Comprehending the synergistic effects of gene ensembles in the context of disease biology, prognosis, and therapy

By

Madhu Sharma

(PhD19204)

Under the Supervision of Dr. Vibhor Kumar

Department of Computational Biology

Indraprastha Institute of Information Technology, Delhi

New Delhi – 110020

September 2024

Comprehending the synergistic effects of gene ensembles in the context of disease biology, prognosis, and therapy

By

Madhu Sharma

(PhD19204)

A Thesis

Submitted in Partial Fulfillment of the Requirements for the Degree Of

Doctor of Philosophy

Under the Supervision of Dr. Vibhor Kumar

Department of Computational Biology

Indraprastha Institute of Information Technology, Delhi

New Delhi – 110020

September 2024

Certificate

This is to certify that the thesis entitled "**Comprehending the synergistic effects of gene ensembles in the context of disease biology, prognosis, and therapy**" being submitted by **Madhu Sharma** to the Indraprastha Institute of Information Technology Delhi, for the award of the degree of **Doctor of Philosophy**, is an original research work, carried out by him under my supervision. In my opinion, the thesis has reached the standards, fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree/diploma.

September 2024

Dr. Vibhor Kumar

Associate Professor

Indraprastha Institute of Information Technology Delhi

New Delhi - 110020

ACKNOWLEDGMENTS

"Alone we can do so little; together we can do so much." — Helen Keller

The completion of this thesis has been a profound journey, and it would not have been possible without the support and encouragement of several remarkable individuals. I am deeply grateful to all those who have contributed to this endeavor.

First and foremost, I want to thank ***Lord Badri and Lord Kedar*** *for guiding me through this challenging period. Your divine presence has been a source of strength and reassurance, helping me navigate the ups and downs with faith and resilience.*

I am profoundly grateful to my supervisor, ***Dr. Vibhor Kumar****. Your expert guidance and profound insights have been the cornerstone of this research. Your constructive criticism and unwavering support have not only refined my work but also deepened my understanding of the subject. Your dedication to excellence has inspired me to push the boundaries of my own capabilities, and I am incredibly fortunate to have had you as a mentor. You stood by me with unwavering patience and wisdom, truly embodying the role of a mentor and a father figure throughout this journey.*

I am also profoundly thankful to ***Prof. GPS Raghava****, Head of the Department of Computational Biology, IIIT Delhi. Your leadership has been instrumental in creating an environment that promotes academic rigor and innovation. Your support in facilitating resources and opportunities has been crucial in successfully completing this thesis. I express my deepest gratitude to IIIT-D Director* ***Prof. Ranjan Bose*** *for his support and efforts at the facility for the research work. I am also immensely grateful to the members of my evaluation committee,* ***Dr. Gaurav Ahuja*** *and* ***Dr. Arjun Ray****, for their thoughtful critiques and constructive suggestions, which have significantly enhanced the quality of my work. Also, thanks to all the department faculty members,* ***Dr. Sriram K.****,* ***Dr. Ganesh Bagler****,* ***Dr. Tavpritesh Sethi****,* ***Dr. Debarka Sengupta****,* ***Dr. Jaspreet Kaur Dhanjal****,* ***Dr. N. Arul Murugan****,* ***Dr. Tarini Shankar Ghosh*** *for their constant support and help.*

*A special thanks to all the administrative staff of IIIT-D, whose assistance with the logistical aspects of my thesis, from paperwork to scheduling, has been indispensable. Your efficiency and support have greatly facilitated the completion of this project. I would also like to express my deepest gratitude to Indian Council of Medical Research (****ICMR****) for their generous financial support throughout the course of my PhD research.*

My heartfelt thanks extend to my friends and colleagues, who have been an indispensable part of this journey. To my lab seniors, ***Dr. Neetesh Pandey****,* ***Dr. Indra Prakash Jha****,* ***Dr. Priyadarshini Rai****,* ***Dr. Smriti Chawla****,* ***Dr. Shreya Mishra****,* ***Dr. Sarita Poonia****,* ***Dr. Omkar Chandra R.****, your mentorship and guidance have been crucial in navigating the complexities of my research. I deeply appreciate the time and effort you invested in helping me refine my work and overcome challenges. To my lab juniors,* ***Niharika Dubey****,* ***Jaidev Sharma****,* ***Durjay Pramanik****,* ***Karuna Kerketta****, and* ***Akshat Bhatt****, your enthusiasm, hard work, and collaboration have made working*

in the lab a rewarding experience. I would like to extend my special thanks to ***Dr. Srishti Gautam****, whose exceptional support and guidance have been pivotal to the success of my research.*

I would like to extend my heartfelt thanks to my friends from other labs, whose support and encouragement have been invaluable throughout my thesis journey. To ***G L Harika****,* ***Shruti Kaushal****,* ***Samriddhi Gupta****,* ***Abhishek Halder****,* ***Sakshi Gujral****,* ***Mansi Goel****,* ***Shubham Choudhury****,* ***Ritu Tomer****,* ***Nisha Bajiya****,* ***Nishant Kumar****,* ***Gayatri Panda****,* ***Diksha Marhawa****,* ***Sukriti Sacher****,* ***Aayushi Mittal****,* ***Sanjay Kumar Mohanty****,* ***Vishakha Gautam****, your camaraderie and willingness to share knowledge and insights have greatly enriched my research experience.*

I also want to give a big shoutout to my amazing friends who've been there for me throughout this whole thesis adventure, even though they aren't part of my academic or professional world. To ***Surojit Jana****,* ***Manashree Jana****,* ***Saanvi Jana****,* ***Nilakshya Biswas****,* ***Tanushree Paul****,* ***Dr. Arvind Yadav****,* ***Krishna Kapoor****,* ***Satinder Kaur*** *and* ***Sanchit Arora****, you reminded me of the joy and lightness in life, which was exactly what I needed. A special thank you to* ***Rahul Mandoliya****. Your support meant the world to me. From our uplifting conversations to your ability to make me laugh during stressful times, you have truly been a rock. Your presence in my life has been a constant reminder of what matters most and has given me strength when I needed it most.*

I would like to express my deepest gratitude to my late father, ***Mr. Mahesh Sharma****, whose support and encouragement were fundamental throughout my academic journey. Though he is no longer with us, his unwavering belief in my potential and guidance have been a constant source of strength and inspiration. I would also like to dedicate a special acknowledgment to my late grandmother,* ***Mrs. Peetha Devi****. This thesis is a tribute to her memory and the lasting impact she had on my life. I am deeply grateful for the foundation she provided and the strength she instilled in me. I am profoundly grateful to my mother,* ***Mrs. Renu Sharma****, and my siblings,* ***Anmol*** *and* ***Anjali****. Your unwavering support and love have been the cornerstone of my journey throughout this thesis. This achievement is as much yours as it is mine, and I cannot thank you enough for the countless ways in which you have supported me.*

Lastly, I would like to acknowledge anyone else who has contributed to this research and thesis, whether directly or indirectly. Your support, whether through professional guidance or personal encouragement, has been deeply appreciated.

Thank you all for being an integral part of this journey. Your collective contributions have been crucial to the completion of this thesis, and I am eternally grateful for your support.

Madhu Sharma
PhD19204

ABSTRACT

It is rare for individual genes to exert influence on biological processes in isolation. Instead, they are controlled by intricate networks of genes that collaborate in a well-organized manner. Complex biological processes and their dysregulation in disease states are governed by the collaborative action of gene ensembles via epigenetic, genetic, and proteomic mechanisms. Through the analysis of their synergistic actions, researchers have the potential to understand the complex interplay of biological systems that are involved in the development, progression, and treatment of diseases. However, despite the abundance of readily accessible high-throughput technology, unraveling disease-related molecular pathways remains difficult. Possible factors contributing to this issue are background noise, batch issues, environmental conditions, individual heterogeneity, and technology limitations. To overcome these limitations, it is necessary to create more advanced, integrated, and personalized diagnostic methods that may provide a thorough understanding of disease biology, leading to enhanced diagnosis and treatment options.

In the field of disease diagnosis and treatment, genes often work together in the form of pathways that provide valuable insights into the fundamental processes of many disorders. In our study, we have examined the challenge of determining the direct relationships between pathways and diseases. We present sci-PDC, a method that leverages single-cell expression data to infer relationships between disease, cell type, and pathways. The use of this approach offers valuable perspectives on the causal connections between these variables and has the potential to make improvement in current precision medicine methods.

Another similar set of gene ensembles known as cancer hallmarks additionally serves a vital role in cancer identification by providing insights into the underlying features of cancer cells. In order to acquire a deeper understanding of the fundamental processes, we analyzed the hallmark properties of cancer in relation to canonical pathways. As we go, our objective is to investigate the drug's mechanism of action in connection to its specificity towards different cell types. Therefore, in this study, we used our technique to investigate the connections between the drug-targeted pathways and the distinctive characteristics of different single-cell cancer transcriptomes.

In genomic conformations, gene ensembles often collaborate via spatial organization, therefore exerting an influence on cellular activities and phenotype. These higher order chromatin topologies facilitate the integration of genes, and their regulatory elements, thereby enabling synchronised gene expression and regulation. This research used a unique methodology that included analyzing Topologically associating domains (TAD) activity in order to investigate the diversity of cancer and patients' responsiveness to drugs. Our study's results unequivocally show that TAD activity may function as a biomarker for estimating survival in the midst of tumor heterogeneity and predicting drug responsiveness.

Regulation of transcriptome and genomic conformations are often profoundly affected by epigenetic markers through mechanisms such as DNA methylation. The functional integrity of gene ensembles may be compromised by dysregulation of such epigenetic mechanisms, which can also lead to a number of diseases. Our work offers an elucidation of the computational difficulties associated with DNA methylation analysis, which arise from the inherent bias present in various approaches of profiling. Moreover, this study assesses the efficacy of deconvolution and machine learning methodologies in the examination of cell-free DNA (cfDNA) methylation, hence indicating their potential use in the early identification of cancer.

Overall, our suggested methodologies have the potential to leverage the synergistic effects of gene ensembles via diverse genomic and epigenomic patterns in order to provide a holistic comprehension of disease biology, hence enhancing diagnostic and therapeutic approaches.

LIST OF PUBLICATIONS

Publications and Preprints

1. **Sharma, Madhu,** Rohit Kumar Verma, Sunil Kumar, and Vibhor Kumar. "Computational challenges in detection of cancer using cell-free DNA methylation." ***Computational and Structural Biotechnology Journal*** 20 (2022): 26-39**.**
2. **Sharma, Madhu,** Indra Prakash Jha, Smriti Chawla, Neetesh Pandey, Omkar Chandra, Shreya Mishra, and Vibhor Kumar. "Associating pathways with diseases using single-cell expression profiles and making inferences about potential drugs." ***Briefings in Bioinformatics*** 23, no. 4 (2022): bbac241.
3. Neetesh Pandey[+], **Madhu Sharma**[+], Arpit Mathur, George Anene Nzelu, Muhammad Hakimullah, Indra Prakash Jha, Omkar Chandra, Shreya Mishra, Ankur Sharma, Roger Foo, Amit Mandoli, Ramanuj DasGupta, Vibhor Kumar. "Deciphering the phenotypic heterogeneity and drug response in cancer cells using genome-wide activity and interaction of chromatin domains." ***bioRxiv*** (2023). ([+]**Equal Contribution/co-first**)

Other publications and Preprints

4. Chandra, Omkar, **Madhu Sharma**, Neetesh Pandey, Indra Prakash Jha, Shreya Mishra, Say Li Kong, and Vibhor Kumar. "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.
5. Shreya Mishra, Neetesh Pandey, Smriti Chawla, **Madhu Sharma**, Omkar Chandra, Indra Prakash Jha, Debarka SenGupta, Kedar Nath Natarajan, Vibhor Kumar. "Matching queried single-cell open-chromatin profiles to large pools of single-cell transcriptomes and epigenomes for reference supported analysis." ***Genome Research*** (2023).
6. Chandra, Omkar, Durjay Pramanik, Srishti Gautam, **Madhu Sharma**, Niharika Dubey, Biswarup Mahato, and Vibhor Kumar. "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): 2024-05.

INDEX

LIST OF FIGURES

3.5. Drug response prediction using cell line specific hallmark pathway associations.

4. Chapter 4
   4.1. Investigation of chromatin interaction patterns and domain activity in cell lines originating from patients with head and neck cancer.
   4.2. Drug response and TAD-activity association patterns.
   4.3. Analysis of topologically associated domains (TADs) in various cancer samples from TCGA and the relationship with patient survival.
   4.4. Correspondance between the connection of TAD activity with survival across TCGA patients and the association between TAD and response towards drugs in CCLE cell lines is examined.
   4.5. Providing a comprehensive case study on the topologically associated domain (TAD) that encompasses the EGFR gene.

5. Chapter 5
   5.1. A brief description of the approach followed.
   5.2. This heatmap illustrates the variability that exists across biomarkers across a certain particular type of cancer.
   5.3. This boxplot illustrates the performance of a single biomarker for sample label prediction (breast cancer or non-cancerous) by comparing the false positive rate (FPR) to the sensitivity measures.
   5.4. Implementing various deconvolution methods to analyze the DNA methylation patterns of cancerous and non-cancerous samples.

CHAPTER 1

INTRODUCTION

1.1 Background

1.1.1 Overview of gene ensembles

Gene ensembles, often referred to as gene clusters or gene sets, constitute collections of genes that exhibit shared biological roles, regulatory patterns or chromosomal positions. They are essential in many biological study domains, such as transcriptomics, genomics, and bioinformatics.

Variety of Criteria: Genes may be categorized into different groups depending on a range of criteria, each providing distinct perspectives.

- *Functional similarities*: Functional similarities arise when genes participate in the same biological pathways and cellular activity or possess analogous molecular activities, resulting in their grouping together (Garcia-Moreno et al. 2022). For instance, genes implicated in the glycolysis process.
- *Physical proximity*: Physical proximity refers to the near positioning of genes on a chromosome. When genes are placed close together, they have a higher likelihood of being co-regulated and are often organized into sets. This may highlight possible patterns of co-expression (García-Cortés et al. 2020).
- *Literature curated ensembles*: Literature-curated ensembles refer to gene sets that researchers create based on certain biological themes or disease markers that have been observed in the literature. These sets might exhibit a high degree of specificity in relation to a particular field of inquiry (Wang et al. 2013).

- *Regulation based gene ensembles*: Regulatory based gene ensembles are defined as sets of genes that are co-regulated by certain transcription factors or other regulatory components (Figure 1.1) (Groenewoud et al. 2022; Chandra et al. 2024).

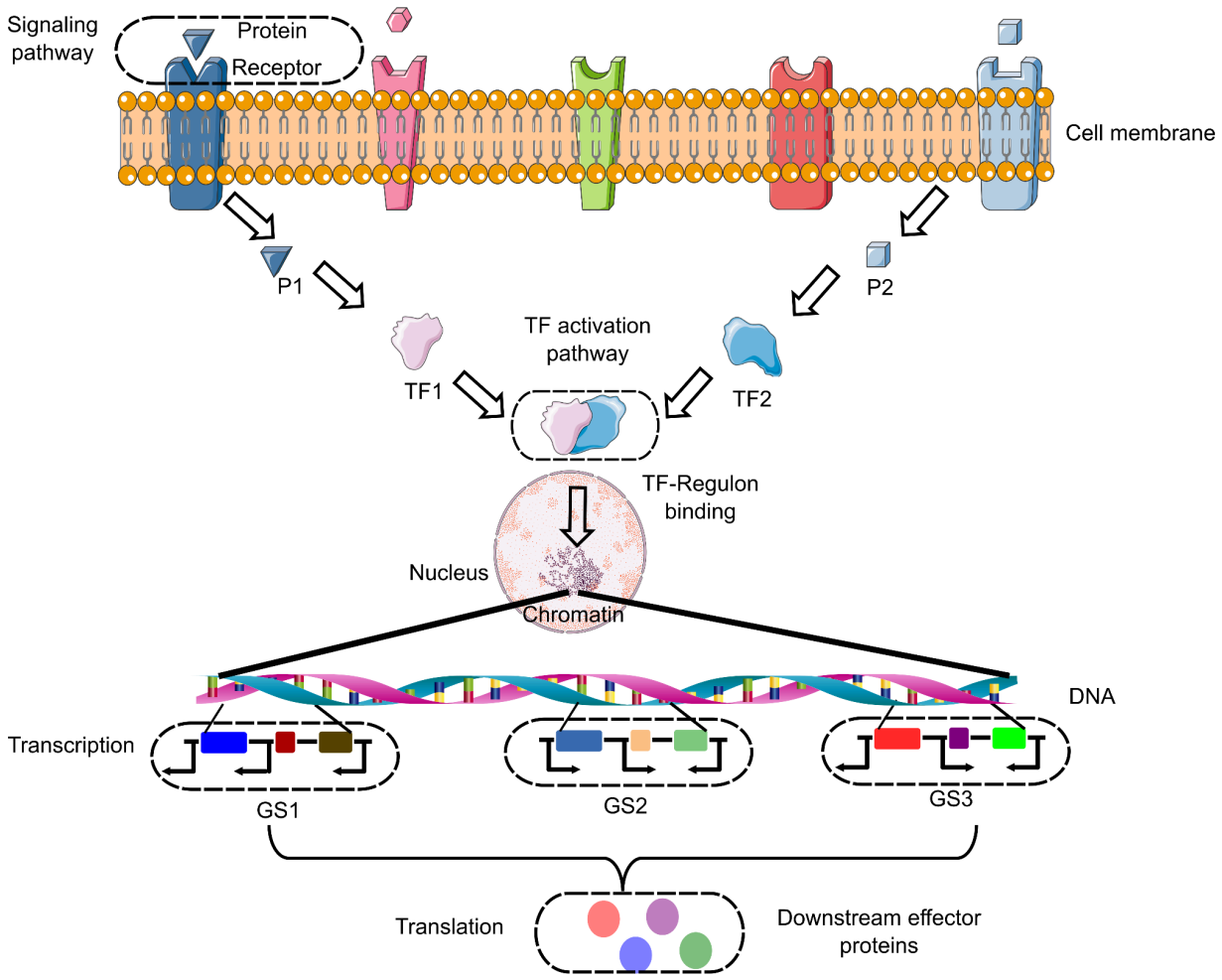

Figure 1.1: A schematic illustration depicting synergistic behavior of genes in regulation. Proteins P1 and P2, upon activation by their corresponding cell membrane receptors, subsequently activate transcription factors TF1 and TF2. By binding to their target regulome in the nucleus, these transcription factors induce the expression of gene sets GS1, GS2, and GS3. The translation of these gene sets yields effector proteins that together influence the observable phenotype in the cell. The coordinated expression of genes ensures that the appropriate combination of genes is activated, thereby enabling a precise cellular response and a desirable phenotype.

Traditional Methods for Gene Ensemble Selection:

A number of conventional methods have been used for selecting gene ensembles in view of their potential implications in particular diseases, therapies, or biological processes. Such as,

- *Differential gene expression analysis (DGE)*: DGE analysis is a popular gene ensemble selection method. This approach finds genes with significantly different expression levels between two or more experimental conditions or treatments, such as disease vs control. It is believed that the molecular mechanisms behind the condition under study are influenced by DEGs (Joly et al. 2021; Yaari et al. 2013).
- *Co-expression network analysis*: Co-expression network analysis is another popular method for identifying gene ensembles. One prominent approach for generating these networks is Weighted Gene Co-expression Network Analysis (WGCNA) (Langfelder & Horvath 2008). WGCNA clusters genes into modules with highly correlated expression profiles, indicating co-regulation or functional relatedness. Associating these modules with external traits helps identify gene ensembles involved in biological processes (Yin et al. 2019).
- *Genetic variation analysis*: Genome-wide association studies (GWAS) identifies SNPs and other genetic variations linked to certain characteristics or diseases (Sun et al. 2019). This method can identify genetically altered gene ensembles by correlating genetic variation to gene expression patterns (e.g., eQTL analysis) (Zhu et al. 2019). Such variation-associated gene ensembles may affect disease susceptibility or treatment response.
- *Machine learning approaches*: A variety of machine learning techniques have also been implemented to select gene ensembles in recent years. These approaches can find gene ensembles that predict disease diagnosis, prognosis, and therapy response in large, complex data sets. Training models on gene expression data allows these algorithms to rank genes by significance to the predictive model, emphasizing those that best discriminate conditions or predict clinical outcomes (Abbas & El-Manzalawy 2020; Silva et al. 2022).

A brief description of gene ensembles' resources and databases:

- *KEGG (Kyoto Encyclopedia of Genes and Genomes)*: A knowledge repository for comprehensive gene function study, KEGG links genomic and higher-level functional

knowledge (Kanehisa & Goto 2000). KEGG offers Java graphical tools for the exploration and comparison of genomic maps and amending expression charts, as well as computational tools for sequence and graph comparisons. KEGG databases are publicly accessible at http://www.genome.ad.jp/kegg/.

