–S74.

[363] **Roussy, M.**, **M. Bilodeau**, **L. Jouan**, **P. Tibout**, **L. Laramée**, **E. Lemyre**, **F. Léveillé**, **F. Tihy**, **S. Cardin**, **C. Sauvageau**, *et al.* (2018). Nup98-bptf gene fusion identified in primary refractory acute megakaryoblastic leukemia of infancy. *Genes, Chromosomes and Cancer*, **57**(6), 311–319.

[364] **Chang, K.-C.**, **G.-C. Huang**, **D. Jones**, and **Y.-H. Lin** (2007). Distribution patterns of dendritic cells and t cells in diffuse large b-cell lymphomas correlate with prognoses. *Clinical cancer research*, **13**(22), 6666–6672.

[365] **Saad, M. N.**, **M. S. Mabrouk**, **A. M. Eldeib**, and **O. G. Shaker** (2019). Studying the effects of haplotype partitioning methods on the ra-associated genomic results from the north american rheumatoid arthritis consortium (narac) dataset. *Journal of Advanced Research*, **18**, 113–126.

[366] **Catapano, M.**, **M. Vergnano**, **M. Romano**, **S. K. Mahil**, **S.-E. Choon**, **A. D. Burden**, **H. S. Young**, **I. M. Carr**, **H. J. Lachmann**, **G. Lombardi**, *et al.* (2020). Il-36 promotes systemic ifn-i responses in severe forms of psoriasis. *Journal of Investigative Dermatology*, **140**(4), 816–826.

[367] **Bedrosian, A. S.**, **A. H. Nguyen**, **M. Hackman**, **M. K. Connolly**, **A. Malhotra**, **J. Ibrahim**, **N. E. Cieza-Rubio**, **J. R. Henning**, **R. Barilla**, **A. Rehman**, **H. L. Pachter**, **M. V. Medina-Zea**, **S. M. Cohen**, **A. B. Frey**, **D. Acehan**, and **G. Miller** (2011). Dendritic cells promote pancreatic viability in mice with acute pancreatitis. *Gastroenterology*, **141**(5), 1915–26.e1–14.

[368] **McKeithen, D. N.**, **Y. O. Omosun**, **K. Ryans**, **J. Mu**, **Z. Xie**, **T. Simoneaux**, **U. Blas-Machado**, **F. O. Eko**, **C. M. Black**, **J. U. Igietseme**, *et al.* (2017). The emerging role of asc in dendritic cell metabolism during chlamydia infection. *PLoS One*, **12**(12), e0188643.

[369] **Lessard, S.**, **E. S. Gatof**, **M. Beaudoin**, **P. G. Schupp**, **F. Sher**, **A. Ali**, **S. Prehar**, **R. Kurita**, **Y. Nakamura**, **E. Baena**, *et al.* (2017). An erythroid-specific atp2b4 enhancer mediates red blood cell hydration and malaria susceptibility. *The Journal of clinical investigation*, **127**(8), 3065–3074.

[370] **Wu, X.** and **F. Xu** (2023). Dendritic cell-based immunotherapy: A potential therapeutic option for chronic hepatitis b virus infection. *Clinical and Translational Discovery*, **3**(1), e173.

[371] **Love, P. E.**, **C. Warzecha**, and **L. Li** (2014). Ldb1 complexes: the new master regulators of erythroid gene transcription. *Trends in Genetics*, **30**(1), 1–9.

[372] **Miyai, M.**, **M. Hamada**, **T. Moriguchi**, **J. Hiruma**, **A. Kamitani-Kawamoto**, **H. Watanabe**, **M. Hara-Chikuma**, **K. Takahashi**, **S. Takahashi**, and **K. Kataoka** (2016). Transcription factor mafb coordinates epidermal keratinocyte differentiation. *Journal of Investigative Dermatology*, **136**(9), 1848–1857.

[373] **Ruzhynsky, V. A.**, **M. Furimsky**, **D. S. Park**, **V. A. Wallace**, and **R. S. Slack** (2009). E2f4 is required for early eye patterning. *Developmental Neuroscience*, **31**(3), 238–246.

[374] **de Melo, J.**, **C. Zibetti**, **B. S. Clark**, **W. Hwang**, **A. L. Miranda-Angulo**, **J. Qian**, and **S. Blackshaw** (2016). Lhx2 is an essential factor for retinal gliogenesis and notch signaling. *Journal of Neuroscience*, **36**(8), 2391–2405.