- *Reactome*: Reactome is a pathway database that is reviewed by experts, and curated manually (https://reactome.org/). The database offers user-friendly tools for bioinformatics that facilitate the visualization, comprehension, and analysis of pathway information. These tools are designed to serve several domains, such as fundamental biology, clinical studies, genomic analysis, and systems biology (Croft et al. 2011).
- *Gene Ontology (GO)*: The primary objective of the GO Consortium is to construct a contemporary and exhaustive computational representation of biological systems, spanning from the molecular scale to more comprehensive pathways, cellular structures, and organism-level frameworks (Ashburner et al. 2000). The Gene Ontology information repository offers a computer depiction of our existing scientific understanding of the roles of genes, including the protein and non-coding RNA molecules, across various organisms ranging from humans to bacteria (https://geneontology.org/).
- *MSigDB (Molecular Signatures Database)*: MSigDB is a highly used collection of well annotated gene collections that include a variety of biological processes (Liberzon et al. 2011). These gene ensembles are essential for the meaningful and informative analysis of large-scale genomic data. MSigDB provides an extensive selection of more than 6700 gene ensmbles, together with a comprehensive update of canonical pathways and the experimental markers sourced from literature, as well as streamlined annotations. MSigDB can be explored at https://www.gsea-msigdb.org/gsea/msigdb/.
- *BioCarta*: BioCarta is a repository of gene interaction maps. The database comprises high-resolution graphics depicting various biological signaling and interaction pathways (Adriaens et al. 2008). Each diagram is extensively connected to product and details of sites dedicated to certain genes. The user-friendly interface of this resource provides crucial information for over 120,000 genes across several species. The database consists of 296 pathways that are updated regularly, but with occasional interruptions in availability (http://www.biocarta.com/).

- *WikiPathways*: WikiPathways is a collaborative scientific platform that enables researchers from across the world to contribute, modify, and use knowledge on biological pathways (Agrawal et al. 2024). WikiPathways is an innovative approach to pathway databases which enhances and complements existing initiatives, such as KEGG, Pathway Commons and Reactome. Furthermore, it proposes a new paradigm for pathway databases. Searching, browsing, interacting with authors and communities, editing, and exporting pathway information in various formats are just a few of the many tools available on the website for interacting with the database (https://www.wikipathways.org/).

There are several bioinformatics resources available for free and non-commercial usage, such as MSigDB, Reactome, KEGG etc. This enables researchers to utilize a diverse array of datasets and tools without requiring a commercial license.

1.1.2 Rational behind gene ensembles analysis

Concentrating on specific genes across huge data sets may be daunting and fails to reflect the broader context. Gene ensembles enable researchers to discern clusters of genes that are collectively engaged in a certain biological activity. Researchers may streamline their study through the use of curated gene ensembles that are derived from established biological knowledge, allowing them to concentrate on pertinent pathways or functions. This process aids in eliminating possibly extraneous genes and pinpointing crucial contributors in the system being examined.

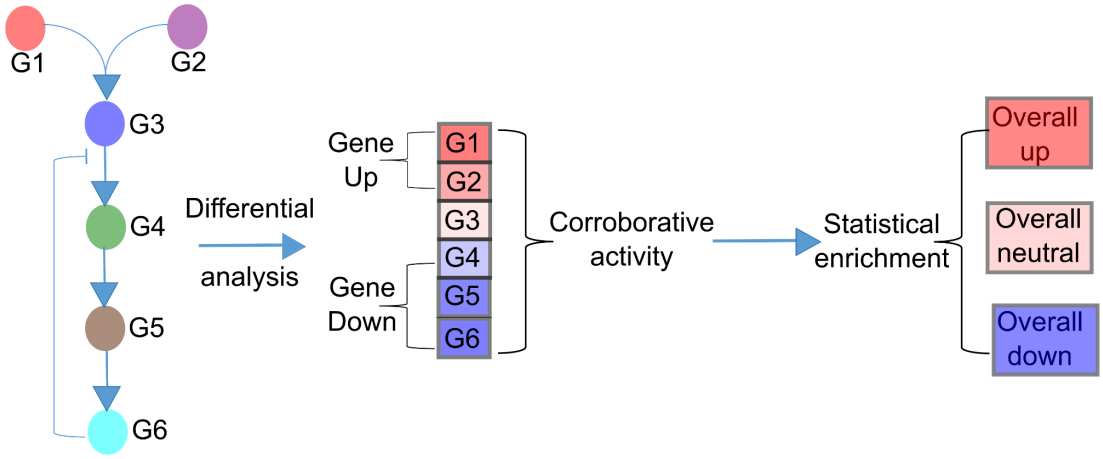

Figure 1.2: Comparing autonomous vs synergistic effects. This illustration depicts the cascade of gene expression, wherein genes G1 and G2 stimulate G3, resulting in the further activation of genes G4, G5, and G6. The proactive participation of G6 initiates a feedback inhibition loop, reflecting the dynamic control of gene expression. The autonomous approach analyzes gene expression for each gene individually, independently evaluating its upregulation or downregulation. Conversely, the synergistic mechanism examines the collective activity of all participating genes, evaluating their interactions and assessing overall differential expression by statistical enrichment.

1.1.2.1 Autonomous vs synergistic gene effects

To comprehend the way genes interact and affect biochemical reactions and phenotypes, it is essential to grasp the ideas of autonomous and synergistic effects of genes (Figure 1.2). Autonomous effects pertain to the influence of an individual gene on a specific feature or phenotype, without being influenced by other genes. These effects manifest when the gene operates alone, without any interaction or modification from other genes. Conditions such as cystic fibrosis or sickle cell anaemia are the result of mutations in a single gene, namely CFTR and HBB, respectively (Bobadilla et al. 2002; Carlice-Dos-Reis et al. 2017). These mutations have independent consequences. In the field of developmental biology, cell-autonomous genes exert their influence inside the specific cell where they are active, impacting the fate and behaviour of the cell without being influenced by neighbouring cells (Gilbert 2000). Synergistic effects, on the other hand, pertain to the collective influence of many genes that interact to generate a phenotype that deviates from what would be anticipated based on the cumulative impact of their individual impacts. During these interactions, the genes collaborate, and their combined effect is more significant than the aggregate of their separate contributions. Epistasis is a situation in which the impact of one gene is altered by one or more other genes (Table 1.1). For instance, in the determination of coat colour in mice, the impact of the agouti gene may be concealed by the existence of a recessive allele at a different genetic location (Morgan et al. 1999). Moreover, several genes operate together in biological pathways to control responses of cells to external stimuli, signal transmission, metabolism, and other functions (Carthew 2021; van Dam et al. 2018; Armingol et al. 2021). Disrupting a single gene may have a profound impact on the whole network, resulting in a synergistic effect.

Aspect	Autonomous Effects	Synergistic Effects
Definition	Isolated effect of a single gene	Cummulative effect multiple collaborating genes
Gene functionality	Operates independently, direct phenotype association	Collaborative, interdependent partnership
Phenotypic response	Concise and unambiguous	Elaborate and enhanced or modulated
Examples	Monogenic diseases, cell-autonomous genes	Epistasis, gene networks, polygenic traits
Study approach	Can be examined independently	Demands examination of biological connections and networks

Table 1.1: Assessing the relevant properties of autonomous and synergistic processes

1.2.2.2 Leveraging the synergistic effects of gene ensembles for disease therapy

The use of gene ensembles in disease therapy has many benefits, capitalising on the intricacy of genetic relationships to provide therapies that are both more efficient and tailored to individual patients. Below is an elaborate summary of these benefits:

- *Comprehensive Understanding of Disease Mechanisms*: Gene ensembles provide a thorough comprehension of disease mechanisms by revealing the biological pathways and networks implicated in disease processes. Comprehending the interactions between several genes in these pathways helps in identifying crucial regulatory sites that might be specifically targeted for therapeutic purposes (Cordell 2009; Motsinger & Ritchie 2006). Furthermore, gene ensemble analysis (GEA) takes into account the wider genetic framework rather than only concentrating on individual gene changes. This method acknowledges the complex character of a variety of diseases, so providing a comprehensive and holistic perspective.

- *Personalized Medicine*: Genetic profiles linked to distinct therapeutic responses may be identified by GEA, which facilitates personalised medicine. This enables the customisation of treatment strategies according to each individual's distinct genetic composition (Wang et al. 2015). Gene ensembles may assist in identifying biomarkers for disease diagnosis, prognosis, and therapy response, hence enabling more accurate and individualised healthcare (Angione 2019).
- *Enhanced Treatment Effectiveness*: By targeting many genes simultaneously, it is possible to take advantage of synergistic effects, in which the combined impact of numerous drugs is more potent than the sum of their individual effects. Gene ensemble based drugs have the ability to target numerous pathways, which may effectively decrease the chances of resistance formation. This is in contrast to single-target therapies, which often face the problem of resistance (Ng et al. 2014; Goswami et al. 2015). However, gene interactions within such genes are complicated and can alter non-target genes or cellular activities inadvertently. Off-target effects may cause phenotypic alterations, immunological responses, or cellular process dysregulation. To reduce off-target consequences, gene ensemble specificity must be carefully assessed as gene editing methods develop (Atkins et al. 2021).
- *Implications in the complex and polygenic disorders*: Several diseases, including diabetes, cancer, and many cardiovascular disorders, result from intricate interplay among several genes. GEA is very suitable for comprehending and specifically addressing these complex diseases influenced by several factors. Gene ensembles facilitate the investigation of polygenic features, which include the contribution of numerous genes to the susceptibility and course of diseases. This allows for the development of more comprehensive treatment approaches (Gilbert-Diamond & Moore 2011; Glasspool et al. 2006).
- *Integration with Emerging Technologies*: GEA may be used to analyse single-cell RNA sequencing data (scRNA-seq), offering valuable knowledge on the diversity of cells and suggesting therapeutic targets particular to each cell type (Wilson et al. 2015). Gaining knowledge about gene ensembles and their connections might be advantageous in formulating CRISPR-based gene editing techniques, enabling accurate change of several genes concurrently for therapeutic intentions (Shmakov et al. 2018).

Utilising gene ensembles for disease management is a potent method for comprehending and addressing complex disorders. Through the examination of gene relationships and collective impacts, scientists and physicians may create therapies that are more efficient, tailored to individual needs, and have reduced risks. This approach not only improves our comprehension of disease processes but also creates new opportunities for drug development, personalised therapy, and the treatment of complex disorders.

1.1.3 Statistical approaches for gene ensemble analysis

GEA address the problem of analyzing extensive gene expression datasets by using specific gene ensembles. These sets categorize genes according to their common biological function, participation in a pathway, or other relevant characteristics. Statistical methods for GEA are designed to discover gene collections that show substantial enrichment or depletion in a certain condition when compared to a baseline. An outline of the primary statistical approach categories in GEA is provided below:

1.1.3.1 Over-Representation Analysis (ORA)

ORA is a statistical approach used to determine whether a certain collection of genes is enriched in an ordered list of genes, relative to the level that would be predicted by chance. This category uses statistical techniques such as the hypergeometric distribution, chi-square or Fisher's exact test to evaluate the likelihood of detecting the amount of enrichment seen in the data (Maleki et al. 2020).

Example:

- Tools for pathway analysis: Tools such as Ingenuity Pathway Analysis (IPA) and PANTHER use ORA techniques to evaluate pathway enrichment using expression of genes (Anon n.d.; Mi & Thomas 2009).

1.1.3.2 Functional Class Scoring (FCS)

The objective of FCS approaches is to give a score to every gene collection that accurately represents its overall relationship with a phenotypic or experimental condition. This class uses diverse statistical techniques to compute the score. Some often used methods are Kolmogorov-Smirnov (KS) statistics, rank aggregation techniques, and so on. The KS statistic

quantifies the disparity between the cumulative distribution functions (CDFs) of the gene ensemble and the background gene collection. Rank aggregation techniques merge data from several ranked lists, such as those obtained from separate studies, to get a combined score for each gene collection (Maleki et al. 2020).

Example:

- Gene Set Enrichment Analysis (GSEA): GSEA is a widely used approach for ORA that takes into account a comprehensive ranked set of genes, rather than relying on a predetermined threshold for differential expression (Subramanian et al. 2005). The algorithm computes an enrichment score that quantifies the extent to which the genes in the given collection are concentrated at either the highest or lowest of the sorted list.

1.1.3.3 Pathway Topology (PT) Methods

PT techniques integrate knowledge of the topological architecture of biological processes, such as interactions among proteins, into the study. These strategies often use graph-based methods and network modelling tools to evaluate enrichment inside the framework of pathway topology (Maleki et al. 2020).

Example:

- GSNCA (Gene Sets Net Correlations Analysis): GSNCA is a statistical test that examines the differential co-expression of many genes, taking into consideration the whole correlation architecture between them. GSNCA detects disparities in co-expression networks and identifies genes associated with prominent and specialised pathway regulators, as well as genes that are most influenced by the biological variation between two situations (Rahmatallah et al. 2014).

The choice of the most appropriate statistical method for GEA relies on several aspects, such as the nature of gene collections, the purpose of the study, and the specifics of the data (Damian & Gorfine 2004). The selection of methodology may be impacted by variables for instance the array of genes, the count of the gene ensembles, and the existence of preexisting biological information. For example, when doing curated pathway analysis, ORA approaches may be appropriate. When dealing with gene ensembles obtained via co-expression analysis, it may be more suitable to use FCS approaches. However, there are other considerations that need to be taken into account, such as the greater possibility of false positives when analysing several gene

ensembles. Researchers must use statistical adjustments to account for multiple testing (Das et al. 2020). Through a thorough comprehension of the many statistical methodologies that are accessible and a meticulous evaluation of the research question and data attributes, scientists may effectively use GEA techniques to derive significant biological knowledge from various gene expression data sets.

1.1.4 Current paradigms in disease biology, prognosis, and therapy

GEA is a potent method used in the study of disease biology, prognosis, and treatment. It offers valuable information on the combined actions of genes and aids in comprehending the fundamental processes of diseases. A thorough examination of GEA's applicability in several fields is provided below:

1.1.4.1 Utilising gene ensembles to comprehend disease biology pathways and mechanisms

- *Pathway analysis*: GEA aids in the identification of dysregulated biological pathways in diseases. For instance, in the context of cancer, GEA may specifically identify and emphasise the pathways that are entailed in the control of cell cycle, programmed cell death (apoptosis), and the dissemination of cancerous cells throughout the body (metastasis) (Liu et al. 2023).
- *Functional insights*: Researchers may get valuable information on the functional roles of certain genes in the development of diseases by analysing gene ensembles. This is especially beneficial for complex disorders in which numerous genes contribute to the pathogenesis. Such as in Alzheimer's condition, GEA may assist in revealing networks associated with amyloid-beta formation, tau phosphorylation, and neurological inflammation (Andrews et al. 2023).
- *Biomarker analysis*: GEA has the capability to find groups of genes that may be used as biomarkers for certain diseases. These biomarkers may assist in the early detection and comprehension of disease development. For instance, GEA has been used to investigate crucial signalling pathways implicated in many cancer types, such as the PI3K/AKT pathway involvement in breast cancer or the Wnt signalling pathway in the colorectal cancer (Imkeller et al. 2022; Madsen et al. 2021).

1.1.4.2 Prognostic models for disease utilising ensembles of genes

- *Prognostic biomarkers*: GEA may uncover groups of genes that are associated with disease prognosis. These groups of genes may function as prognostic biological markers, aiding in an assessment of disease progression and rate of survival. For instance, GEA has the capability to detect gene ensembles that are linked to the likelihood of experiencing myocardial infarction or stroke, hence assisting in the forecast of unfavourable cardiovascular incidents (Wang & Xian 2022).
- *Risk stratification*: Risk stratification may be achieved by examining gene expression patterns and their corresponding gene ensembles, allowing for the classification of patients into various risk groups. This allows for individualised surveillance and control of disease. For instance, the GEA technique has been used to create gene expression profiles that may accurately forecast the probability of recurrence and the effectiveness of treatment, as shown by the Oncotype DX and MammaPrint tests (Yao et al. 2022).

1.1.4.3 Gene ensembles for identifying targets and development of drugs in therapeutics

- *Therapeutic targets*: GEA aids in the identification of crucial pathways and genes that may be specifically targeted for therapeutic action. Researchers may design drugs that target particular pathways by identifying dysregulated gene ensembles. For instance GEA can detect dysregulated immunological pathways in diseases like lupus or rheumatoid arthritis, providing guidance for the design of precise immunotherapies (Shchetynsky et al. 2017; Wang et al. 2022).
- *Tailored therapies*: GEA facilitates the creation of personalised drugs that are specifically designed according to an individual's gene expression profile. This strategy makes sure that patients are provided with therapies that are very probable to be efficacious for their particular genetic composition. For example, GEA has the capability to identify individuals who have a high probability of responding positively to certain targeted medicines, such as tyrosine kinase inhibitors for lung cancer or PARP inhibitors for ovarian cancer (Gurule et al. 2021; Hodgson et al. 2018).

1.1.5 Challenges and Considerations of the field

GEA is an effective technique with wide applications in disease biology, prognosis, and therapy. However, there are multiple obstacles and variables that researchers have to overcome when implementing GEA. These obstacles span from technological and computational concerns to biological and interpretational complications. Here's a brief overview:

1.1.5.1 Barriers during analysis

- *Quality of gene ensembles*: The analysis's effectiveness greatly depends on the quality of sets of genes used. Researchers must use discernment when evaluating the primary sources and criterion utilised in the construction of the gene ensembles they integrate into their research. Gene ensembles that include biases or errors might result in deceptive outcomes (D. W. Huang et al. 2009a).
- *Data interpretation*: Data interpretation is a crucial aspect when analysing GEA data. Although enriched collections of genes provide insights into potentially important pathways, they may not always demonstrate the direct participation of each gene within the set (Song & Black 2008). Additional validation tests involving particular genes or pathways may be required.
- *Multiple testing corrections*: Multiple testing corrections are necessary when analysing many gene ensembles to reduce the elevated possibility of discovering false positives arising from randomness. Researchers must use statistical adjustments for repeated testing to address amplification of significance (Blüthgen et al. 2005).
- *Data standardisation and batch effects*: Gene expression data may be susceptible to technical variances or batch effects. Effective data standardisation techniques are essential to prevent inaccurate findings in GEA (D. W. Huang et al. 2009a).

1.1.5.2 Contemplations

- *Choice of method*: When selecting a technique for GEA, there are several statistical methodologies available, such as Pathway Topology (PT) methods. Functional Class Scoring (FCS), and Over-Representation Analysis (ORA). The selection of methodology is contingent upon the research inquiry, the nature of the gene ensembles employed, and the features of the data (Benjamini & Hochberg 2018).

- *Integrated analyses*: GEA is often used in tandem with differential expression study to provide a more thorough comprehension of the data. Researchers must take into account how GEA findings align with the wider scope of the investigation (Araki et al. 2012).
- *Biological knowledge*: Utilising pre-existing biological information about pathways and gene functions may improve the understanding of GEA findings and provide direction for more research (Goeman & Bühlmann 2007).
- *Computational resources:* GEA may need significant computing power, particularly when working with extensive datasets or intricate network-based techniques. When structuring their study, researchers must take into account the computing resources that are currently accessible (Gui et al. 2011).

Through recognition of these obstacles and factors, scientists may efficiently use GEA to get significant biological understanding from gene expression information. Employing a critical perspective, together with meticulous research design and data analysis, will facilitate the realisation of the whole capabilities of this effective tool in enhancing our comprehension of intricate biological processes and disorders.

1.2 Coactivity of gene ensembles across transcriptome profiles

1.2.1 Significance of transcriptomic analysis

The central dogma of molecular biology emphasises the significance of mRNA as the intermediate agent connecting DNA and protein (Crick 1970). Transcriptomic profiles include all the RNA transcripts that exist in a cell, tissue, or organism at a certain moment. These profiles provide a comprehensive perspective on gene expression, including the detection and measurement of mRNA, as well as other RNA types such as rRNA, tRNA, and non-coding RNAs. Examining these profiles is crucial for comprehending the ever-changing characteristics of gene expression and control (Samuels et al. 2021).

The importance of transcriptome analysis is amplified when it is considered in the context of disorders. Various diseases, such as cancer, neurological disorders, and infectious diseases, are characterised by changes in gene expression. Through the comparison of transcriptome patterns between sick and healthy tissues, researchers may identify pathways and genes that exhibit dysregulation in disease states. This information may elucidate the molecular foundations of the

disease, establish possible biomarkers for prognostic and diagnostic purposes, and unveil new treatment targets (Ospina et al. 2024). Furthermore, transcriptome analysis plays a crucial role in personalised therapy. Healthcare practitioners may customise therapies for patients by analysing the gene expression patterns of each person, taking into account the unique molecular features of the patient's condition (Chiang et al. 2023).

1.2.2 High-throughput transcriptomic profiling

High-throughput methods have greatly transformed the field of transcriptomics by enabling thorough and intricate examination of RNA transcripts in diverse biological samples (D'Agostino et al. 2022). These technologies allow researchers to measure the levels of gene expression, identify new transcripts, and comprehend the intricacies of gene control (Srivastava et al. 2018). The following are the primary high-throughput techniques used for generating transcriptome profiles:

1.2.2.1 Microarrays

Microarrays are an early high-throughput technology used for studying gene expression. These arrays are composed of a matrix of DNA probes that are immobilised on a stable substrate, with each probe specifically targeting a particular gene. The RNA is isolated from the sample, transformed into complementary DNA (cDNA), and then tagged with fluorescent markers. The labelled cDNA is attached to the microarray, where it forms a connection with matching probes. The fluorescence intensity at each probe location is quantified, reflecting the degree of expression of genes. Microarrays are a sensible choice for analysing established genes and are well-suited for monitoring the expression of genes across a large number of samples. Nevertheless, their agile range and sensitivities are restricted, and they are incapable of detecting newly constituted transcripts and alternative splicing instances (Govindarajan et al. 2012).

1.2.2.2 RNA Sequencing (RNA-Seq)

A more thorough and precise assessment of transcriptomes is possible using next-generation sequencing (NGS) technologies like RNA-Seq. The process involves the extraction of RNA, followed by the conversion of the extracted RNA into a library of cDNA fragments. These fragments are then subjected to sequencing utilising high-throughput sequencing technologies.

The obtained sequencing reads are matched against a reference genome or a transcriptome, and the level of abundance of every single transcript is measured. RNA-Seq has the ability to identify both previously known and unknown transcripts, alternative splicing occurrences, and has a broad range of variability and exceptional sensitivity. Nevertheless, it incurs more costs and demands greater computational power and bioinformatics proficiency in contrast to microarrays (Koch et al. 2018).

1.2.2.3 Digital Gene Expression (DGE)

DGE is a sequencing technology that uses tags to offer quantitative information on the expression of genes. Sequences of about 17-20 base pairs are produced and analysed from the 3' end of every transcript. Each tag corresponds to a distinct gene. The frequency of each tag's occurrence is tallied, yielding a digital quantification of gene expression levels. DGE is a highly quantitative method that uses a digital representation to accurately estimate the quantity of transcripts. The drawback is that it might not capture complete transcripts or convey details regarding alternate splicing (Liu et al. 2018).

1.2.2.4 single-cell RNA Sequencing (scRNA-seq)

single-cell RNA sequencing is an advanced method that analyzes gene expression at the individual cell level, using RNA sequencing (RNA-Seq). The process involves the isolation of single cells, subsequently accompanied by extraction of their RNA and its conversion into cDNA libraries. These libraries are then subjected to sequencing. scRNA-seq enables the introspection of the transcription patterns in individual cells, facilitating the investigation of variations between cells and the determination of unique cell types. It offers valuable information about the diversity inside a tissue, uncovering uncommon cell types and transitory conditions in cells (Mishra, Pandey, Chawla, et al. 2023). However, the technique is costlier, has reduced performance per cell in comparison with bulk RNA-Seq, and requires intricate data interpretation (Jovic et al. 2022).

1.2.3 Leveraging transcriptome profiling to associate disease gene ensembles with pathways

Examining the function of gene ensembles in transcriptome profiles is a potent method for comprehending the fundamental molecular systems and biological events associated with different disease states (Mathur et al. 2018). This methodology facilitates the identification of crucial regulating pathways, the detection of possible biomarkers, and the formulation of targeted medicines. Moreover, the integration of pathway and disease GEA offers a more thorough comprehension of the molecular underpinnings of disorders (Lamb et al. 2006). By combining the characteristics of both approaches, this integrative method is able to unearth important insights.

1.2.3.1 Pathway-Disease Interaction Networks

By combining the gene ensembles of pathways and diseases, researchers are able to develop interaction graphs that provide insight into the connections that exist between pathways and disorders (Buphamalai et al. 2021). This allows for the disclosure of:

- *Communal pathways*: In many diseases, equivalent pathways are often dysregulated, which suggests that the underlying mechanisms are similar.
- *Disease specific pathways*: Pathways that possess characteristics distinctive for particular diseases and serve as targets for therapy.

1.2.3.2 Multi-Omics Integration

Integrating transcriptome data from the transcriptome with metabolomics and proteomics information, among other omics data, improves the comprehension of pathway activity and progression of diseases (Jeon et al. 2023).

- *Methods based on networks*: Create complex networks by integrating various forms of omics data to find crucial regulatory nodes and connections.
- *Machine learning techniques*: Utilise machine learning models to include and examine multi-omics data, accurately forecasting states of disease and pathway activity with enhanced precision.

1.3 Cancer hallmark properties as gene ensembles in cancer transcriptomic profiles

1.3.1 Relevance of Cancer Hallmark Properties

Cancer is an intricate and diverse disease that occurs due to the buildup of genetic and epigenetic changes that disturb the normal functioning of cells. In order to comprehend and efficiently counteract the disease, scientists have recognised a collection of essential skills referred to as the Hallmarks of Cancer, which are obtained by cells as they progress towards being malignant(Fares et al. 2020). The hallmarks, first suggested by Douglas Hanahan and Robert Weinberg in 2000, are fundamental basic principles that differentiate cancer cells from normal cells (Hanahan & Weinberg 2000). Gaining a comprehension of these distinguishing characteristics offers a structure for deciphering the complex processes involved in the advancement of tumours and directing treatment approaches.

While cancer hallmarks highlight the essential properties that facilitate the growth, survival, and metastasis of cancer cells, it is crucial to acknowledge that these characteristics emerge from the integrated functioning of several biological systems. These activities are governed not by individual genes but by gene ensembles *i.e.,* collections of genes that collaboratively control cellular functions.

Gene ensembles are essential in facilitating the phenotypic manifestations of cancer hallmarks, including evasion of growth suppressors as well as sustaining proliferative signaling. For example, a gene ensemble engaged in cell cycle regulation could promote uncontrolled cell division, which is a characteristic of persistent proliferative signaling. Consequently, cancer hallmarks represent observable traits of cancer cells, while gene ensembles are the fundamental genetic elements that facilitate the expression of these hallmarks.

It is essential to elucidate that cancer hallmarks are not synonymous with gene ensembles. Hallmarks denote the functional results of intricate biological processes, whereas gene ensembles pertain to the particular genetic elements that drive these processes. Grasping this dichotomy is essential for clarifying the molecular pathways that govern cancer biology.

1.3.2 Major Cancer hallmark properties

- *Sustaining Proliferative Signaling*: Healthy cells carefully control their division and growth in reaction to environmental signals in order to maintain proliferative signalling. Cancer cells, on the other hand, circumvent these restrictions by activating signalling pathways that consistently stimulate their growth (Ciccarone & Ciriolo 2023).
- *Evading Growth Suppressors*: Healthy cells contain intrinsic mechanisms that inhibit unregulated proliferation. Cancer cells often disable these kinds of genes as well as proteins that inhibit proliferation, enabling them to multiply without restraint (Scheel & Schäfer 2023).
- *Resisting Cell Death*: Apoptosis, also known as programmed cell death, is a vital mechanism for eliminating damaged or unnecessary cells. Cancer cells use many methods to elude apoptosis, allowing them to persist even under circumstances that typically trigger cellular death. (Zhu et al. 2022)
- *Enabling Replicative Immortality*: Normal cells have a finite lifetime and ultimately halt their division. On the other hand, cancer cells have the capacity to undergo unlimited division, enabling them to create tumours and possibly spread to other parts of the body (Bailey 2023).
- *Inducing Angiogenesis*: Tumours require an uninterrupted supply of nutrients and oxygen in order to proliferate. Cancer cells stimulate angiogenesis, the process of forming new blood vessels, in order to fulfil these requirements (Saman et al. 2020).
- *Activating Invasion and Metastasis*: Cancer cells may have the capability to infiltrate nearby tissues and navigate to distant locations in the body, resulting in the formation of additional tumours, a process known as metastasis. This particular process is the phase of cancer that is the most fatal (Fares et al. 2020).
- *Avoiding Immune Destruction*: Typically, the immune system identifies and eradicates uncommon cells. Cancer cells use strategies that circumvent immune detection, enabling unrestricted growth (Kim & Cho 2022).
- *Genome Instability and Mutation*: The gradual buildup of genetic alterations in crucial genes is a defining characteristic of cancer. This instability in genes drives the emergence of additional distinctive characteristics (Negrini et al. 2010).

- *Tumor-Promoting Inflammation*: Prolonged inflammation in the tumor microenvironment may aid in the development of tumors by stimulating their development, survival, and infiltration of cells (Roncati & Figueiredo 2023).
- *Reprogramming Energy Metabolism*: Cancer cells undergo a change in metabolism to prioritise fast growth and sustain uncontrolled proliferation, even when oxygen is scarce. This reprogramming involves a metabolic mechanism called aerobic glycolysis, which is often known as the Warburg effect (Kaur et al. 2023).

1.3.3 Relationship between cancer hallmark properties and pathways

Pathways connected to the hallmark properties of cancer are essential for promoting the growth and proliferation of cancer cells. Every hallmark associated with cancer is supported by distinct signaling processes and biological mechanisms. Gaining knowledge about these pathways offers valuable understanding of the formation of individual therapies as well as the detection of potential markers for the diagnosis and treatment of cancer. The relationship between key pathways and each of the hallmarks that define cancer is examined in detail below:

1.3.3.1 Key signaling pathways associated with different cancer hallmark properties

- *Pathways Involved in sustaining proliferative signaling*: These pathways encompass the Ras-ERK, PI3K-AKT-mTOR, and Wnt signaling pathways. Typically, these pathways control the development and division of cells in response to external signals. In the context of cancer, mutations or excessive activation of these pathways result in the perpetuation of growth signals, which in turn facilitate unregulated cell division (Asati et al. 2016).
- *Pathways associated with cell death resistances*: The apoptotic pathway, which is regulated by proteins such as caspases, is a primary focus for cancer cells (Olsson & Zhivotovsky 2011). Alterations or disruption of these pathways can hinder cancer cells from experiencing apoptosis.
- *Pathways involved in facilitating replicative immortality*: Telomerase, an enzyme responsible for preserving the length of telomeres, plays essential role in facilitating ability of cells to replicate indefinitely (Yaswen et al. 2015). Cancer cells frequently

utilize telomerase to bypass the inherent limitation of cell replication known as replicative senescence.

- *Relevant pathways in angiogenesis*: The VEGF pathway plays a vital role in regulating angiogenesis. Cancer cells have the ability to release VEGF, which encourages the formation of new blood vessels (Shibuya 2011). This process guarantees a continuous provision of nutrients and oxygen for the growth of tumors.
- *Pathways implicated in invasion and metastasis*: Epithelial-mesenchymal transition is a process in biology wherein epithelial cells undergo a transformation, acquiring characteristics of mesenchymal cells. This transformation enables them to move and infiltrate neighboring tissues. Signaling pathways such as TGF-β and Notch signaling participate in the process of epithelial-mesenchymal transition (EMT) (Kim et al. 2020). Furthermore, the processes responsible for cell adhesion, motility, and breakdown of the extracellular matrix are also disrupted in metastasis.
- *Pathways involved in cellular metabolism deregulation*: The Warburg effect, a characteristic feature of cancer, entails a transition towards aerobic glycolysis as the primary mechanism for energy generation, even in the presence of oxygen. The shift in metabolism is frequently caused by genetic mutations, such as those in the MYC gene, and changes in signaling pathways, such as AMPK (Faubert et al. 2013).

1.4 Gene ensembles in three-dimensional genomic organization

1.4.1 Key features of 3D genome organization

The spatial configuration of chromosomes within the nucleus, known as the three-dimensional (3D) structure of the genome, is essential for controlling the regulation of genes and preserving the integrity of the genome. This intricate architecture enables the genome to be condensed inside the confined nuclear environment while promoting functional connections across remote genomic sites (Y. Li et al. 2018).

The 3D genome displays a hierarchical framework consisting of multiple key organizational layers:

- *Chromosome Territories*: Inside the cell's nucleus, every chromosome has its own unique territory, which serves as the foundation for the framework of the cell (Kempfer & Pombo 2020).
- *A & B Compartments*: Chromosomes are further differentiated into compartment A, which are rich in genes and euchromatic, and compartment B, which are poor in genes and heterochromatic. Each of these compartments has a distinct activity related to transcription (Cubeñas-Potts & Corces 2015).
- *Topologically Associating Domains (TADs)*: TADs are regions of the genome that exhibit a high degree of spatial proximity and tend to interact with each other more frequently than with other regions. TADs are large areas of chromatin that have a size of several million base pairs and exhibit frequent interactions with one another. These neighborhoods serve as regulators, exerting influence on gene expression within the specific area (Szabo et al. 2019).
- *Chromatin Loops*: Chromatin loops are created through the interactions of distant DNA components either across a TAD or among separate TADs. They facilitate the physical interaction between regulating elements, like the enhancers, and the genes they control, hence impacting the genes' functionality (Mohanta et al. 2021).

1.4.2 Approaches for analyzing 3D genome organization

1.4.2.1 Chromosome Conformation Capture Technologies

- 3C (Chromosome Conformation Capture): The 3C method utilizes cross-linking DNA, enzymatic digestion, and fragment ligation to capture the physical interactions between different areas of the genome. The ligation products are subsequently measured via PCR. This method is utilized to investigate particular relationships across established genomic regions (J. Han et al. 2018).
- Hi-C (Chromatin Conformation Capture): Hi-C is an advanced form of 3C that incorporates next-generation DNA sequencing to identify contacts throughout the whole genome. The process includes cross-linkage, digestion, biotinylation of fragmented ends, the ligation process, and sequencing. Offers a thorough perspective on genomic interactions, uncovering chromosomal territories, TADs, and loop domains. This method

is capable of detecting interactions across many different levels throughout the entire genome without any bias (Belton et al. 2012).

- 4C (Circular Chromosome Conformation Capture): 4C characterizes all interactions between a certain gene and every single component of the genome. Following the 3C technique, inverse PCR is employed for amplifying the ligated products, which are subsequently subjected to sequencing. This tool is valuable for detecting all genomic areas that are associated with a specific locus (Zhao et al. 2006).
- 5C (Chromosome Conformation Capture Carbon Copy): The multiplexed ligation-mediated amplifying technique is used by 5C in order to investigate many interactions at the same time. Initially, a repository of ligation intermediates is compiled, and then high-throughput genome sequencing is performed on the samples. It is suitable for thorough interaction identification inside certain genomic domains (Dostie et al. 2006).

1.4.2.2 Techniques based on microscopy

- FISH (Fluorescence In Situ Hybridization): FISH is an approach that makes use of fluorescently tagged probes to bind to specific sequences of DNA. This allows for visualization of the probes' position inside the nucleus by using fluorescent microscopy. It provides spatial knowledge of the exact position of certain genes or areas, unveiling their spatial arrangement and interactions. Advantages include enhanced spatial precision and the ability to observe certain locations directly (Shakoori 2017).
- Super-Resolution Microscopy: The super-resolution microscopy techniques like PALM (Photoactivated Localization Microscopy) and STORM (Stochastic Optical Reconstruction Microscopy) are able to overcome the diffraction limitation of light. This allows them to provide extremely detailed images of chromatin structure. It enables the observation of the structure of chromatin at a resolution of nanometers, providing intricate insights into the architecture of the genome (B. Huang et al. 2009).
- Live-Cell Imaging: It is a technique that involves the use of fluorescently labeled proteins and complex microscopes to study the movement and changes in chromatin inside alive cells over an extended amount of time. It investigates the kinetic properties of chromatin,

including the motion of specific genetic regions and the establishment/termination of chromatin associations (Shroff et al. 2024).

1.4.2.3 Strategies contingent on sequencing

- *ChIA-PET (Chromatin Interaction Analysis by Paired-End Tag Sequencing)*: It is a method that utilizes chromatin immunoprecipitation (ChIP) with 3C (chromosome conformation capture) to identify connections facilitated through certain proteins. Following the process of cross-linkage and digestion, fragments of DNA are subjected to immunoprecipitation using antibodies specific to a protein of interest. Subsequently, the DNA fragments are joined together using ligation and then sent to sequencing. This application involves the identification of protein-mediated chromosomal interactions, which establish connections between regulatory components and targeted genes. It offers a valuable understanding of the function of individual proteins in the structure of chromatin (Li et al. 2014).
- *HiChIP*: It is a technique that combines the principles of Hi-C and ChIP-Seq. It detects and analyzes interactions among chromatin that are facilitated by certain proteins. This is achieved by isolating chromatin aggregates using immunoprecipitation and then conducting Hi-C on these. This application is used to identify and analyze protein-centric interactions. It offers high-resolution interaction evidence and requires fewer resources compared to ChIA-PET (Mumbach et al. 2016).
- *SPRITE (Split-Pool Recognition of Interactions by Tag Extension)*: The basic idea of SPRITE involves the use of tags and pooling to interact with DNA sections via a sequence of split-and-pool stages, which then proceed via sequencing. This approach does not depend on the use of close vicinity ligation. This tool is used to detect and analyze intricate chromatin structures by identifying relationships between several genomic locations (Quinodoz et al. 2022).

1.4.3 TADs as gene ensembles in gene regulation, disease development, and therapy

The arrangement of the genome spatially is crucial for regulating gene expression. Contemporary chromatin interaction map studies have identified TADs and sub-domains as the fundamental

components of the 3D genome. Multiple studies have noted that TADs effectively insulate a specific section of the genome from the influence of nearby sites. TADs often enable periodic interactions between cis-regulatory parts and their designated promoters, ensuring persistent expression of genes that might not take place otherwise (Tena & Santos-Pereira 2021; Galupa & Heard 2017). TADs have been proposed as a framework for chromatin organization to regulate the nearby cis-regulatory landscapes. They are defined as "extensive genomic regions consisting of multiple long-range sequences of regulation that collectively influence a number of targeted genes". However, the question of whether TADs represent a distinct operational component in the chromatin scaffold has become more complex due to the improved precision of Hi-C experiments. The researchers have discovered nesting structures within the subTAD group by observing the corresponding shielding among them. Undoubtedly, TADs may be utilized to investigate the connections that exist between the three-dimensional genome and gene regulation, as well as their modifications throughout time in connection with growth, disease, and the evolutionary process (James et al. 2024; Okhovat et al. 2023).

1.4.4 Role of chromatin patterns in cancer gene regulation

Cancer is a consequence of the unregulated proliferation of aberrant cells caused by genetic alterations that are acquired or passed down through our ancestors. Every kind of malignancy has a distinct array of molecular changes inside the cancerous cells. Recent technological advancements have facilitated the use of molecular profile analysis, enabling medical professionals to differentiate between molecular changes in cells with cancer and normal cells (Pulumati et al. 2023). Molecular profiling employs many methods to identify biomarkers associated with cancer. These findings provide valuable insights to clinicians on the potential sensitivity or resistance of tumors to therapy. A cancer marker is indicative of the presence of cancer in the body. A cancer signature might originate from the tumor, or it might be a direct physiological reaction to the presence of cancer in the body. Personalized medicine utilizes data with regard to an individual's lifestyle, surroundings, and physiology to avoid, diagnose, and cure diseases. Chromatin-associated proteins have a complex and very context-dependent function in cancer (Pandey et al. 2021; Sharma et al. 2020). Only a few number of chromatin modifications have the ability to independently initiate cancer development. On the contrary, these are often modified in combination with important tumor suppressors and regulators of the cell cycle, such

as CDKN2A and p53 (Pacifico & Leone 2007). Several mutated chromatin molecules have high tissue specificity. However, some chromatin regulatory bodies, such as MLL3/4-UTX from the COMPASS family, might have a wide-ranging impact in suppressing tumors in different types of cancers (Sze & Shilatifard 2016).

1.5 Ensembles of genes centered upon the epigenome

Epigenome-based gene collections are ensembles of genes that are categorized in accordance with their modulation by epigenetic changes. The epigenome comprises chemical changes to the histone proteins and DNA that impact the expression of genes without affecting the core sequence of DNA (Mishra, Pandey, Rawat, et al. 2023). These alterations include DNA methylation, the modifications of histones, and changes in chromatin accessibility, among other factors (Wang & Chang 2018). Gene arrays built on the epigenome may provide valuable information on the regulation systems that control the activities of cells and disease conditions, including cancer.

1.5.1 Major epigenetic modifications

1.5.1.1 DNA methylation

It is the procedure of adding the methyl group onto the 5-carbon of the cytosine residues, usually seen within CpG dinucleotides. It is frequently associated with the inhibition of the process of transcription. Gene expression may be silenced by hypermethylation of the promoter regions, while activation of genes or instability in genomics can result from hypomethylation (Kiselev et al. 2021).