[375] **Han, H.**, **H. Qin**, **Y. Yang**, **L. Zhao**, **T. Shen**, and **Q. Pang** (2023). Effect of overexpression of klf4 on the growth and development of hair follicles in mice. *Development Genes and Evolution*, **233**(2), 137–145.

[376] **Inoue, Y.**, **C.-W. Liao**, **Y. Tsunakawa**, **I.-L. Tsai**, **S. Takahashi**, and **M. Hamada** (2022). Macrophage-specific, mafb-deficient mice showed delayed skin wound healing. *International Journal of Molecular Sciences*, **23**(16), 9346.

[377] **Yan, Q.**, **B.-j. Chen**, **S. Hu**, **S.-l. Qi**, **L.-y. Li**, **J.-f. Yang**, **H. Zhou**, **C.-c. Yang**, **L.-j. Chen**, and **J. Du** (2021). Emerging role of rnf2 in cancer: from bench to bedside. *Journal of cellular physiology*, **236**(8), 5453–5465.

[378] **Kim, H.**, **S.-H. Lee**, **M.-N. Lee**, **G. T. Oh**, **K.-C. Choi**, and **E. Y. Choi** (2013). p53 regulates the transcription of the anti-inflammatory molecule developmental endothelial locus-1 (del-1). *Oncotarget*, **4**(11), 1976–1985.

[379] **Lee, S.-H.**, **D.-Y. Kim**, **F. Jing**, **H. Kim**, **C.-O. Yun**, **D.-J. Han**, and **E. Y. Choi** (2015). Del-1 overexpression potentiates lung cancer cell proliferation and invasion. *Biochem. Biophys. Res. Commun.*, **468**(1-2), 92–98.

[380] **Lee, S. J.**, **J.-H. Jeong**, **S. H. Kang**, **J. Kang**, **E. A. Kim**, **J. Lee**, **J. H. Jung**, **H. Y. Park**, and **Y. S. Chae** (2019). MicroRNA-137 inhibits cancer progression by targeting del-1 in Triple-Negative breast cancer cells. *Int. J. Mol. Sci.*, **20**(24).

[381] **Lécuyer, H.**, **D. Borgel**, **X. Nassif**, and **M. Coureuil** (2017). Pathogenesis of meningococcal purpura fulminans. *Pathog. Dis.*, **75**(3).

[382] **Peiser, L.**, **M. P. J. De Winther**, **K. Makepeace**, **M. Hollinshead**, **P. Coull**, **J. Plested**, **T. Kodama**, **E. R. Moxon**, and **S. Gordon** (2002). The class a macrophage scavenger receptor is a major pattern recognition receptor for neisseria meningitidis which is independent of lipopolysaccharide and not required for secretory responses. *Infect. Immun.*, **70**(10), 5346–5354.

[383] **Catapano, M.**, **M. Vergnano**, **M. Romano**, **S. K. Mahil**, **S.-E. Choon**, **A. D. Burden**, **H. S. Young**, **I. M. Carr**, **H. J. Lachmann**, **G. Lombardi**, **C. H. Smith**, **F. D. Ciccarelli**, **J. N. Barker**, and **F. Capon** (2020). IL-36 promotes systemic IFN-I responses in severe forms of psoriasis. *J. Invest. Dermatol.*, **140**(4), 816–826.e3.

[384] **Burden, A. D.** and **B. Kirby** (2016). Psoriasis and related disorders.

[385] **Zhang, S.**, **H. D. Coughlan**, **M. Ashayeripanah**, **S. Seizova**, **A. J. Kueh**, **D. V. Brown**, **W. Cao**, **N. Jacquelot**, **A. D'Amico**, **A. M. Lew**, **Y. Zhan**, **C. J. Tonkin**, **J. A. Villadangos**, **G. K. Smyth**, **M. Chopin**, and **S. L. Nutt** (2021). Type 1 conventional dendritic cell fate and function are controlled by DC-SCRIPT. *Sci Immunol*, **6**(58).