1.5.1.2 Histone modifications

Modifications of histone proteins that occur after they have been translated, such as acetylation, methylation, ubiquitination, and phosphorylation. Such modifications have the ability to either stimulate or inhibit transcription, depending on the specific kind and position of those modification. As an example, the process of adding three methyl groups to lysine 4 of histone H3 (known as H3K4me3) is connected to the activation of gene transcription. Besides, the

incorporation of three methyl groups to lysine 27 of histone H3 (known as H3K27me3) is related with the repression of gene expression (Bannister & Kouzarides 2011).

1.5.1.3 Chromatin accessibility

Chromatin's architectural state, regardless of whether it is closed or open, impacts the availability of DNA to gene transcription factors along with other protein regulators (Chandra et al. 2023). Euchromatin, which refers to open chromatin areas, is often linked with active transcription of genes. On the other hand, heterochromatin, which refers to closed sections, is coupled with gene suppression (Klemm et al. 2019).

1.5.2 DNA Methylation as a key epigenetic modification

DNA methylation is a crucial epigenetic alteration that has a vital function in controlling the expression of genes and preserving the integrity of the genome. Methylation is the procedure of incorporating a group of methyl to the 5-carbon of the cytosine residues, particularly in CpG dinucleotides. These modifications has the potential to impact a wide range of biological functions and is especially significant in relation to growth, disease, and pharmaceutical therapies (Li & Tollefsbol 2010).

1.5.2.1 Significance of DNA methylation-based gene ensembles

- *Transcriptional Repression*: is the process where DNA methylation often takes place on CpG islands situated inside the promoter areas of genes. When these sites undergo methylation, it obstructs the transcriptional machinery's ability to access the DNA, resulting in the suppression of gene expression (Dhar et al. 2021).
- *Context-Dependent Effects*: Although DNA methylation is often associated with the suppression of gene activity, its impact may vary depending on the specific circumstances. For example, methylation occurring inside the gene bodies might be associated with ongoing transcription (Q. Wang et al. 2022) (Q. Wang et al. 2022).
- *Genomic Stability and Integrity*: Genomic stability and integrity are maintained by the process of DNA methylation, which effectively suppresses the activity of repetitive sequences and transposable elements (TEs) in the genome. This prevents their migration and integration into other genomic sites (Zhou et al. 2020).

- *Development and Differentiation*: DNA methylation patterns undergo dynamic regulation throughout development to selectively activate or suppress certain sets of genes in various tissues (Wu & Sun 2006).

1.5.3 DNA methylation in cell-free DNA (cfDNA)

Having found that conventional clinical diagnostic procedures, such as bone marrow or tissue biopsies, are intrusive in nature and result in sample bias, researchers are seeking for alternative molecular biomarkers. When compared to tissue-based studies, liquid biopsy-based detection approaches have been increasingly popular in recent years due to the fact that they are faster and safer (Feng et al. 2019). One approach that is derived from liquid biopsy makes use of cancer traces that are retrieved from cell-free DNA (cfDNA) (Liu et al. 2020). Various research studies have demonstrated multiple applications of DNA methylation based markers in cfDNA (Spector et al. 2023).

1.5.4 Techniques for generating DNA methylation profiles in cfDNA

Developing DNA methylation profiling entails using many techniques that may precisely identify and measure methylation patterns over the whole genome. These approaches differ in their level of detail, rate of processing, and particular uses (Figure 1.3). Below is a summary of the most frequently employed techniques:

1.5.4.1 Methods based on bisulfite conversion

This approach is considered the most reliable and widely accepted technique for profiling DNA methylation (e.g. WGBS, RRBS etc.). The process of bisulfite treatment transforms cytosines (C) that are not methylated into uracil (U), which is then interpreted as thymine (T) during the sequencing process. Cytosines that have undergone methylation, namely 5mC, remain unaltered (Q. Li et al. 2018).

1.5.4.2 Methods Based on Arrays

Utilizes microarrays to measure the level of DNA methylation at particular CpG sites throughout the whole genome. It offers a high rate of data processing, is cost-efficient, and is well-suited for

extensive research projects employing standardized data. It is commonly used in the field of epidemiology, cancer investigation, and extensive cohort studies (Schumacher et al. 2006).

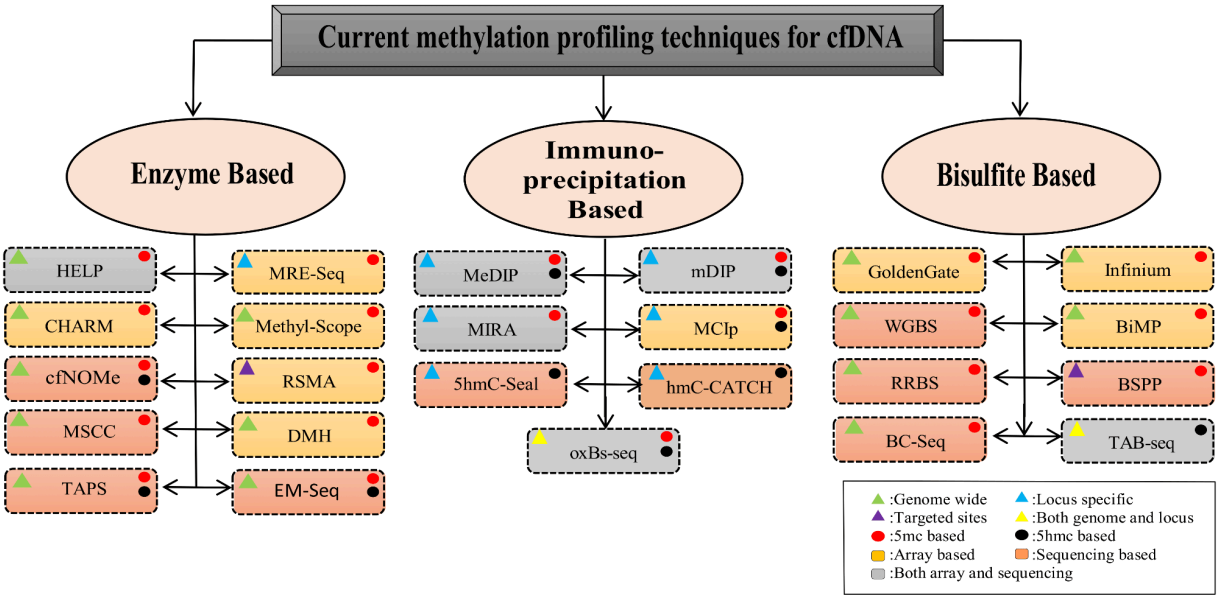

Figure 1.3: A comprehensive examination of methods for characterizing DNA methylation that is also applicable for identifying cell-free DNA (cfDNA). The circular and triangular icons provide more information about various techniques. The expanded form of abbreviations for different methods are as such:- HELP: HpaII-tiny fragment enrichment by ligation-mediated PCR, CHARM: comprehensive high-throughput arrays for relative methylation, cfNOMe: cell-free DNA-based Nucleosome Occupancy and Methylation profiling, MSCC: methyl-sensitive cut counting, qPCR: Quantitative polymerase chain reaction, TAPS: TET-assisted pyridine borane sequencing, MRE-Seq: methylation restriction enzyme sequencing, RSMA: methylation-sensitive restriction enzyme-based assay, DMH: differential methylation hybridization, ddPCR: droplet digital PCR, EM-Seq: Enzymatic Methyl-seq, MeDip: methylation DNA immunoprecipitation sequencing, MIRA: methylated CpG island recovery assay, mDIP: methylated DNA immunoprecipitation, oxBs-seq: oxidative bisulphite sequencing, WGBS: whole-genome bisulphite sequencing, RRBS: reduced representation bisulphite sequencing, BC-Seq: bisulphite conversion followed by capture and sequencing, BiMP: bisulphite methylation profiling, BSPP: bisulphite padlock probe, TAB-seq: TET-assisted bisulphite sequencing).

1.5.4.3 Methods based upon enrichment

These techniques selectively amplify methylated segments of DNA prior to sequencing, enabling researchers to target specific areas of interest and minimize the expenses associated with whole-genome bisulfite sequencing (e.g. MeDIP-seq) (Chiou & Bergey 2018).

1.5.4.4 Methods for Long-Read Sequencing

It eliminates the requirement for the conversion of bisulfite by directly detecting DNA methylation using the technique of nanopore sequencing. It offers the possibility of sequencing with long reads and permits real-time monitoring of DNA methylation and sequence. It is perfect for integrating methylation alongside genetic variations and researching methylation patterns in repetitive regions (Y. Liu et al. 2021).

1.5.5 DNA Methylation role in disease diagnostics and therapy

1.5.5.1 Cancer Diagnostics

Early Detection and Screening: Patterns of DNA methylation may be used for the purpose of early detection as well as screening so that tumors can be identified at an earlier stage. Using liquid biopsy methods, for instance, it is possible to determine whether or not certain cancer suppressor genes have been hypermethylated in body fluids such as blood or urine (Kwon et al. 2023).

Example: The methylation state of the SEPT9 gene is utilized in blood tests for the purpose of discovering colorectal cancer during an earlier stage (Y. Wang et al. 2018).

1.5.5.2 Epigenetic Therapies

DNA Methyltransferase Inhibitors (DNMTis): These drugs hinder the activity of DNA methyltransferases, which results in the reactivating of genes that were previously silenced. They are mostly employed in the management of specific kinds of cancer (Zhang et al. 2022).

Example: Myelodysplastic syndromes and acute myeloid leukemia are both treated with DNMTis named azacitidine and decitabine, both of which have been authorized by the Food and Drug Administration. DNA methyltransferases are covalently trapped by DNMTis, which is integrated within DNA amid the replication process. This results in the destruction of DNA methyltransferases, followed by DNA demethylation (Sorrentino et al. 2021).

1.5.5.3 Personalized Medicine

Targeted Therapy: Knowledge of the precise patterns of methylation of a patient's tumor could assist in the selection of personalized therapies (Gupta et al. 2023). Methylation biomarkers may

be utilized to assess the efficacy of therapy and make appropriate modifications to treatment strategies.

Example: Patients diagnosed with glioblastoma who have hypermethylation in their MGMT gene get greater advantages from the use of alkylating drugs (Qi & Tan 2020).

1.5.5.4 Combination Therapies

Synergistic Approaches: The use of epigenetic medications with other therapies, such as chemotherapy and immunotherapy, may augment the effectiveness of therapy. An instance of enhancing anti-tumor reactions is the amalgamation of DNMTis together with immune system checkpoint inhibitors (Zhang et al. 2022).

Example: Current clinical studies are investigating the potential of combining azacitidine along with inhibitors of PD-1 and PD-L1 for the treatment of different types of malignancies (Yao et al. 2021).

1.6 Scope of thesis work

Single-gene centric methodologies have mostly proven inadequate in tackling the intricacies of diseases due to their tendency to overgeneralize the complicated and interrelated structure of biological systems. Diseases, particularly complex disorders such as cancer, diabetes, and neurological conditions, arise via multifactorial interactions that include a multitude of genes, regulatory components, and environmental variables. By concentrating just on individual genes, the extensive network of associations that provide the foundation of cellular functioning and disease causes is disregarded. This reductionist methodology is unable to fully comprehend the redundancy and resilience inherent in biological systems, where numerous genes may compensate for one another, and regulatory structures are capable of adapting to perturbations. Moreover, research focusing on individual genes fails to consider the ever-changing nature of gene expression and regulation, which are regulated by epigenetic changes and signaling networks. Hence, comprehending and accurately focusing on diseases need a comprehensive systems biology methodology that takes into account the whole range of genetic, epigenetic, and environmental factors within the wider biological framework.

Furthermore, despite significant technological progress, effectively addressing illnesses completely using a systems biology approach continues to be difficult owing to many crucial

variables. Biological systems exhibit a high level of complexity and dynamism, characterized by intricate regulatory networks that pose challenges in terms of modeling and comprehensive understanding. High-throughput methods produce huge quantities of data across several biological levels, and the integration of these diverse datasets necessitates the use of modern bioinformatics programs and evolving computational models. Moreover, the presence of genetic diversity among patients, together with the impact of epigenetic and environmental variables, adds to the complexity of diseases and makes it challenging to find universally effective treatment strategies. Obstacles in translating medical advancements, such as the time-consuming process of assessing and approving novel medicines, as well as difficulties in effectively delivering therapies, impede development. The broad application of systems biology techniques in clinical settings is limited by economic restrictions, as well as the requirement for specialized infrastructure and interdisciplinary knowledge. To tackle these issues, it is necessary to make more progress in computational and integrative methodologies, strengthen customized medical strategies, and improve infrastructure and multidisciplinary cooperation. In order to address these constraints, it is imperative to develop more sophisticated, integrated and customized diagnostic techniques that can provide an exhaustive understanding of disease biology, resulting in improved diagnosis and treatment opportunities.

1.6.1 Using single-cell expression profiles to associate pathways to diseases and make inferences about potential drugs

In the domain of disease identification and therapy, genes commonly work together in a number of different pathways that provide a valuable understanding of the underlying mechanisms that are involved in a variety of disorders. In this work, we have investigated the challenging task of establishing direct connections between pathways and diseases. We introduce sci-PDC, a technique that makes use of data on the expression of a single-cell in order to infer links between diseases, cell types, and pathways connecting them. The use of this methodology provides insightful insights into the underlying causal relationships that exist between these factors and offers an opportunity to influence the area of precision medicine.

1.6.2 Using tumor specific single-cell profiles to explore relationships between biological pathways and cancer hallmark properties for precision therapy

Another set of gene ensembles, cancer hallmarks, help identify cancer early by revealing its core properties. The environment of cellular subtypes may affect how these signature traits emerge and behave. Cellular distinctiveness stresses the need for precision medicine via personalized therapies that target malignant cells' genetic vulnerabilities. Due to its snapshot nature, scRNA-seq-based studies must add a temporal component to adequately understand association direction and poising. This work shows how to directly link cancer hallmark properties and pathways. The method uses scRNA-seq and RNA-velocity characterization. This strategy may help us understand how drugs influence cells in different tumor microenvironments and reveal signature traits and pathways unique to each malignant cell type.

1.6.3 Utilising gene ensembles defined by chromatin domains with transcriptomics to understand drug-response and phenotypic heterogeneity in cancer cells

In the context of genomic conformations, groups of genes often cooperate via spatial arrangement, hence impacting cellular processes and phenotype. These complex chromatin structures help to combine genes and associated regulatory components, allowing for coordinated expression of genes and control. This study utilized an innovative method that used TAD activity to investigate the diversity of cancer and the ability of patients to respond to drugs. Our study's results strongly indicate that TAD activity could potentially used as a predictive biomarker for determining the way a patient will respond to a treatment and estimate their chances of survival when there is tumor heterogeneity.

1.6.4 Introspecting computational challenges associated with cell-free cancer diagnosis using ensemble of DNA methylation markers

Epigenetic markers, similar to DNA methylation, often have a significant impact on the regulation of transcription as well as genomic conformations. Disordered regulation of epigenetic pathways may undermine the physiological authenticity of gene ensembles, potentially resulting in multiple diseases. Our study provides an unambiguous description of the computational

challenges involved in analyzing DNA methylation. These challenges stem from the inbuilt prejudices seen in different profiling methods. This work evaluates the effectiveness of the deconvolution process and machine learning techniques in analyzing the methylation state of cfDNA, suggesting its possible implementation in the earlier detection of cancer.

In summary, our proposed methods may harness the combined impacts of gene ensembles via various genomic and epigenomic patterns (Figure 1.4). This can lead to a comprehensive understanding of disease biology, hence improving diagnostic and therapeutic strategies.

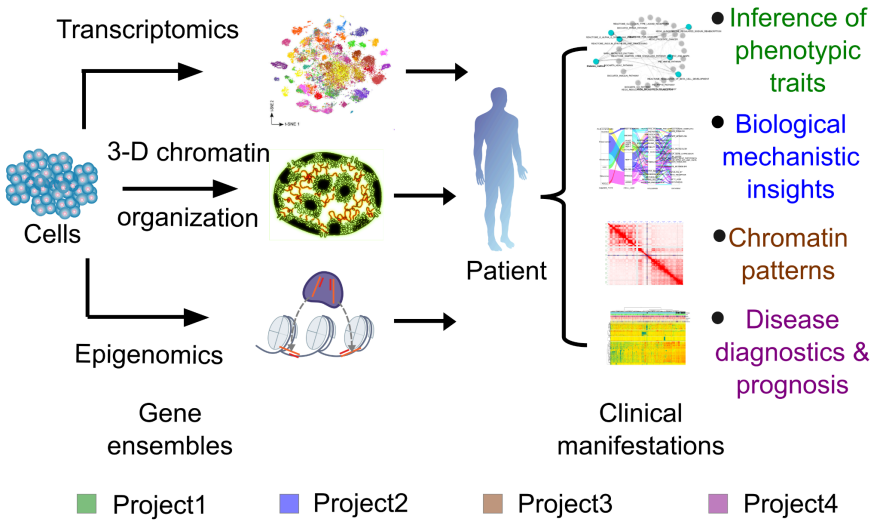

Figure 1.4: Coherence among the projects. Figure illustrating the manner in which our suggested strategies may be able to exploit the combined impacts of gene ensembles through a variety of genomic and epigenomic patterns. This can lead to a complete knowledge of disease biology, which in turn can improve diagnostic and treatment efforts.

CHAPTER 2

USING SINGLE-CELL EXPRESSION PROFILES TO ASSOCIATE PATHWAYS TO DISEASES AND MAKE INFERENCES ABOUT POTENTIAL DRUGS

2.1 Introduction

2.1.1 Dynamics of pathways in disease progression and therapy

Throughout history, diseases have consistently posed a significant burden on contemporary human civilization. Despite the present focus of research and medical healthcare facilities on diagnosing and treating diseases, the efforts to prevent and develop effective therapies remain mainly ineffective (Hofker et al. 2014). Although there are several high throughput technologies available, it is difficult to understand disease-related biological pathways. Possible factors contributing to the situation may include ambient noise, variations among individuals, environmental circumstances, problems with batches, and the intricate nature of the diseases. In recent times, there has been a specific emphasis on examining the fundamental process of diseases via the use of genetic pathway activity or enrichment analysis (Emmert-Streib & Glazko 2011; de Leeuw et al. 2018) . Given that active genetic pathways often govern cellular responses to external stimuli, analysing their activity profile might provide more logical insights into the underlying processes of different disorders (Liu & Chance 2013). Researchers often attempt to gain understanding of disease-related pathways by analysing mutations revealed by genome-wide association studies (GWAS)(Lappalainen & MacArthur 2021).

2.1.2 Leveraging cell type specificity for precision medicine

Specific cell types often exhibit causative or consequential links with diseases. It is possible to find more context-specific disease-related genes by using cell-type specificity (Stoney et al. 2018; Guan et al. 2020; Ju et al. 2013; Mar et al. 2011). Furthermore, the advanced capabilities of scRNA-seq technologies make it possible for scientists to analyse early disease onset, progression, treatment resistance, and immune escape by examining cellular subpopulations in disease microenvironments (Stetson et al. 2021; Goldman et al. 2019; Aissa et al. 2021; Wang et al. 2020; Jiang et al. 2021; Nomura 2021). Additionally, there has been a consistent increase in the number of disease samples with single-cell expression profiles (Zhao et al. 2021). After identifying genes that are expressed differently, many widely used tools have been developed for analysing pathways in bulk gene-expression profiles. These tools include EnrichNet (Glaab et al. 2012), WebGestalt (Zhang et al. 2005), DAVID (D. W. Huang et al. 2009b), GSEA (Subramanian et al. 2005) , g:Profiler (Raudvere et al. 2019) , and others. The analysis of pathway function is approached in a distinctive manner by each of these strategies. Furthermore, some techniques such as Consensus Pathway Analysis (CPA) enable researchers to do comparative analysis of outcomes across various approaches and studies (Nguyen et al. 2021). In addition, many approaches such as UniPath (Chawla et al. 2021) , PAGODA (Fan et al. 2016) , and AUCcell (Aibar et al. 2017) have been developed to convert single-cell expression patterns into gene set enrichment scores for each cell. Nevertheless, while utilizing single-cell expression profiles, the majority of the previously suggested methods often struggle to appreciate the direct relationships between pathways and disorders. A number of variables, including the transitive impact of correlation and environmental confounding factors like age, gender, and environment, may contribute to indirect relationships that show up significantly when using conventional techniques (Feizi et al. 2013) . Therefore, it is necessary to develop techniques that can understand the direct relationships between pathway gene sets and diseases with use of single-cell expression profiles. This will allow us to uncover the variations in the connection between diseases and pathways in various cell-types.

The objective of this work was to uncover hidden information in large collections of single-cell profiles by determining the direct relationships between pathways (groups of genes) and diseases utilizing patterns of gene expression at the single-cell level. It has always been a difficult and open issue to infer direct connections among attributes using their scores in various samples. The

characteristics examined in this research are distinct, and the data sets used are acknowledged to have a limited amount of information; as a result, the difficulty level is elevated. Even in data sets that are not sparse, approaches that rely on co-activity or co-expression often uncover indirect correlations between attributes. Therefore, we used a methodical approach by evaluating several techniques and implemented our modifications to ultimately conduct a comparison based on established relationships between pathways and diseases. To begin, we demonstrate the way we enhance the inference among diseases and pathways with cell type context. Next, we evaluate the hypothesised relationships between human and mouse pancreatic beta cells and the disease (diabetes type 2). The approach we use is referred to as sci-PDC, which stands for single-cell based inference of the association between pathway, disease, and cell type. Through this approach, we have also explored the potential for identifying alternative drugs for treating diseases.

2.2 Material And Methods

2.2.1 Utilising single-cell attributes for disease and pathway inference

At first, we conducted an initial assessment by examining the pathway and disease activities of single-cell expression patterns of human pancreatic cells. The data was extracted from the work conducted by Segerstolpe *et al*., which includes single-cell RNA-seq dataset of human pancreatic cells from healthy controls and patients with type 2 diabetes. Single cells were prepared using the Smart-seq2 protocol and sequenced on an Illumina HiSeq 2000. The data providers had already normalized the dataset in the FPKM format (Segerstolpe et al. 2016). The acquisition of gene ensembles or gene sets for certain pathways and diseases was accomplished using Enrichr (Jensen-disease gene set) (Chen et al. 2013) and MSigDB (Subramanian et al. 2005) respectively. The R package UniPath was utilized to convert single-cell expression data into gene set enrichment scores (Chawla et al. 2021). It has been shown that UniPath-based scores outperform other comparable methods, such as PAGODA (Fan et al. 2016), AUCell (Aibar et al. 2017), GSVA (Hänzelmann et al. 2013), etc., in terms of accuracy and speed. The combined p-value, or enrichment score, for a gene set was determined by having at least five non-zero expressions. The recalibrated scores were used to analyze disease-pathway co-occurrence using Spearman correlation.

Moreover, it is to be noted that here, the enrichment score or activity score denotes the degree of enrichment or activity associated with a certain gene ensemble. This score, when applied to a gene ensemble of a pathway, is referred to as the pathway score, indicating the importance of gene participation in that pathway. In a similar manner, when utilized for a disease gene ensemble, the enrichment or activity score is designated as the disease score, reflecting the extent of gene involvement in the disease context. Fundamentally, these phrases represent variants of the same core notion, emphasizing the activity or enrichment of a certain group of genes inside a given biological or disease-related process.

2.2.2 Normalization utilizing single-cell transcriptome atlases

2.2.2.1 Collection of data for normalization

Human cell landscape (HCL): The human cell landscape is a collection of human single-cell transcriptome data obtained via collection of more than 700,000 cells. There are 1462 distinct cell types represented by these cells, which originate from over fifty different human tissues. The HCL is accessible at http://bis.zju.edu.cn/HCL/. We used UniPath to transform the individual cellular expression of all cells into pathway scores, which were then employed for further research (Han et al. 2020).

Mouse cell atlas (MCA): The mouse cell atlas, accessible at http://bis.zju.edu.cn/MCA/index.html, provides single-cell expression profiles of mice. It encompasses more than 520,000 single cells through 1357 distinct cell types across ten various mouse tissues. The scRNA-seq data of mice was converted into activity scores using UniPath (X. Han et al. 2018).

2.2.2.2 Standardization of correlation values and screening parameters

The enrichment scores for a certain disease and pathway gene set were determined using adjusted p-values derived from UniPath. Data preprocessing and filtering were done after the scores of each cell in the corresponding tissue were determined to prevent noisy and irrelevant connections. The gene set for the type of cell was considered if a minimum of 10% of the cells had a disease score and pathway score below 0.25 and there were at least ten pathways. Afterwards, we conducted a co-occurrence analysis on the filtered enrichment scores containing

adjusted p-value in order to estimate significant associations between pathway and disease gene sets.

Three criteria were established to evaluate the co-occurrence of disease and pathway which were rank specificity, specificity score, and correlation. The log transformed enrichment scores were used to compute the Spearman correlation. The specificity score was calculated using a null model-based permutation test to assess the relevance of the variation in disease-pathway co-occurrence. The disease-pathway values of correlation were normalized based on rankings in the context of cell types to achieve rank specificity. In the case of diabetes, we first determined the correlation rankings of complete pathway gene set with the diabetes gene set specifically in pancreatic beta cells. Furthermore, we computed the rank of pathways with diabetes correlation across 1462 cell types from human landscape dataset (Han et al. 2020). We computed the cell specificity score for each pathway and diabetic gene set by determining the proportion of 1462 cell types where the relevant rank appeared superior. Afterwards, disease-pathway connections that had a rank specificity below 0.25, a specificity score below 0.05 and a correlation of above 0.3 were selected to undergo further analysis. In order to ensure that there is minimal overlap across gene sets, we additionally computed the Jaccard index amongst them.

2.2.3 Inferencing direct associations using probabilistic graph model

2.2.3.1 Inference through Bayesian networks

Our dataset may include complex associations including mediation, inter-causal dependence, and some other confounding variables, thus we used Bayesian network (BN) structure learning across pathways to do association analysis.

For any set of *N* feature or the random variables $\{X_1, X_2, \ldots, X_N\}$, joint probability distribution (*JPD*) is given by, as per the chain rule of probability:

$$P(X_1, X_2, \ldots, X_N) = \prod_{i=1}^{N} P(X_i \mid X_{i+1}, X_{i+2}, \ldots, X_N) \quad (1)$$

Provided a directed acyclic graph (*DAG*) of the features, the parents of any feature X_i represented by $Pa(X_i)$ is the group of features in the *DAG* that were found to have a directed edge to X_i. For a Bayesian network induced by a *DAG*, the joint probability distribution factorizes as

$$P(X_1, X_2, \ldots, X_N) = \prod_{i=1}^{N} P(X_i \mid X_{i+1}, X_{i+2}, \ldots, X_N) = \prod_{i=1}^{N} P(X_i \mid Pa(X_i)) \quad (2)$$

For instance, provided a two node Bayesian network with configuration $X_1 \rightarrow X_2$, the joint probability distribution is

$$P(X_1, X_2) = P(X_1)\, P(X_2 \mid X_1). \quad (3)$$

In this study, we implemented bnlearn to deduce BN by utilizing the enrichment scores of different gene sets (Nagarajan et al. 2014). We used the proportion of edge overlaps across several datasets to optimise the values of the parameters in bnlearn tools. The Bayesian Information Criterion (BIC) scoring function and Hill Climbing (HC) structure learning approach are used here. Next, in order to enhance the resilience of the graphical model, we used bootstrapping to train and applied majority voting on 1001 Bayesian Networks (BNs). Quantile-based discretization of the data was used to shorten the computing time required to train the network.

2.2.3.2 Inference using Markov network

Sparse and high-dimensional undirected graphical representations between features represent Markov networks. We estimated the Markov network between features or gene sets using XMRF package (Wan et al. 2016).

The Wan *et al.,* XMRF technique assumes that conditional distributions on nodes conform to univariate exponential families by using Markov random field models. With a vector $X = \{X_1, X_2, \ldots X_p\}$ and every component X_i having value, we construct an undirected graph G=(V, E) via p nodes representing the p features. The goal is to ensure that conditional distributions satisfy Markov independence assumptions concerning each node in the graph. The exponential Family Graphical Model is formed by the joint density, taking into account the exponential family distribution of X_s on nodes conditioned at all remaining nodes and the Hammersley-Clifford theorem.

$$P(X) = exp\left\{ \sum_{s \in V} (\theta_s T(X_s) + h(X_s)) + \sum_{(s,t) \in E} \theta_{st} T(X_s)\, T(X_t) - A(\theta) \right\} \quad (4)$$

Here $T(.)$ represents the sufficient statistics function and θ_{st} highlights the weight of edge between node *s* and *t*. $A(\theta)$ shows the log normalization term.

The XMRF approach uses projected or proximal gradient descent to carry out neighbourhood selection graph estimation. It individually scans the neighbourhood of each node. With parameters stability = "STAR" (Stability Approach to Regularisation Selection), N=1001 iterations, nlams=15 (required number of lambdas for regularisation), and beta=0.3 (network sparsness parameter), the 'GGM' Markov random field (MRF) approach was used.

2.2.4 Literature based high-throughput validation

To gain confidence within the resultant disease-pathway associations, we made use of PubMed abstract-based validation. We also utilised this validation to examine the effectiveness of several applicable methodologies, including rank specificity, specificity score, Bayesian, correlation, and Markov. The disease phrase and its accompanying pathway keywords were utilized as input in such scenario. Optimal matching of pathway phrases with a particluar disease in a potentially relevant abstract was achieved by processing the terms to eradicate stop words. The PubMed database was queried in research paper abstracts from 1990 to 2022 using the Bio.Entrez tool (Chang et al. 2010). By programmatically querying PubMed with specific terms, the Bio.Entrez package facilitates high-throughput PubMed literature searches. It retrieves large sets of article IDs and subsequently downloads metadata (e.g., titles, abstracts) for numerous publications at once. This procedure enables users to automate extensive literature searches and data extraction in an efficient, batch-processing manner. The same settings were also used to search for co-occurrence of random pathway words for the same disease as a control. We chose the top 10 pathways via each technique and contrasted their abstract occurrences with regard to random pathways with the help of a box plot in order to compare the performance of the multiple strategies used to calculate disease-pathway connections.

2.2.5 Variations with regard to species

We used a similar methodology, as previously mentioned, to delve further into the many facets of disease-pathway relationships in pancreatic-beta cells across different species. The dataset consisting of UMI counts from human and mouse beta pancreatic cells (GSE84133) was

analyzed for co-occurrence using probabilistic graphical models utilizing the Bayesian and Markov models after translation into pathway space.

2.2.6 Age-related Variability

Additionally, the age-dependent change of disease-pathway associations in human cells was examined using the same approach. The data was acquired from the GEO database (GSE81547), which includes gene expression data of pancreatic beta cells from two age groups: young-adult (age: 21–22 years) and elderly (age: 38–54 years) (Enge et al. 2017). Here, we investigated the pathway connections associated with diabetes mellitus in order to gain a deeper understanding of the impact of age.

2.3 Results

We assessed the interdependence between pathways and diseases by analysing the co-occurrence of enrichment scores from respective gene sets. Figure 2.1 displays a concise representation of the process implemented by sci-PDC. Based on their single-cell expression patterns, enrichment scores of pathway and disease gene sets were generated for each cell with help of UniPath. We utilized single-cell expression patterns of cells belonging to the same category to assess the co-occurrence of enrichment of pathway and disease gene sets. While Spearman's correlation might provide a simple means of measuring co-occurrence, it can also indicate indirect connection as a result of possible confounding variables, mediation, and inter-causal dependence (Xiao et al. 2022) . After additional investigation, we discovered that filtering procedures and methods based on Probabilistic Graphical Models (PGM) may be more useful for determining direct relationships (Friedman 2004). We illustrate our analysis in addition to sci-PDC implementation via the example of diabetes disease.

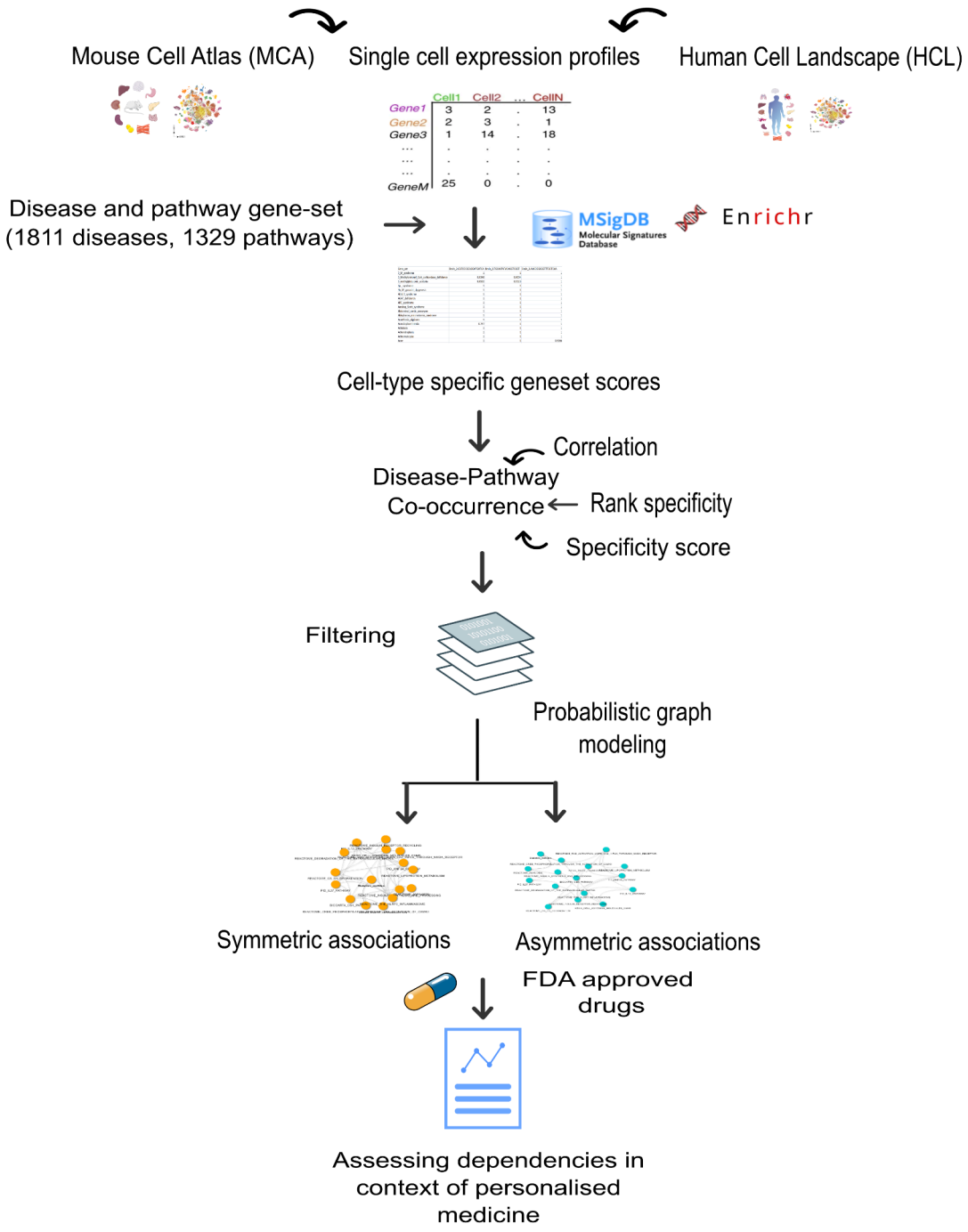

Figure 2.1: Workflow of the pipeline. Initially, we transformed the single-cell expressions obtained from the human-cell landscape (HCL), mouse cell atlas (MCA), and several more single-cell collections of datasets into matrices representing gene set enrichment scores. Here, Pathways are equivalent to rows, whereas each column indicates one individual cell. We compute the associations between pathway and disease gene set scores by analyzing cells of every cell type individually. In addition, we standardize the correlations along with rankings by

comparing them to the comparable values of numerous cell types utilizing HCL or MCA files. In order to develop the probabilistic graph model (PGM), we choose pathways that exhibit good correlation, specificity (shown by a low p-value), & rank-specificity (indicated by a low rank based p-value) for both disease and the type of cell. The sci-PDC program offers the choice of Markov and Bayesian network inference methods. These methods may be used to estimate direct dependencies while gaining a deeper understanding of the best possible drug for a certain condition.

2.3.1 Inferring the relationship between diseases and pathways

Our preliminary examination of the single-cell expression data of human beta-pancreatic cells uncovered a plethora of intriguing findings regarding the associations between diabetes and pathways (Segerstolpe et al. 2016). The top ten pathways that showed the highest correlation with diabetes mellitus within human pancreatic beta cells were very significant. It is well recognised that diabetes mellitus happens due to problems in the production, secretion, and survival of beta cells in the pancreas that produce insulin (Fu et al. 2013; Cantley & Ashcroft 2015; Kahn 1998). Nevertheless, the process of insulin production and its related pathways may be influenced by several other elements, including as hormones, nutrition, physical activity, and the environment. As a result, further links between diabetes and other conditions were examined by analysing the association between disease pathways. For instance, ghrelin has been recognised as a important regulatory factor of energy homeostasis and appetite, and it has been observed to be impaired in metabolic syndromes, type-2 diabetes mellitus, and obesity (Pulkkinen et al. 2010; Alamri et al. 2016).

Although the preliminary findings were consistent with the literature, we discovered that the relationship between gene ensembles of pathways and diseases might be affected by several factors, such as the count of genes in the gene sets. When there is a large overlap in the number of genes between two groups of genes, the correlation may consistently be strong in any kind of cell. Therefore, in order to account for these covariances, we utilized correlation analysis between pathway and disease gene sets utilising single-cell expression profiles from the null model of UniPath. Therefore, we determined the specificity score for a cell type by comparing the connection between pathway and disease gene set with the correlation obtained from one thousand subsets of cells through the null model. In addition, we standardised the correlation values between pathways and a gene set associated with a disease for each type of cell by comparing them to the corresponding values in different cell types. Consequently, we eliminated

the influence of the covariate by using two methods. Initially, we normalised correlation values against null models. Subsequently, we compared the rankings of these values in comparison to different cell types. We followed identical steps for all mouse cell types, using the mouse cell atlas as a reference.

2.3.2 Modelling cell type specific networks

The purpose of the sci-PDC normalisation phases was to enhance the estimation of cell-type specificity in the relationship across pathways and disease. Figure 2.2A displays the correlation and p-values, which are determined by rank specificity, for a couple of pathways and their association with diseases in different cell types. It is clear from the outcomes shown in Figure 2.2A that there are significant disease pathway connections within certain cell types. Our results are consistent with previous findings that thalassemic erythrocytes have aberrant lipid levels per cell (Rachmilewitz et al. 1976; Kalofoutis et al. 1980). In a similar vein, it is also shown that astrocytes, pyruvate metabolism, Parkinson's disease, and the TCA cycle are related(Anandhan et al. 2017) . Figure 2A demonstrates the association between liver cirrhosis and apoptosis in hepatocytes (Malhi & Gores 2008) . Additionally, diabetes mellitus is often associated with disrupted TCA cycle in pancreatic beta cells (Haythorne et al. 2019). In addition, we examined the relationships between pathways and many additional disorders, such as Thalassemia, hyperuricemia, hemolytic anaemia, and diarrhoea. We next conducted a thorough validation of the resulting pathways using high throughput literature-based methods (Figure 2B). Figure 2.2B indicates that the paths generated using the sci-PDC pipeline have stronger backing from literature compared to randomly generated null pathways for the same disease.