[386] **Giaccherini, M.**, **R. Farinella**, **M. Gentiluomo**, **B. Mohelnikova-Duchonova**, **E. F. Kauffmann**, **M. Palmeri**, **F. Uzunoglu**, **P. Soucek**, **D. Petrauskas**, **G. M. Cavestro**, **R. Zykus**, **S. Carrara**, **R. Pezzilli**, **M. Puzzono**, **A. Szentesi**, **J. Neoptolemos**, **L. Archibugi**, **O. Palmieri**, **A. C. Milanetto**, **G. Capurso**, **C. H. J. van Eijck**, **H. Stocker**, **R. T. Lawlor**, **P. Vodicka**, **M. Lovecek**, **J. R. Izbicki**, **F. Perri**, **R. Kupcinskaite-Noreikiene**, **M. Götz**, **J. Kupcinskas**, **T. Hussein**, **P. Hegyi**, **O. R. Busch**, **T. Hackert**, **A. Mambrini**, **H. Brenner**, **M. Lucchesi**, **D. Basso**, **F. Tavano**, **B. Schöttker**, **G. Vanella**, **S. Bunduc**, **Á. Petrányi**, **S. Landi**, **L. Morelli**, **F. Canzian**, and **D. Campa** (2023). Association between a polymorphic variant in the CDKN2B-AS1/ANRIL gene and pancreatic cancer risk. *Int. J. Cancer*, **153**(2), 373–379.

[387] **Luo, J.**, **R. Bai**, **Y. Liu**, **H. Bi**, **X. Shi**, and **C. Qu** (2022). Long non-coding RNA ATXN8OS promotes ferroptosis and inhibits the temozolomide-resistance of gliomas through the ADAR/GLS2 pathway. *Brain Res. Bull.*, **186**, 27–37.

[388] **Haro, M. A.**, **A. M. Dyevoich**, **J. P. Phipps**, and **K. M. Haas** (2019). Activation of B-1 cells promotes tumor cell killing in the peritoneal cavity. *Cancer Res.*, **79**(1), 159–170.

[389] **Laidlaw, B. J.** and **J. G. Cyster** (2021). Transcriptional regulation of memory B cell differentiation. *Nat. Rev. Immunol.*, **21**(4), 209–220.

[390] **Frissora, F.**, **H.-C. Chen**, **J. Durbin**, **S. Bondada**, and **N. Muthusamy** (2003). IFN-gamma-mediated inhibition of antigen receptor-induced B cell proliferation and CREB-1 binding activity requires STAT-1 transcription factor. *Eur. J. Immunol.*, **33**(4), 907–912.

[391] **Yang, C.**, **H. Lee**, **S. Pal**, **V. Jove**, **J. Deng**, **W. Zhang**, **D. S. B. Hoon**, **M. Wakabayashi**, **S. Forman**, and **H. Yu** (2013). B cells promote tumor progression via STAT3 regulated-angiogenesis. *PLoS One*, **8**(5), e64159.

[392] **Andrews, A. J.** and **K. Luger** (2011). Nucleosome structure(s) and stability: Variations on a theme. *Annu. Rev. Biophys.*, **40**(Volume 40, 2011), 99–117.

[393] **Miles, B.** and **P. Tadi** (2024). Genetics, somatic mutation.

[394] **Duan, S.** and **M. Pagano** (2011). Linking metabolism and cell cycle progression via the APC/CCdh1 and SCFβTrCP ubiquitin ligases. *Proc. Natl. Acad. Sci. U. S. A.*, **108**(52), 20857–20858.

[395] (2009). Different gene regulation strategies revealed by analysis of binding motifs. *Trends Genet.*, **25**(10), 434–440.

[396] **Pearson, H.** (2006). What is a gene? http://dx.doi.org/10.1038/441398a. Accessed: 2024-3-15.

[397] **Leal-Esteban, L. C.** and **L. Fajas** (2020). Cell cycle regulators in cancer cell metabolism. *Biochim. Biophys. Acta Mol. Basis Dis.*, **1866**(5), 165715.

[398] **Russell, S.**, **J. Bennett**, **J. A. Wellman**, **D. C. Chung**, **Z.-F. Yu**, **A. Tillman**, **J. Wittes**, **J. Pappas**, **O. Elci**, **S. McCague**, **D. Cross**, **K. A. Marshall**, **J. Walshire**, **T. L. Kehoe**, **H. Reichert**, **M. Davis**, **L. Raffini**, **L. A. George**, **F. P. Hudson**, **L. Dingfield**, **X. Zhu**, **J. A. Haller**, **E. H. Sohn**, **V. B. Mahajan**, **W. Pfeifer**, **M. Weckmann**, **C. Johnson**, **D. Gewaily**, **A. Drack**, **E. Stone**, **K. Wachtel**, **F. Simonelli**, **B. P. Leroy**, **J. F. Wright**, **K. A. High**, and **A. M. Maguire** (2017). Efficacy and safety of voretigene neparvovec (AAV2-hRPE65v2) in patients with RPE65-mediated inherited retinal dystrophy: a randomised, controlled, open-label, phase 3 trial. *Lancet*, **390**(10097), 849–860.