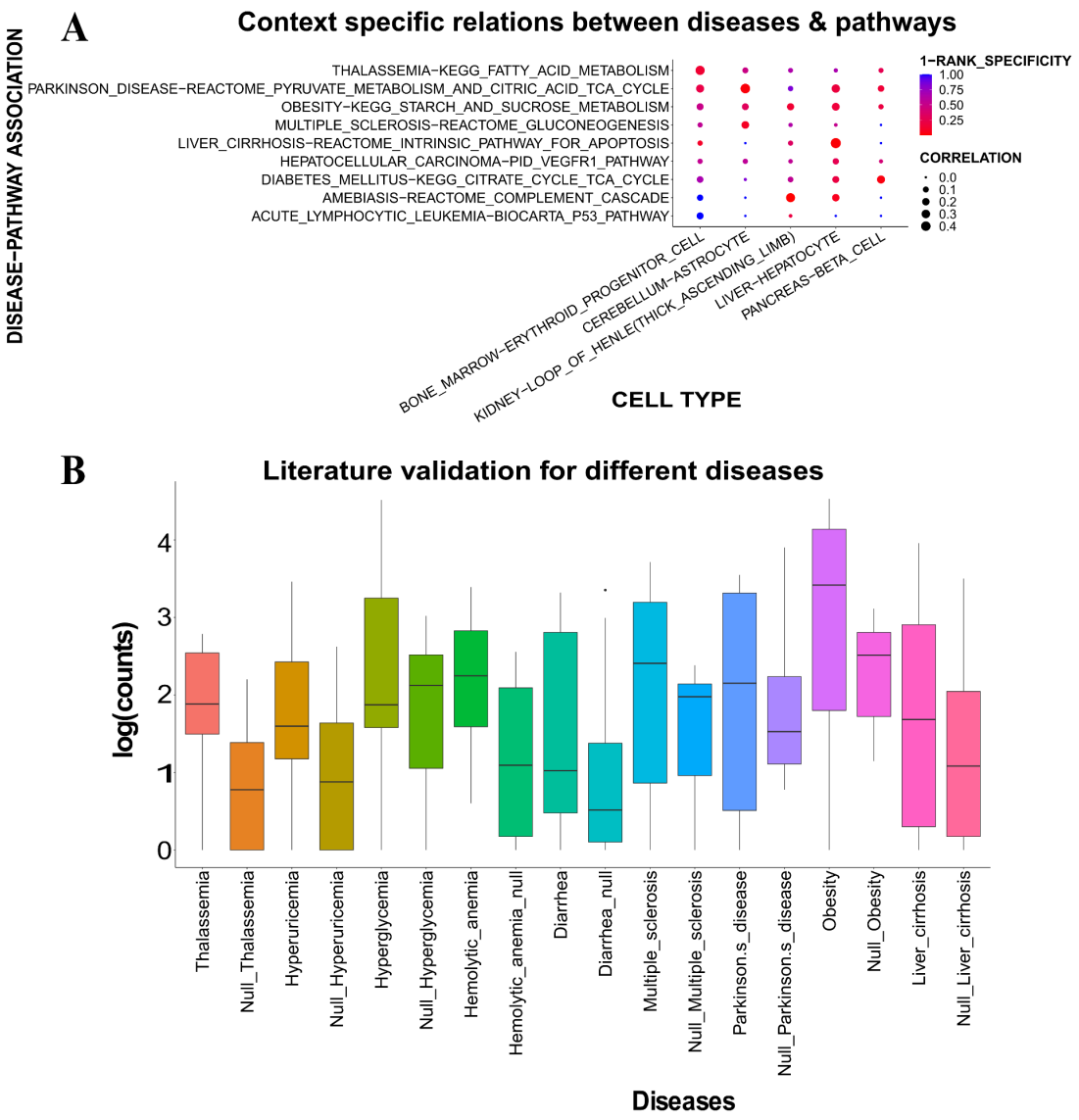

Figure 2.2: The objective of the study was to determine the relationship across pathways and disease for human cells by normalizing the coefficient of correlation employing a comprehensive database of single-cell transcriptome profiles. **A)** Connections between pathways and diseases were observed in five different cell types. **B)** High-throughput validation is performed through leveraging co-occurrence within PubMed abstracts to predict connections between pathways together with diseases. The y-axis represents the frequency of occurrences where the disease and anticipated pathways are associated in the abstracts. The findings are analyzed to determine the occurrence of random correlations using a null model between pathways and the same diseases.

While the significance and specificity calculations revealed several significant and well-established connections between different diseases and pathways, it is important to note that

these associations may only be linear and might potentially be indirect as well. Therefore, in order to examine the non-linear and direct relationship (without any conditional reliance) between diseases and pathways, we conducted a study using Bayesian and Markov model-based methods. Bayesian network inference emphasises the presence of one-way links in the network, whereas the Markov model is useful for representing two-way relationships that are not specifically directed. Figure 2.3A illustrates the pathway associations of diabetes mellitus in both the Bayesian network and Markov model. Figure 2.3B illustrates the weight of connections between nodes related to diabetes in the Bayesian network of the human pancreatic-beta cells. Among all other pathways, the HNF3B route has been identified as the most essential pathway in terms of edge strength in the Bayesian network. FOXA2, formerly referred to as HNF3B, has been discovered to impact many pathways associated with insulin. Lantz et al. have shown that a mutation in FOXA2 leads to an increase in insulin production and a complete loss of insulin response to glucose stimulation (Lantz et al. 2004). This mutation has been identified as a major factor in the progression of type 2 diabetes mellitus (Wolfrum et al. 2003). Similarly, studies have shown that STAT3 plays a significant role in the progression of insulin resistance and leads to damage to islet beta cells, ultimately leading to decreased insulin output (Mashili et al. 2013; Velayos et al. 2017; Yang et al. 2017).

Similarly, inside the Markov model, the pathway known as amine compound SLC transporters was assigned the most significance compared to other pathways related to diabetes. Several mutations related with SLC have been shown to impact insulin homeostasis and increase the risk of type 2 diabetes mellitus (Zhang et al. 2019; Lin et al. 2015; SIGMA Type 2 Diabetes Consortium et al. 2014). In addition to other factors, the CREB pathway has been identified as a significant regulator of diabetes mellitus in Markov model. Numerous studies indicate that the failure of CREB's negative feedback regulation can be a major factor contributing to beta-cell dysfunction in type 2 diabetes (Blanchet et al. 2015; Cho et al. 2012) .

In addition, a network-based analysis was conducted on Alzheimer's disease using cerebellar astrocyte cells and on Thalassemia utilising erythroid progenitor cells. A few pathways, such the PPAR (Heneka et al. 2011) and RIG-1 (retinoic acid-inducible gene-I) mediated IFN alpha-beta (de Rivero Vaccari et al. 2014) (Figure 2.3C) pathways, seemed to be significant for Alzheimer's disease based on Bayesian modelling predictions. The Bayesian modelling identified many significant pathways linked with Thalassemia, including FGFR signalling(Al-Hakeim &

Alhillawi 2018) and erythrocyte membrane transport (Olivieri et al. 1994) (Figure 2.3D). A comparable network-based methodology has also been employed for many more diseases, such as Obesity, Parkinson's disease, Multiple sclerosis, Pancreatic cancer and Diarrhoea.

2.3.3 Literature based validations of disease pathway associations

In order to verify the correctness of the findings from the sci-PDC analysis, we conducted a search to identify instances when disease and pathway related phrases appeared together in the literature. A literature based validation was conducted by looking for pathway phrases that co-occur with a particular disease in the PubMed abstracts. A comparison was conducted on a group of diabetes related pathways using several approaches of associating disease with pathways, such as rank specificity, specificity score, Bayesian inference, correlation, and Markov network inference. Our analysis revealed that the pathways identified using each of the approaches outperformed the null distribution. This suggests that there is a strong indication of disease-pathway relationships. Furthermore, while comparing other association strategies, it was shown that the links derived from the Bayesian graph of filtered pathways of diabetes mellitus performed better than other techniques in terms of literature authentication. The differential enrichment of the pathway scores between normal and diabetic pancreatic beta cells was also computed. The pathways that showed higher enrichment scores in diabetic pancreatic beta cells have less empirical evidence supporting their association with diabetes compared to the findings obtained using our technique (Figure 2.4). The findings indicate that the variation in pathway enrichment in diseased cells may be attributed to indirect relationships. However, the Bayesian network based method revealed a more immediate association with diseases.


Bayesian
Markov
A
REACTOME_SYNTHESIS_OF_BILE_ACIDS_AND_BILE_SALTS_VIA_7ALPHA_HYDROXYCHOLESTEROL
REACTOME_PLATELET_CALCIUM_HOMEOSTASIS
REACTOME_SYNTHESIS_OF_BILE_ACIDS_AND_BILE_SALTS
REACTOME_ION_CHANNEL_TRANSPORT
REACTOME_GPCR_LIGAND_BINDING
Diabetes_mellitus
REACTOME_AMYLOIDS
REACTOME_LATENT_INFECTION_OF_HOMO_SAPIENS_WITH_MYCOBACTERIUM_TUBERCULOSIS
REACTOME_REGULATION_OF_BETA_CELL_DEVELOPMENT
PID_HNF3B_PATHWAY
ST_STAT3_PATHWAY
REACTOME_GABA_SYNTHESIS_RELEASE_REUPTAKE_AND_DEGRADATION
REACTOME_NITRIC_OXIDE_STIMULATES_GUANYLATE_CYCLASE
B
Bayesian model for Diabetes mellitus: Pathways edge weight
REACTOME_SYNTHESIS_SECRETION_AND_DEACYLATION_OF_GHRELIN
REACTOME_REGULATION_OF_BETA_CELL_DEVELOPMENT
REACTOME_PI3K_CASCADE
REACTOME_INSULIN_SYNTHESIS_AND_PROCESSING
REACTOME_AMYLOIDS
PID_PTP1B_PATHWAY
PID_HNF3B_PATHWAY
PID_HNF3A_PATHWAY
NABA_SECRETED_FACTORS
KEGG_MTOR_SIGNALING_PATHWAY
BIOCARTA_HDAC_PATHWAY
0.02273
0.09474
0.02837
0.01064
0.00641
0.05
0.07207
0.008547
0.004808
0.01613
0.01923
jaccard
0.075
0.050
0.025
Pathways
0.0
0.3
0.6
0.9
Edge weight
C
Bayesian model for Alzheimer's: Pathways edge weight
REACTOME_RIG_I_MDA5_MEDIATED_INDUCTION_OF_IFN_ALPHA_BETA_PATHWAYS
REACTOME_RESPONSE_TO_ELEVATED_PLATELET_CYTOSOLIC_CA2
REACTOME_PLATELET_ACTIVATION_SIGNALING_AND_AGGREGATION
REACTOME_HEMOSTASIS
PID_MYC_REPRESS_PATHWAY
NABA_MATRISOME_ASSOCIATED
NABA_MATRISOME
NABA_ECM_REGULATORS
KEGG_PPAR_SIGNALING_PATHWAY
0.003745
0.01429
0.01003
0.006088
0.003891
0.006369
0.007414
0.00464
0
jaccard
0.010
0.005
0.000
Pathways
0.0
0.2
0.4
0.6
Edge weight
D
Bayesian model for Thalassemia: Pathways edge weight
REACTOME_TRANSMISSION_ACROSS_CHEMICAL_SYNAPSES
REACTOME_TRANSMEMBRANE_TRANSPORT_OF_SMALL_MOLECULES
REACTOME_TOLL_RECEPTOR_CASCADES
REACTOME_SIGNALING_BY_FGFR_IN_DISEASE
REACTOME_SIGNALING_BY_FGFR
REACTOME_SIGNALING_BY_ERBB2
REACTOME_NEUROTRANSMITTER_RECEPTOR_BINDING_AND_DOWNSTREAM_TRANSMISSION_IN_THE_POSTSYNAPTIC_CELL
REACTOME_NEURONAL_SYSTEM
REACTOME_FACTORS_INVOLVED_IN_MEGAKARYOCYTE_DEVELOPMENT_AND_PLATELET_PRODUCTION
REACTOME_CYTOSOLIC_TRNA_AMINOACYLATION
REACTOME_ACTIVATED_TLR4_SIGNALLING
KEGG_SYSTEMIC_LUPUS_ERYTHEMATOSUS
KEGG_PORPHYRIN_AND_CHLOROPHYLL_METABOLISM
KEGG_CALCIUM_SIGNALING_PATHWAY
BIOCARTA_AHSP_PATHWAY
0
0.004386
0
0
0
0
0
0
0.03509
0
0
0.005435
0.02381
0
0.09434
jaccard
0.075
0.050
0.025
0.000
Pathways
0.0
0.3
0.6
0.9
Edge weight

Figure 2.3: Probabilistic graph models have been employed to infer direct relationships between diseases with pathways by analyzing enrichment values in human pancreatic-beta cells. **A)** Bayesian network was constructed to infer the relationships between pathways and disease. Only the inferred dependencies with an edge value greater than 0.3 are displayed. The vertices located adjacent to the disease node (diabetes) can be seen in blue color. The Markov network connecting pathways and diabetes gene sets has also been displayed. Only the top 13 associated paths are shown in the Markov network to prevent perceptual clutter. The pathways and the associated enrichment scores were utilized for both Bayesian & Markov networks. **B)** For the purpose of identifying pathways that are directly linked to diabetes, weights on the edges in the Bayesian network are learned together with Jaccard index of overlap across the appropriate pathway and disease gene sets. **C)** The Bayesian network was also used to determine the weight of the edges in the pathways directly associated with Alzheimer's within cerebellum astrocytes. **D)** The edge weights represent the strength of the connections between directly linked pathways in the Bayesian network model for Thalassemia. These weights were determined using the single-cell expression profiles from erythrocyte progenitors in the bone marrow.

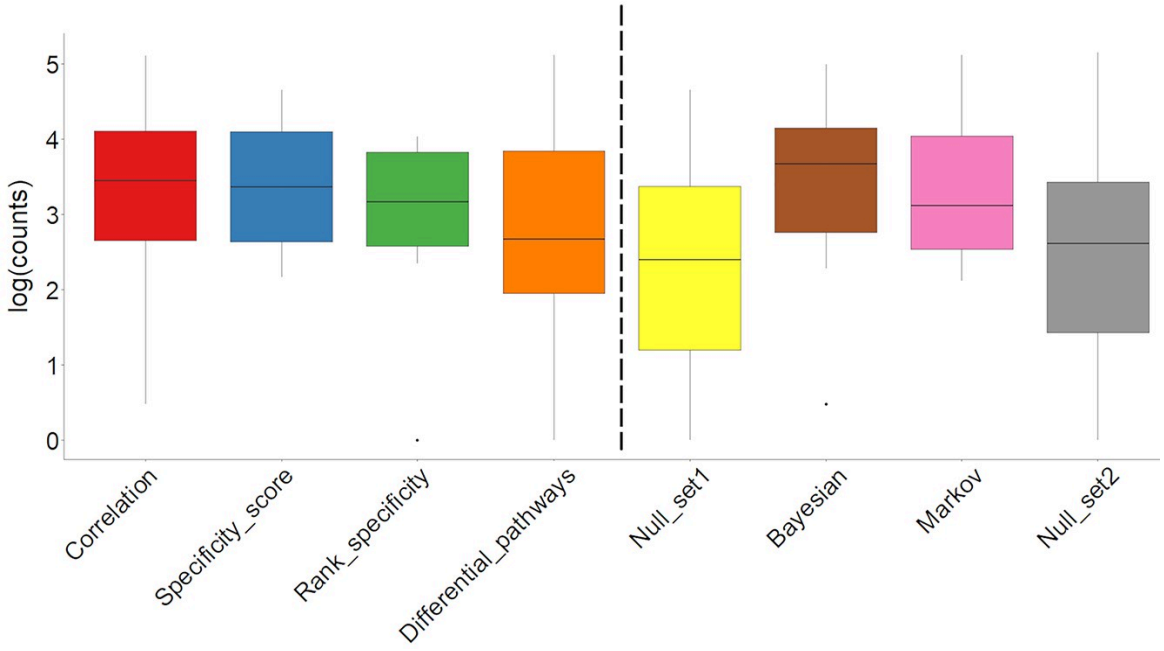

Figure 2.4: Comparative analysis of disease and pathway dependencies, estimated by several approaches, on the single-cell expression pattern of human pancreatic-beta cells. The techniques being evaluated in this analysis include Spearman correlation coefficient, specificity (calculated as 1 minus the p-value), rank-based specificity (calculated as 1 minus the p-value rank), Bayesian network, and Markov network. Please note that the PubMed based co-occurrence for diabetes is only displayed for the topmost pathways (top ten) that were identified using various methodologies. The PubMed co-occurrence of null model counts across random pathways is also shown. The count

is additionally displayed for pathways that exhibited greater activity within diabetic human pancreatic-beta cells compared to normal pancreatic-beta cells (labeled as “Differential_pathways”).

2.3.4 Variations across species

In order to assess variability in relationships between diseases and pathways across other species, we examined the linkages between pathways in the pancreatic-beta cells of both humans and mice. We examined pathways that matched the filtering requirements by using Markov and Bayesian models for symmetrical and asymmetric network associations, respectively. An intriguing finding from the data is that whereas several pathways have comparable connections with diabetes mellitus in the both humans and mice, a few of others seem to be connected in different manner (Figure 2.5A). For instance, pathways such as insulin generation and processing, HDAC route, and PTP1B pathway show comparable connections in both species, and several studies indicate that these pathways might be viable therapeutic targets for diabetes mellitus (Vieira et al. 2017; Dewanjee et al. 2021) . Several additional pathways, including amyloids, the HNF3B pathway, and ghrelin, showed differences in relevance between the two species. Cefalu *et al*. state that while there are many suitable mice models for the human diseases, however there is no one mouse model that encompasses all the symptoms of type 2 diabetes in human. The mouse models develop diabetes as a result of extreme obesity, but they do not exhibit the same islet pathology (islet amyloid) that occurs in humans with type 2 diabetes (Cefalu 2006). In conclusion, these findings provide evidence that in order to understand the range of characteristics seen in diabetes patients, it is necessary to use a range of animal models when studying the progression of the disease (Michael Conn 2013; Kottaisamy et al. 2021; Al-Awar et al. 2016) .

2.3.5 Variability across age

Utilizing single-cell profiles to investigate the association between diseases and pathways allows us to analyze individuals separately. The presence of genetic predisposition and environmental variables may contribute to heterogeneity in these relationships (Mishra et al. 2021; Jha et al. 2021). Age is a component that interacts with a person's genetic background and influences carbohydrate-associated metabolism (Chentli et al. 2015) . Therefore, in order to delve more into the advantages of using single-cell profiles, we endeavored to examine the impact of age on the

connection between diabetes-pathways in human pancreatic-beta cells. The study examined and compared diabetes-related pathways derived from co-occurrence along with graphical models in beta-pancreatic cells for two age groups: young adults (21-22 years) and older persons (38-54 years) (Figure 2.5B) (Enge et al. 2017).

Analysis of Bayesian networks in human pancreatic-beta cells from both young and elderly individuals identified many well-known pathways, including the insulin, PPARA, and HDAC pathways. In addition, several pathways in network were shown to have exclusive connections towards diabetes node. Specifically, HNF3B pathway was exclusively associated with old adult beta-pancreatic cells, whereas the HNF3A pathway was specifically related to young adult pancreatic-beta cells. In a research conducted through Abderrahmani *et al*., six distinct variations of HNF3B were examined and it was shown that HNF3B is unlikely to be a significant factor in the development of maturity-onset diabetes of the young (MODY) (Abderrahmani et al. 2000). The absence of HNF3B pathway node in the young adult pancreatic-beta network may be attributed to the fact that this form of diabetes often occurs before the age of thirty.

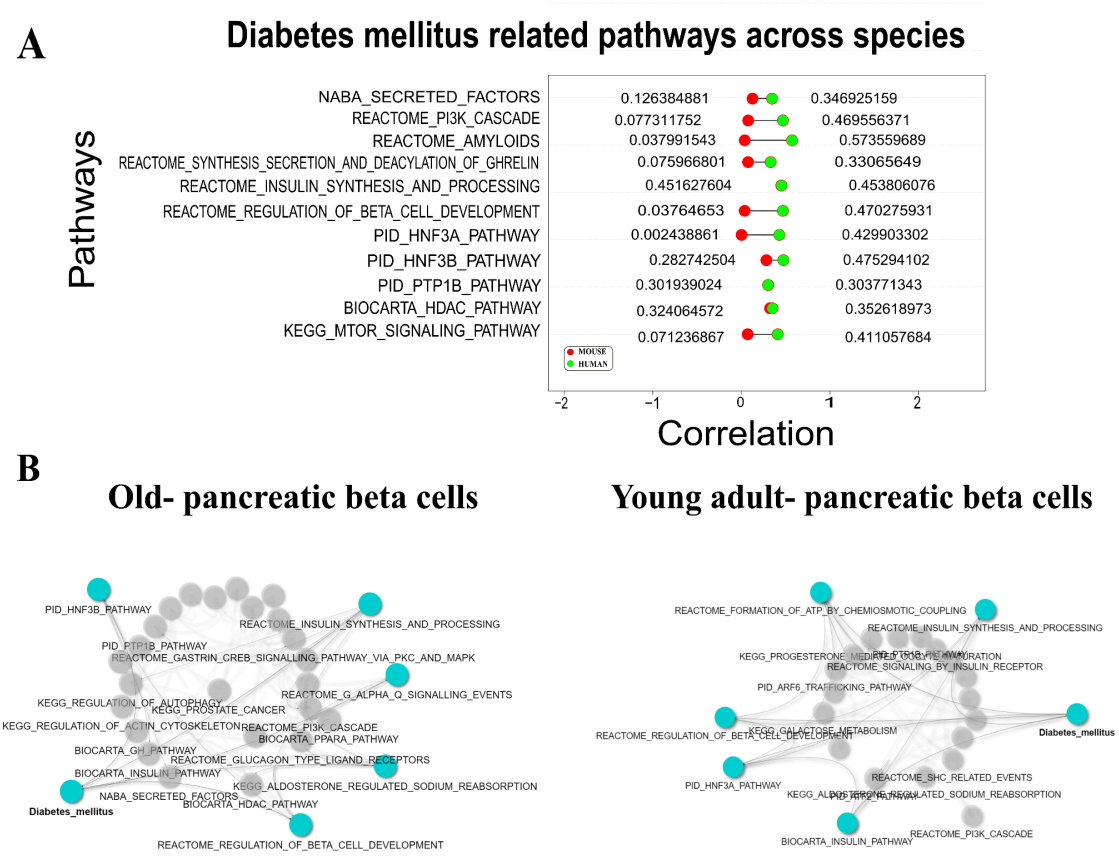

Figure 2.5: Analysis of hypothesized dependencies between diabetes disease and pathways reveals variations across different animals and age groups. **A)** The estimated dependencies between pathways and diabetes are displayed using correlation values. These values are based on the single-cell expression profiles via pancreatic beta cells in both mice (depicted using red dots) and humans (depicted using green dots). **B)** Constructed Bayesian networks by inferring relationships between disease and pathways gene sets based on the respective enrichment scores within human pancreatic-beta cells from young and old people. Only associations with a value greater than the threshold of 0.3 are shown. Following the application of thresholding, only the vertices (Pathways) that were immediate neighbors to the diabetes node have been highlighted in blue.

2.3.6 Opportunities for exploring the effects of drugs

Our approach avoided obscuring aberrations resulting from heterogeneity in bulk data by emphasizing closely related diseases and pathways using single-cell expression profiles. Validation by unbiased study of PubMed abstracts enhanced the reliability of sci-PDC prediction. In order to thoroughly investigate the potential and efficacy of our model, we conducted a search for viable drugs and relevant clinical trials with the help of Therapeutic Target Database (Zhou et al. 2022). We obtained several consistent results on the projected appropriateness of drugs (Figure 2.6). For example, one such finding from sci-PDC shows a significant correlation between the GABA production pathway with diabetic retinopathy, which is a kind of diabetic neuropathy condition. Moreover, the Jaccard index across diabetic retinopathy and GABA production pathway has been demonstrated to be zero. We have come upon a relevant drug named divalproex sodium, that has been shown to be effective by acting on GABA associated pathways. Sodium valproate, a component of divalproex sodium, has shown substantial improvements in the level of pain reported by individuals with Diabetic Neuropathy, which also includes Diabetic retinopathy (Kochar et al. 2004; Cohen et al. 2015). Through the use of identical methodologies, we have discovered a gene set responsible for the breakdown of the extracellular matrix (ECM) that is linked to both pancreatitis and pancreatic cancer. Prinomastat, a pharmaceutical compound, acts as an inhibitor of MMPs, which play a role in the breakdown of the extracellular matrix (ECM). Previous studies have indicated the therapeutic efficacy of prinomastat for treating pancreatic cancer (Zucker et al. 2000; Alves et al. 2001) . Our investigation produced a hypothesis suggesting that MMP inhibitors, such as prinomostat, may have potential benefits for pancreatitis.

Furthermore, the connection between Baclofen and neurotransmitter release in MS (multiple sclerosis) was shown using sci-PDC (Figure 2.6). Oral Baclofen is the first-line medication recommended by the Consortium of Multiple Sclerosis Centres (CMSC) for MS spasticity (Erwin et al. 2011; Haselkorn et al. 2005; Rekand & Grønning 2011). Baclofen inhibits the release of stimulating chemical messengers in the brain and promotes muscular relaxation by attaching to receptors on certain nerve cells (Thomas et al. 2010). Phentermine, a commonly used medicine for managing obesity, has been shown to impact glycemic indices. Our methodology has also brought attention to the correlation between phentermine and the glucose transport pathway (Cosentino et al. 2013) . Our investigation revealed an association between the use of the phentermine drug and the occurrence of disorders such as diarrhoea and Parkinson's disease. Phentermine consumption has been linked to the development of several disorders, as shown in recent studies (University of Illinois 2017; Anon n.d.). In addition, our research also revealed a drug-related connection between diarrhoea and the GPCR-signalling pathway. Diphenoxylate and loperamide are drugs that effectively cure diarrhoea by specifically binding to opioid receptors, which are a kind of Gut-residing antagonistic G-protein coupled receptors. (Mackerer et al. 1976; Mercer & Coop 2011; Dhawan et al. 1996). Therefore, the disease-pathway connections identified by sci-PDC seem to be valuable for identifying prospective drugs.

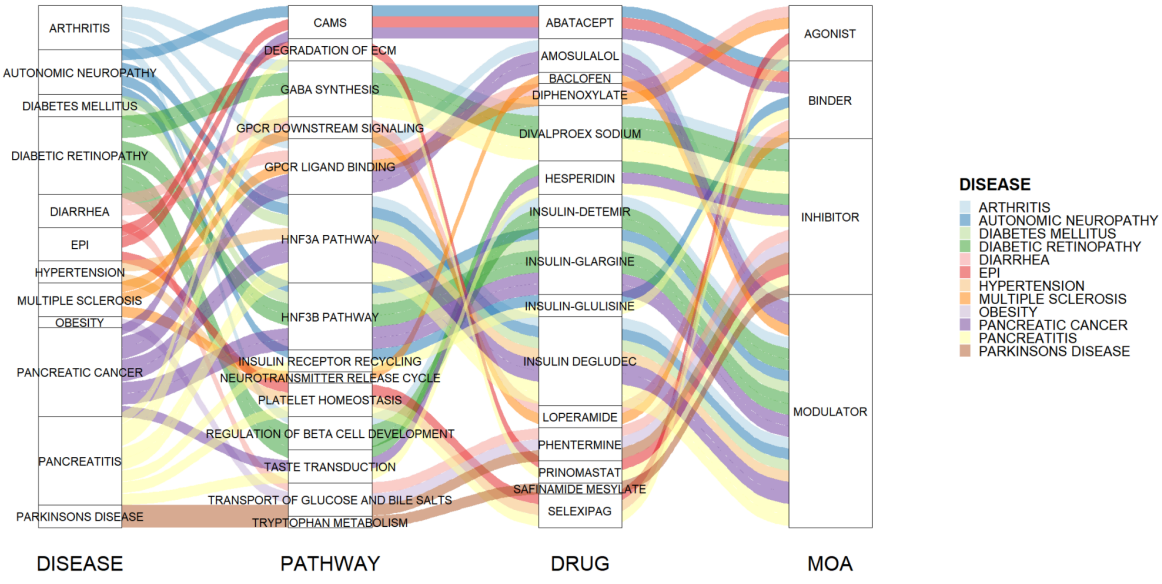

Figure 2.6: Demonstration of the top accurate anticipated connections among disorders and pathways, along with potential established drugs. The drug class is often referred to as an modulator or inhibitor, specifically in terms of its mode of action (MOA). The study examined the relationship between Arthritis, Diabetes mellitus, Autonomic neuropathy, Diabetic retinopathy, Hypertension, Pancreatic cancer, EPI (Exocrine pancreatic insufficiency), Parkinson's disease, Pancreatitis, Multiple sclerosis, Diarrhea, and Obesity with multiple pathways which satisfied the filtering criteria. Furthermore, an examination was conducted on drugs that specifically target interconnected pathways with the goal of enhancing comprehension of the relationships between diseases, pathways, and drugs.

While there are a few studies that provide evidence for these concepts, it would need several further validation experiments. Furthermore, the connections anticipated by sci-PDC may potentially suggest the tissue-specific adverse effects of certain medications that target pathways. For example, we anticipated the connection between GABA production and pancreatitis. Sodium valproate, a medication that affects the GABA system, has been identified as a cause of pancreatitis (Huang et al. 2019). Therefore, this technique may be used to create hypotheses for investigating the beneficial and detrimental impacts of drug-targeting pathways inside certain cell types and tissues. In summary, our approach using single-cell profiles may assist in narrowing down the search area for identifying alternative drugs that target specific cell types associated with a certain condition, for instance diabetes and pancreatic-beta cells. Additionally, it can aid in understanding the impact of these treatments on other characteristics within the context of different cell types.

2.4 Discussion

The issue of using single-cell expression profiles to infer direct associations between pathways and disease gene set enrichment is highlighted here. A more direct and non-traditional approach to solving this challenge has been shown. Utilising single-cell heterogeneity, we have deduced dependence networks between pathways-diseases. To clarify the significance and specificity of a connection across pathways, diseases, with cell types, we use extensive single-cell atlases to standardise the estimations of dependencies. In the past, many techniques have used large-scale gene expression or mutation profiles to establish connections between pathways and phenotypes (or diseases). Nevertheless, many previous analyses that utilised bulk sample gene expression profiles have been hindered by several issues. These include the presence of cellular heterogeneity in the samples, which impedes the identification of related cell types and obscures

the results of differential expression. Additionally, there is a lack of sufficient samples due to challenges in recruiting patients.

Furthermore, a gene set that has been enhanced by the differential-expression analysis does not always have a direct correlation with the disease being studied. Moreover, when using mutation profiles, the task of drawing such conclusions becomes considerably more complex. Therefore, our method offers a practical resolution to problems related to the specificity of cell types and the direct interconnections between pathways and diseases. sci-PDC has several resemblances to the estimate of gene networks using single-cell expression patterns. However, in this case, we have used gene set enrichment instead of gene expression, which has the benefit of being less influenced by drop-out or sparsity in single-cell expression profiles. To enhance the robustness of the sci-PDC approach, we have implemented various strategies (Sharma, Jha, et al. 2022). These include determining the enrichment score of a gene set by considering the minimum number of genes, filtering cells based on their quality, and assessing the significance of disease to cell-type. Additionally, we have normalised the data using a diverse and extensive collection of cell type profiles to emphasise specificity. Furthermore, we have shown the significance of our resulting connections via a high-throughput validation method based on a search of PubMed literature.

We have also demonstrated how differentiability in the relationship between pathogenicity and mouse and human cell types may be inferred using sci-PDC. Monitoring the fluctuations in the relationship between diseases and pathways is also useful for determining the accuracy of using mice models to examine the connections between specific pathways and diseases. For instance, our findings showed that a few pathways exhibit gene set dependence with disease in both humans and mice. To target these pathways for the treatment of diabetes, it may be possible to use a mouse model. However, in the case of other pathways (Amyloids, HNF3B, Gherlin) where the level of reliance varies across different species and the mouse model may not be suitable for assessing them as potential therapeutic targets for diabetes.

Multiple studies indicate that the mechanisms behind ageing and diabetes are comparable in terms of their pathophysiology. As individuals age, their capacity to regulate glucose levels may decline, perhaps influenced by factors such as physical fitness and body obesity. In addition, it is possible for individuals to acquire asymptomatic hyperglycemia as they become older. This condition leads to a slow and cumulative disruption of metabolic processes, which significantly

raises the likelihood of various chronic illnesses associated with ageing over an extended period of time (Chia et al. 2018).

One advantage of utilizing single-cell expression profiles is the ability to simulate heterogeneity in cells from an individual at a given moment. Therefore, it allows for the identification of differences among various people. Thus, using sci-PDC on a patient's cells has the potential to facilitate advancements in precision medicine. The data obtained from our study on both young and old pancreatic-beta cells consistently showed a pattern. In addition, it also brought to light differences in the connection between certain pathways and diabetes. Additionally, whereas sci-PDC may identify direct relationships between pathways and diseases, it does not always suggest causation. Nevertheless, pathways with a greater direct reliance are more likely to be causally related than random gene sets. Nevertheless, sci-PDC narrows the field of search for actual causation variables that may be examined in *in vivo* or *in vitro* models.

CHAPTER 3

USING TUMOR SPECIFIC SINGLE-CELL PROFILES TO EXPLORE RELATIONSHIPS BETWEEN BIOLOGICAL PATHWAYS AND CANCER HALLMARK PROPERTIES FOR PRECISION THERAPY

3.1 Introduction

3.1.1 Significance of cancer hallmark properties and their pathway association

While cancer manifests in a wide variety of ways, it seems to adhere to certain fundamental principles. The hallmarks of cancer, first suggested in 2000 by Hanahan and Weinberg (Hanahan & Weinberg 2000), delineate a collection of six fundamental properties that cancer cells acquire as they undergo their transition from normal cells. In the latest development of this paradigm, other emergent traits have been included in the key hallmarks of cancer. These traits include not only unrestricted growth but also the capacity to escape growth suppressors, fight cell death, and acquire unrestricted replicative potential (Hanahan 2022). Each hallmark signifies a basic capacity that cancer cells develop for growth and propagation, often via distinct genetic and molecular modifications. Gaining insight into the origins of these hallmark properties is crucial for developing effective cancer treatments.

Pathways are intricate signaling networks that are responsible for the coordination of biological processes. Alterations or modifications in these pathways may result in an early development of the hallmarks. Researchers may enhance their comprehension of the fundamental processes of

cancer formation by pinpointing the distinct pathways linked to each hallmark in various forms of cancer (Iorio et al. 2018). Subsequently, this information can be used to facilitate the development of targeted therapies. For example, certain drugs operate by blocking particular proteins within these pathways, thereby interrupting the hallmark they influence and preventing the progression of cancer (Vogelstein & Kinzler 2004). Thus, examining the relationship between cancer hallmarks and their associated pathways is a fundamental aspect of contemporary cancer research, providing a potent framework for understanding the disease and creating more potent therapies.

3.1.2 Cell line specific nature for hallmark properties and its role in therapy

According to Hanahan and Weinberg (2011), cancer hallmark qualities do, in fact, show cell type-specific traits that represent the various biological behaviors of distinct tumor types (Hanahan & Weinberg 2011). The expression of each hallmark feature might vary based on the specific cell type from which it originates and the distinct microenvironment in which it exists. For instance, the tissue of origin influences the angiogenesis process, as different cancer types secrete distinct angiogenic factors and interact with the adjacent stromal cells in a unique manner (Ferrara 2002). Furthermore, the immune evasion strategies that tumors employ are frequently customized to the immune topography of their particular tissue environment, as evidenced by the dissimilar patterns observed in malignancies such as melanoma and lung cancer (Pardoll 2012). The progress of high-throughput technology has made it possible for scRNA-seq to revolutionize our understanding of cellular diversity. These developments enable researchers to analyze enormous quantities of cells simultaneously, thereby capturing the complete transcriptome of each cell with greater precision and less technical noise. Detailed information such as this offers an unparalleled perspective on the unique characteristics of cell types within intricate tissues (Choi & Kim 2019).

Although scRNA-seq, is an excellent method for determining current levels of cellular heterogeneity and intricate pathway connections, its fundamental limitation is that it can only offer a static image of the cellular states at a particular moment in time (Tanay & Regev 2017; Yuan et al. 2017). This constraint is intriguing when examining the initial growth and advancement of cancer hallmarks, which often include transient and arbitrary events. The development from normal to pre-cancerous to malignant states requires interactions with the

microenvironment and other cells, which scRNA-seq cannot completely understand without combining with other omics technologies (Wagner et al. 2016; Chen et al. 2023). To identify early stages of disease growth, it is imperative to develop methods that can identify cells that are engaging in subtle but critical transitions toward cancer states prior to the onset of clinical symptoms. This enables targeted therapies that are customized to individual genetic and molecular characteristics, possibly enhancing therapy efficacy and patient prognosis.

This work addressed the task of predicting direct dependencies between biological pathways and cancer hallmark features by combining their single-cell expression and RNA-velocity patterns. RNA velocity effectively overcomes the drawbacks of conventional scRNA-seq by offering a dynamic perspective on cellular activities, which is determined by the current ratio of unspliced and spliced mRNA transcripts (La Manno et al. 2018). This method has been extensively utilized to investigate cellular differentiation, tracing lineage, and dynamic events such as tissue growth and disease progression (Gao et al. 2022). The temporal characteristics of gene regulation are revealed by RNA velocity, which provide information on the direction and speed of cellular alterations. However, it is subject to certain constraints, such as the difficulty of accurately quantifying unspliced transcripts, which can induce biases, and its reliance on the type of scRNA-seq method (e.g., full-length vs. 3' biased). These factors can affect the accuracy of velocity estimates and their interpretation across various experimental designs (Bergen et al. 2021).

Furthermore, research has shown that RNA velocity measurements recover a large portion of the information lost by pseudotime-ordered single-cell data, which is present in the real temporal couplings (Qiu et al. 2020). Such real-world functionalities of RNA velocity have facilitated its application in diverse research studies aimed at investigating dynamic cellular processes and developmental trajectories. This has significantly contributed to the comprehension of cell differentiation, tissue development, and disease progression (Bergen et al. 2020; Dorrity et al. 2020; Riba et al. 2022; Adewale et al. 2024). Here, we use an atlas of malignant single-cell profiles to reveal direct relationships between cancer hallmark features and pathways in context of different cell types. Comparing two breast cancer cell lines, we examine the relationships that have been inferred between cancer hallmark features and pathways. Furthermore, our technique has shown great promise in predicting the drug response across multiple cancer cell lines.

3.2 Material And Methods

3.2.1 Associating cancer hallmark properties with pathways in cell line specific context

The biological traits of the tumor, which are in turn influenced by a number of signaling pathways, define the hallmark features of cancer. In order to further examine these connections in cancer, we obtained single-cell transcriptome datasets of several cancer cell lines from the CancerSEA database (http://biocc.hrbmu.edu.cn/CancerSEA/). CancerSEA is a database of single-cell expression profiles of different cancer cell lines containing 93,475 individual cancer cells, from 27 different types of human cancers (Yuan et al. 2019). To comprehend the fundamental mechanistic workflow behind tumors, we examined 21 hallmark properties of cancer and their association with 3051 canonical pathways acquired from MSigDB using our methodology. The cancer hallmark geneset was acquired from Liang *et al*., (2020) and Zhang *et al*., (2020) (Liang et al. 2020; Zhang et al. 2020). The R package UniPath was utilized to convert cancer single-cell expression data into gene set enrichment scores (activity profiles) (Chawla et al. 2021). Hallmark and pathway activities of human single-cell expression profiles from the human cell landscape (HCL) were also computed to improve accuracy throughout the normalization process (Han et al. 2020). Human single-cell transcriptome profiles were collected from HCL (human cell landscape) (http://bis.zju.edu.cn/HCL/), which includes more than 700,000 single cells showing 1462 distinct cell types from over fifty different human tissues. The subsequent step was data processing and filtering to remove noise and irrelevant associations once the scores for each cell type of the corresponding tissue were determined.

Enrichment scores normalization and selection criteria for the co-occurrence of cancer hallmark-pathway associations

A) Cell line specific filtration: Inclusion of the gene set for a given cell type was contingent upon the presence of a minimum of ten pathways and a minimum of 10% of the cells in that category of cell type with a score lower than 0.25 for both hallmark and pathway. This filtration ensures the inclusion of just those cell types that exhibit enough coverage of both hallmark and pathway scores. The threshold of 10 pathways and 10% of cells with low scores is used to concentrate on cell types that have sufficient data variability to

evaluate the relevance of pathway and hallmark associations. In addition, this reduces the inclusion of cell types with sparse or insufficient data.

B) Cooccurrence-based filtration: Using filtered enrichment scores, cooccurrence analysis was used to determine significant associations between hallmark properties and pathway. Three criteria were used to evaluate the co-occurrence of hallmark and pathway: correlation, specificity score, and rank specificity.
    a) Correlation: The Spearman correlation was calculated using gene set enrichment values after log transformation.
    b) Specificity score: The specificity score was determined by conducting a permutation test based on a null model to evaluate the statistical significance of the difference in the frequency of hallmark and pathway.
    c) Rank specificity: This was achieved by normalizing the hallmark and pathway specificity score in the context of numerous cell types on the basis of their respective rankings in human cell landscape activity profiles.

    Furthermore, connections between hallmarks and pathways that had a rank specificity of less than 0.25, a specificity score of less than 0.05, and an absolute correlation of 0.3 were chosen for further analysis to investigate their relationships. By eliminating weaker or irrelevant connections, these criteria aid in ensuring that only strong and statistically significant hallmark-pathway associations are taken into account for further investigation. For example, whereas correlation assesses the strength and direction of a relationship, a lower specificity score (e.g., < 0.05) suggests that the association is likely significant rather than coincidental. Likewise, rank specificity aids in determining whether the identified hallmark-pathway associations are consistently exclusive to distinct cell types or if they are prevalent across different types of cells. This ensures that the associations are not generic and are pertinent to the unique biological environment of the cell type under consideration.

C) Sample set-based filtration: To reduce overlap between any pair of gene sets, we additionally calculated overlap coefficient and Jaccard index. This step minimizes redundancy and ensures that the studied pathway sets are different, hence enhancing the clarity and specificity of the co-occurrence analysis. This also mitigates the overrepresentation of specific pathways in the analysis.

D) PGM filtration: In both Bayesian and Markov networks, hallmark-pathway associations greater than 50 percentile were considered relevant associations. This step makes sure that only the hallmark-pathway relationships that are most closely related are taken into account. It directs the analysis towards connections that are more probable to be biologically significant and of relevance for subsequent investigation.