[399] **Haberle, V.** and **A. Stark** (2018). Eukaryotic core promoters and the functional basis of transcription initiation. *Nat. Rev. Mol. Cell Biol.*, **19**(10), 621.

[400] **Patil, V.**, **R. L. Ward**, and **L. B. Hesson** (2014). The evidence for functional non-CpG methylation in mammalian cells. *Epigenetics*, **9**(6), 823.

[401] (). Burden, a.d. and kirby, b. (2016) psoriasis and related disorders. in griffiths, c., barker, j., bleiker, t., chalmers, r. and creamer, d., eds., rook's textbook of dermatology, 9th edition, john wiley & sons, new delhi, 1-48. - references - scientific research publishing. https://www.scirp.org/reference/referencespapers?referenceid=2521316. Accessed: 2024-6-6.

[402] **Kristensen, L. S.**, **M. S. Andersen**, **L. V. W. Stagsted**, **K. K. Ebbesen**, **T. B. Hansen**, and **J. Kjems** (2019). The biogenesis, biology and characterization of circular RNAs. *Nat. Rev. Genet.*, **20**(11).

[403] **Arner, E.**, **C. O. Daub**, **K. Vitting-Seerup**, **R. Andersson**, **B. Lilje**, **F. Drabløs**, **A. Lennartsson**, **M. Rönnerblad**, **O. Hrydziuszko**, **M. Vitezic**, **T. C. Freeman**, **A. M. N. Alhendi**, **P. Arner**, **R. Axton**, **J. Kenneth Baillie**, **A. Beckhouse**, **B. Bodega**, **J. Briggs**, **F. Brombacher**, **M. Davis**, **M. Detmar**, **A. Ehrlund**, **M. Endoh**, **A. Eslami**, **M. Fagiolini**, **L. Fairbairn**, **G. J. Faulkner**, **C. Ferrai**, **M. E. Fisher**, **L. Forrester**, **D. Goldowitz**, **R. Guler**, **T. Ha**, **M. Hara**, **M. Herlyn**, **T. Ikawa**, **C. Kai**, **H. Kawamoto**, **L. M. Khachigian**, **S. Peter Klinken**, **S. Kojima**, **H. Koseki**, **S. Klein**, **N. Mejhert**, **K. Miyaguchi**, **Y. Mizuno**, **M. Morimoto**, **K. J. Morris**, **C. Mummery**, **Y. Nakachi**, **S. Ogishima**, **M. Okada-Hatakeyama**, **Y. Okazaki**, **V. Orlando**, **D. Ovchinnikov**, **R. Passier**, **M. Patrikakis**, **A. Pombo**, **X.-Y. Qin**, **S. Roy**, **H. Sato**, **S. Savvi**, **A. Saxena**, **A. Schwegmann**, **D. Sugiyama**, **R. Swoboda**,

H. Tanaka, **A. Tomoiu**, **L. N. Winteringham**, **E. Wolvetang**, **C. Yanagi-Mizuochi**, **M. Yoneda**, **S. Zabierowski**, **P. Zhang**, **I. Abugessaisa**, **N. Bertin**, **A. D. Diehl**, **S. Fukuda**, **M. Furuno**, **J. Harshbarger**, **A. Hasegawa**, **F. Hori**, **S. Ishikawa-Kato**, **Y. Ishizu**, **M. Itoh**, **T. Kawashima**, **M. Kojima**, **N. Kondo**, **M. Lizio**, **T. F. Meehan**, **C. J. Mungall**, **M. Murata**, **H. Nishiyori-Sueki**, **S. Sahin**, **S. Nagao-Sato**, **J. Severin**, **M. J. L. de Hoon**, **J. Kawai**, **T. Kasukawa**, **T. Lassmann**, **H. Suzuki**, **H. Kawaji**, **K. M. Summers**, **C. Wells**, **FANTOM Consortium**, **D. A. Hume**, **A. R. R. Forrest**, **A. Sandelin**, **P. Carninci**, and **Y. Hayashizaki** (2015). Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. *Science*, **347**(6225), 1010.

[404] **Office of the Commissioner** (2020). FDA approves innovative gene therapy to treat pediatric patients with spinal muscular atrophy, a rare disease and leading genetic cause of infant mortality. https://www.fda.gov/news-events/press-announcements/fda-approves-innovative-gene-therapy-treat-pediatric-patients-s Accessed: 2023-8-1.

[405] **Liu, G.**, **Z. Chen**, **I. G. Danilova**, **M. A. Bolkov**, **I. A. Tuzankina**, and **G. Liu** (2018). Identification of mir-200c and miR141-Mediated lncRNA-mRNA crosstalks in Muscle-Invasive bladder cancer subtypes. *Front. Genet.*, **9**, 422.

[406] **Kaplon, J.**, **L. van Dam**, and **D. Peeper** (2015). Two-way communication between the metabolic and cell cycle machineries: the molecular basis. *Cell Cycle*, **14**(13), 2022–2032.

[407] **Bruce, A. W.**, **I. J. Donaldson**, **I. C. Wood**, **S. A. Yerbury**, **M. I. Sadowski**, **M. Chapman**, **B. Göttgens**, and **N. J. Buckley** (2004). Genome-wide analysis of repressor element 1 silencing transcription factor/neuron-restrictive silencing factor (REST/NRSF) target genes. *Proceedings of the National Academy of Sciences*, **101**(28), 10458–10463.

[408] **Rosnoblet, C.**, **D. Legrand**, **D. Demaegd**, **H. Hacine-Gherbi**, **G. de Bettignies**, **R. Bammens**, **C. Borrego**, **S. Duvet**, **P. Morsomme**, **G. Matthijs**, and **F. Foulquier** (2013). Impact of disease-causing mutations on TMEM165 subcellular localization, a recently identified protein involved in CDG-II. *Hum. Mol. Genet.*, **22**(14), 2914–2928.

[409] **Bhan, A.** and **S. S. Mandal** (2014). Long noncoding RNAs: emerging stars in gene regulation, epigenetics and human disease. *ChemMedChem*, **9**(9), 1932–1956.

[410] (). Website. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6388056/. Accessed: NA-NA-NA.

[411] **Nogrady, B.** (2020). How cancer genomics is transforming diagnosis and treatment. http://dx.doi.org/10.1038/d41586-020-00845-4. Accessed: 2023-8-1.

[412] (2020). Toward a mechanistic understanding of DNA methylation readout by transcription factors. *J. Mol. Biol.*, **432**(6), 1801–1815.

LIST OF PAPERS BASED ON THESIS

1. **C. Omkar**, P. Durjay, G. Srishti, S. Jaidev, S. Madhu, N. Dubey, M. Biswarup, and K. Vibhor, Explainable models using transcription factor binding and epigenome patterns at promoters reveal disease-associated genes and their regulators in the context of cell-types, bioRxiv, 2024.
   url: https://www.biorxiv.org/content/10.1101/2024.05.06.592622v1.

2. **C. Omkar**, M. Sharma, P. N, J. IP, M. S, K. SL, and K. Vibhor, Patterns of transcription factor binding and epigenome at promoters allow interpretable predictability of multiple functions of non-coding and coding genes, "Computational and Structural Biotechnology Journal", 2023. url: https://www.sciencedirect.com/science/article/pii/S2001037023002453.

Additional Publications

1. P. Neetesh, **C. Omkar**, M. Shreya, and K. Vibhor, Improving chromatin-interaction prediction using single-cell open-chromatin profiles and making insight into the cis-regulatory landscape of the human brain, Frontiers in Genetics, 2021. url: https://www.frontiersin.org/articles/10.3389/fgene.2021.738194/full.

2. M. Shreya, P. Neetesh, C. Smriti, S. Madhu, **C. Omkar**, P. Indra, S. Debarka, N. Kedar, and K. Vibhor, Matching queried single-cell open-chromatin profiles to large pools of single-cell transcriptomes and epigenomes for reference supported analysis, Genome Research, 2023. url: https://genome.cshlp.org/content/33/2/218.full.

3. N. Pandey, S. Madhu, M. Arpit, N. George Anene, H. Muhammad, J. Indra Prakash, **C. Omkar**, and et al., Deciphering the phenotypic heterogeneity and drug response in cancer cells using genome-wide activity and interaction of chromatin domains., bioRxiv, 2023.
   url: https://www.biorxiv.org/content/10.1101/2023.01.15.524115v1.full.pdf.

   S. Madhu, J. Indra Prakash, C. Smriti, P. Neetesh, **C. Omkar**, M. Shreya, and K. Vibhor, Associating pathways with diseases using single-cell expression profiles and making inferences about potential drugs, Brifings in Bioinformatics, 2022. url: https://academic.oup.com/bib/article/23/4/bbac241/6623725?login=true.