3.2.2 RNA-velocity based regulation of hallmark properties of cancer cells

RNA velocity may be used to deduce the direction of modifications in gene expression in scRNA-seq data. By leveraging the abundance of unspliced to spliced RNA transcripts, it offers valuable information regarding the future condition of particular cells (La Manno et al. 2018; Bergen et al. 2021). The transcriptional dynamics model incorporates the rates of transcription (α), splicing (β), and degradation (γ), which result in a compilation of spliced (s) and unspliced (u) mRNA molecules. By using this transcription model, we can formulate the rate equations associated with a single gene. These equations depict the temporal evolution of the predicted quantities of spliced molecules s and unspliced mRNA molecules u :

$$\frac{du}{dt} = \alpha(t) - \beta(t)u(t) \quad (1)$$

$$\frac{ds}{dt} = \beta(t)\,u(t) - \gamma(t)\,s(t) \quad (2)$$

Here, $\alpha(t)$ represents the time dependent rate of transcription, $\beta(t)$ indicates the rate of the splicing, and $\gamma(t)$ reflects the rate of degradation (La Manno et al. 2018; Bergen et al. 2021).

The initial data from scRNA-seq experiments is obtained in the form of raw sequencing data, namely FASTQ files. These files include sequencing reads that are derived from individual cells. To acquire gene expression profiles for each cell, the raw data must undergo processing, alignment to a reference genome, and quantification. Prior to commencing alignment, a quality control assessment was conducted on the raw files using fastp (Chen et al. 2018). Afterward, we used STAR (Spliced Transcripts Alignment to a Reference) to align our reads to the human reference genome (Dobin et al. 2013). This allowed us to identify the specific locations on the human genome from where our reads came. The BAM files, acquired via STAR alignment, were used as input for the velocyto tool to produce spliced and unspliced counts. Velocyto is a package designed to analyze the patterns of gene expression changes in the single-cell RNA sequencing data (La Manno et al. 2018). The velocyto package produced loom files, which were

then used by scVelo to calculate RNA velocity using spliced and unspliced counts (Bergen et al. 2020). Basic pre-processing of the data, including filtering genes and cells and normalizing the counts, was done before computing the moments for velocity estimate. After the initial preprocessing, the gene velocity for each cancer sample was computed using a logarithmically transformed count matrix. The gene velocity matrices were standardized by transforming them into z-scores (standardized scores). The standardized z-score gene velocity matrices were used to compute the combined geneset velocity for each hallmark property and pathway.

To simplify, geneset velocity or combined geneset velocity measures how rapidly genes are activated or repressed by analyzing the rate of change in gene expression within a particular gene ensemble under various conditions. A similar idea is hallmark velocity (or combined hallmark velocity), which measures the dynamics of gene expression within hallmark gene ensemble. Similarly, the enrichment score of genes within a specific biological pathway is referred to as pathway activity, which indicates the extent to which a pathway participates in a given condition. In order to quantify the role of cancer hallmark properties in biological processes, hallmark activity focuses on the degree of activation within cancer hallmark gene ensembles. Finally, geneset activity is a more general term that denotes the activation level of genes within any defined gene ensemble, whether it is a specific pathway, hallmark set, or other group of genes. Although all of these terms quantify some aspect of gene activity or change, their scope and context vary depending on whether they are applied to specific gene sets, hallmark properties, or broader biological processes.

The combined velocity was adjusted using a permutation-based test with a null model to account for the stochasticity and noise in each geneset's velocity. To create a null model, we first selected cells at random from the human cell landscape database (Han et al. 2020). The same procedure was used on the unprocessed fastq data, resulting in the acquisition of a matrix containing the combined geneset velocity for various human cells. To determine the adjusted velocity in a target cell for a hallmark or pathway (gene ensemble), we consider the fraction of cells in the null model that had a lower combined velocity than the corresponding target cell.

3.2.3 Direct association inference using a probabilistic graph model

We used a data-driven probabilistic graph model for association analysis as the data sets can include multifaceted dependencies, including confounding variables and inter-causal dependence (Mourad et al. 2012). The model features comprised of hallmark velocity, hallmark activity, and associated pathways that were derived from co-occurrence analysis. We utilized Hill Climbing (HC) as the structure learning technique from bnlearn to learn the structure of the Bayesian network (BN) across features (Scutari 2009). The Bayesian Information Criterion (BIC) was utilized as the score function. We utilized the XMRF package for estimation of the graph connecting the features for Markov network based inference (Wan et al. 2016). The technique used was 'LPGM', with additional parameters such as stability="StARS" and beta=0.3. Next, in order to enhance the resilience of the graphical model, we used bootstrapping to train and utilize 1001 networks for majority voting. To minimize the computing time required for network learning, the data was discretized using quantiles. Selecting hallmark pathway connections that are shared by Markov and Bayesian networks and were above the 50 percentile in edge weight was done for further investigation. In order to evaluate the importance of these relationships, we also examined their occurrence in GWAS(Genome-wide association studies) (Sollis et al. 2023) based geneset enrichment using Enrichr (Chen et al. 2013).

3.2.4 Literature based high throughput validation

Using PubMed abstract-based validation, we tried to determine the reliability of the resultant hallmark pathway associations and examine the performance of various applied strategies (e.g., correlation, Bayesian, Markov). The input for this case was the hallmark term and its associated pathways. The Bio.Entrez package has been utilized to get data from the PubMed database, namely abstracts of the research papers spanning the years 1980 to 2024 (Chang et al. 2010). Programmatically searching PubMed with specified phrases for several articles at once is possible with the Bio.Entrez package. This approach automates large-scale literature searches and data extraction in batches. The same configurations were used to search for the co-occurrence of random pathway words for the same hallmark as a control. To analyze the performance of various approaches, we selected the top 10 pathways from each of the method and contrasted the frequency of their abstract occurrences with regard to null or random pathways using a box plot.

3.2.5 Predicting drug response through context-specific associations

We endeavored to model the activity of various cancer cell lines against the area under the curve (AUC) value of drugs in order to determine the context-specific significance of hallmarks and associated pathways. In pharmacokinetics, the AUC quantifies the overall drug exposure over a certain period of time. The breast cancer cell lines provided by Gambardella *et al*.. (Gambardella et al. 2022) were used to model 544 drug responses from the CTRP2 version-2 database (Basu et al. 2013). Twenty-five breast cancer cell lines were selected based on the presence of drug response data in the database. We applied the kNN regression model to predict the AUC value of a drug. At first, we chose union pathways that are linked to the hallmark of "resisting cell death" as features for the machine learning model to predict the drug's AUC value. Furthermore, we used the machine learning model to calculate the AUC by using both the whole pathway gene sets and randomly generated pathway sets.

3.3 Results

We used an integrated strategy to quantify the interdependence among pathways and cancer hallmark properties through analysing their activity and velocity. Figure 3.1 illustrates a summarised depiction of the executed workflow. Using UniPath, we computed the activity of pathway and hallmark gene sets for each cell based on their cancerous single-cell expression profiles. In addition, we computed the RNA velocity of each cell to include the transient aspect of stochastic events in single-cell expression profiles. The relationships between hallmark properties and pathways were determined by a co-occurrence analysis of activity profiles. After additional investigation, we observed that filtering techniques and approaches based on a Probabilistic Graphical Model (PGM) would be more useful for determining direct relationships. Here, we use two breast cancer cell lines as an example to illustrate our investigation and findings.

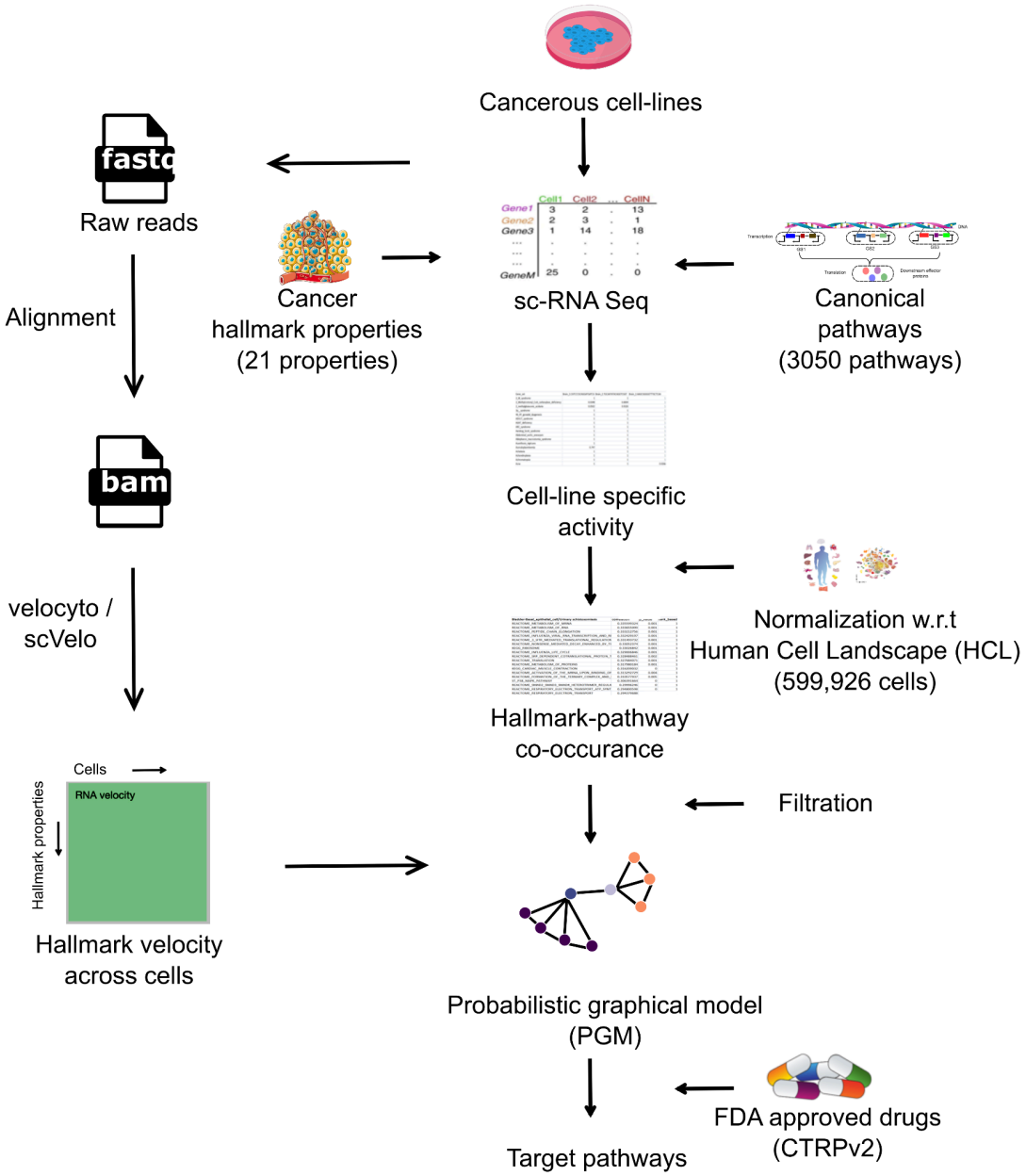

Figure 3.1: A brief workflow of the methodology. At first, we transformed single-cell expression profiles into gene sets into their respective enrichment scores (activity) with gene ensembles in rows and cancerous samples into columns. In addition to activity, we also calculated their combined RNA velocity. The integrated collection of gene set activity and RNA velocity was used to train a probabilistic graph model for obtaining hallmark pathway associations.

3.3.1 Establishing pathway and cancer hallmark specificity across cancer cell lines

We found numerous intriguing associations between cancer hallmark properties and pathways based on our initial examination of single-cell expression profiles of various cancerous cell lines from CancerSEA database. In addition to investigating the diabetes mellitus context specific to cell types in relation to biological pathways earlier in our prior method (sci-PDC), we also examined the impact of other confounding variables, such as age and species, on disease prognosis (Sharma, Jha, et al. 2022). As we proceed, we want to investigate the drug's mode of action with regard to cell type specificity. To illustrate our approach, we utilised the popular, FDA-approved diabetes drug Metformin as a case study. Recent potential applications of metformin include antiaging, neuroprotective, and anticancer effects that are being studied (Hua et al. 2023; Novelle et al. 2016; Sportelli et al. 2020). Thus, in this study, we used our methodology to investigate the connections between the hallmarks of different cancer cell lines and the Metformin-targeted pathways. For cell line-based association inference, hallmark-pathway relationships with a rank specificity of less than 0.25, a specificity score of less than 0.05, and an absolute correlation of more than 0.3 were taken into consideration. Co-occurrence analysis of metformin targeted pathways and the hallmark property (evading cell death) exhibited context specificity across various cancer cell lines (Figure 3.2A). For example, in the breast cancer cell line (MDA-MB-231) the cancer hallmark property "evading cell death" was shown to be associated with the pathway "pyruvate metabolism." One of the main regulators of glycolysis, pyruvate kinase, converts phosphoenolpyruvate to pyruvate. Research has shown that metformin causes breast cancer cells to die by downregulating PKM2 (pyruvate kinase isoenzyme type M2) and triggering apoptotic pathways (Silvestri et al. 2015).

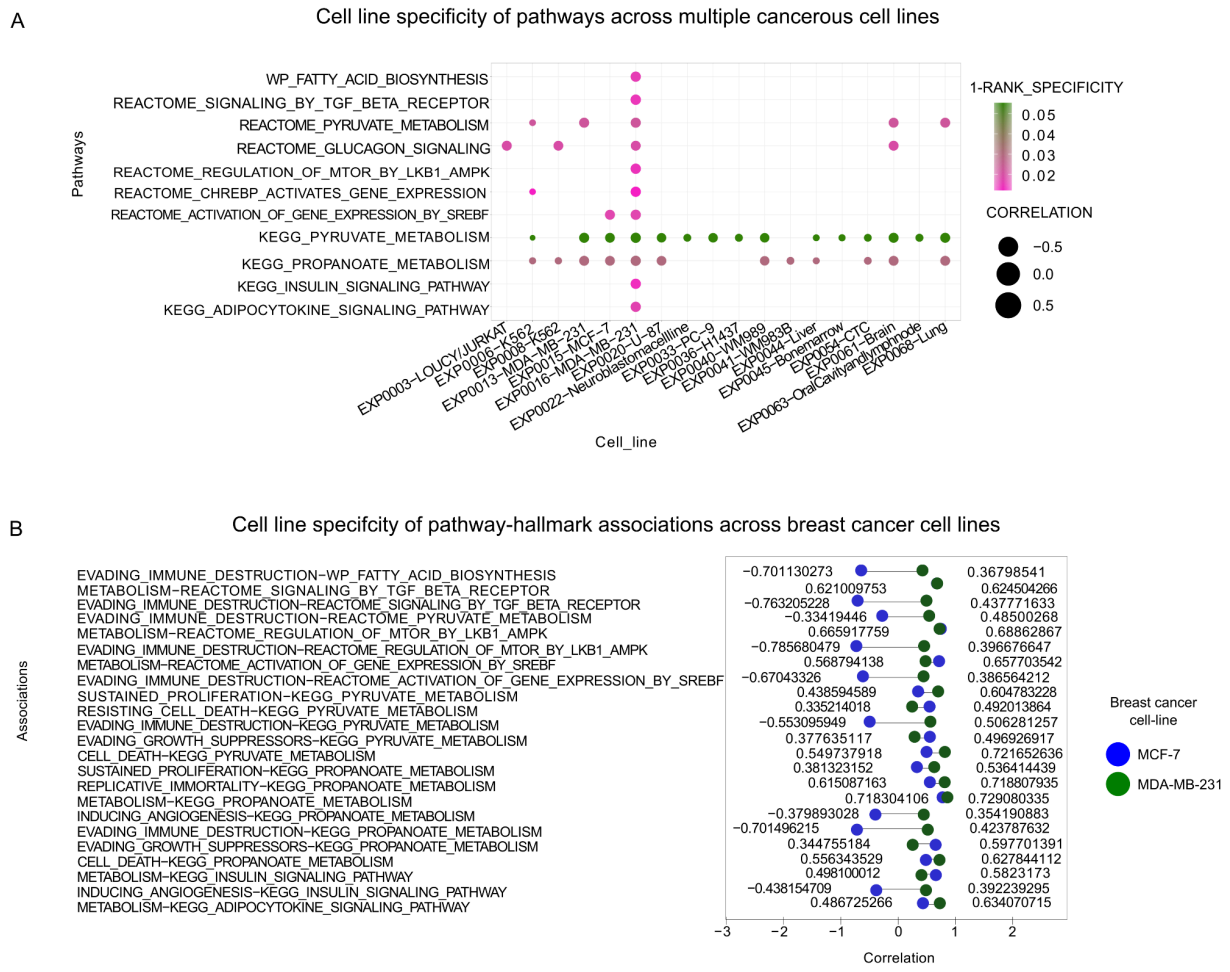

Figure 3.2: Variability within the cell line context in the anticipated associations between a cancer hallmark property and the pathways that a drug (Metformin) targets. A) Assessing the correlation between hallmark feature (Resisting cell death) of cancer cell lines and particular pathways. The study demonstrated the connections between pathways targeted by metformin and the hallmark propertie (Resisting cell death) of cancer for several types of cancer cell lines. B) Inferred connections of hallmark properties and pathways between the two most commonly studied cell lines of breast cancer. The dependencies were determined by calculating the correlation values between metformin targeted pathways and hallmark features using the single-cell expression profiles of MCF-7 (shown with blue dots) and MDA-MB-231 (shown with green dots).

In order to further establish the context-specific associations between these cancer hallmark properties and Metformin-targeted pathways, we performed an analysis of two commonly employed breast cancer cell lines, MCF-7 & MDA-MB 231. The MDA-MB-231 cell line is utilized as a representation of triple negative breast cancer (TNBC) due to its lack of expression of human epidermal growth factor receptor-2 (HER-2) or estrogen receptor (ER). Conversely, the

MCF-7 cell line is employed as a model for ER+ breast cancers because it expresses ER, rendering these cells responsive to estrogen (Shi et al. 2017). In addition to distinct morphological characteristics, numerous studies have reported distinctive differences between these two cell lines (Gest et al. 2013; Mladkova et al. 2010; Kwiatkowska et al. 2016). We looked at their correlation and rank specificity score as acquired after normalisation and filtering in order to further analyse the cell line specificity of cancer hallmark and pathway associations. Figure 3.2B illustrates how pathways such as TGF beta receptor signalling and AMPK-regulated MTOR were found to be similarly associated with hallmark property metabolism in both the cell lines. Ariaans *et al*., have shown that metformin can induce a metabolic transition to increased glycolysis by inhibiting mTOR, resulting in an earlier onset of glucose deprivation and more cell mortality in MCF7 and MDA-MB231 breast cancer cells (Ariaans et al. 2017). Though certain pathways act similarly in the two cell lines, some exhibit different behaviour. In one case, distinct behaviours were seen against the role of fatty acid synthesis in hallmark evading immune destruction. The pathway fatty acid synthesis was found to be positively correlated with the MDA-MB-231 (TNBC) cell line while negatively correlated with MCF-7 (ER+). Receptor-positive MCF-7 cells do not exhibit the same predilection for lipogenesis as MDA-MB-231 cells, according to a research by Balaban *et al*.(Balaban et al. 2018). These associations emphasise the significance of cell specificity in the diagnosis and treatment of cancer.

3.3.2 Coherence between activity and regulation based on RNA velocity

scRNA-seq details a static profile of cellular activity at a specific moment in time. RNA velocity can overcome this constraint of scRNA-seq by offering a dynamic perspective of cellular processes, which is determined by the relative abundance of spliced and unspliced mRNA transcripts. The incorporation of these two characteristics can offer us a comprehensive methodology that can be utilized to monitor the initial progression of cancer-defining characteristics or detect the poisness of cellular states. To lay the foundation for integrated geneset velocity and activity-based hallmark pathway associations, we examined the coherence between these two sets, i.e., hallmark velocity: pathway activity and hallmark activity: pathway activity. Figure 3.3 shows the results of a Spearman correlation analysis between the two sets of breast cancer cell lines MCF-7 & MDA-MB-231, which were utilized to investigate the

coherence. We categorize associations between different scenarios into five distinct categories: NN (Negative hallmark velocity correlation: Negative pathway activity correlation), PP (Positive hallmark velocity correlation: Positive pathway activity correlation), PN (Positive hallmark velocity correlation: Negative pathway activity correlation), NP (Negative hallmark velocity correlation: Positive pathway activity correlation), and FALSE (Zero hallmark velocity correlation: Zero pathway activity correlation). The filtered associations refer to the associations that meet the criteria of the co-occurrence filter, while the unfiltered connections represent all the associations related to hallmark pathways. Figure 3.3 clearly demonstrates that the filtration process effectively eliminated numerous insignificant false associations. Furthermore, it was noted that the MDA-MB-231 cell line predominantly exhibits filtered hallmark pathway associations in the PP category. This indicates that cancer hallmark properties tend to accumulate through elevated pathway activities. The presence of these relationships could potentially account for the extremely metastatic characteristics of MDA-MB-231 (TNBC) (Conner et al. 2024). In contrast, the NP category was found to be more common in MCF-7. Observations of this kind could potentially shed light on the function of intermediate cofactors and the indirect connections that exist between hallmark properties and pathways.

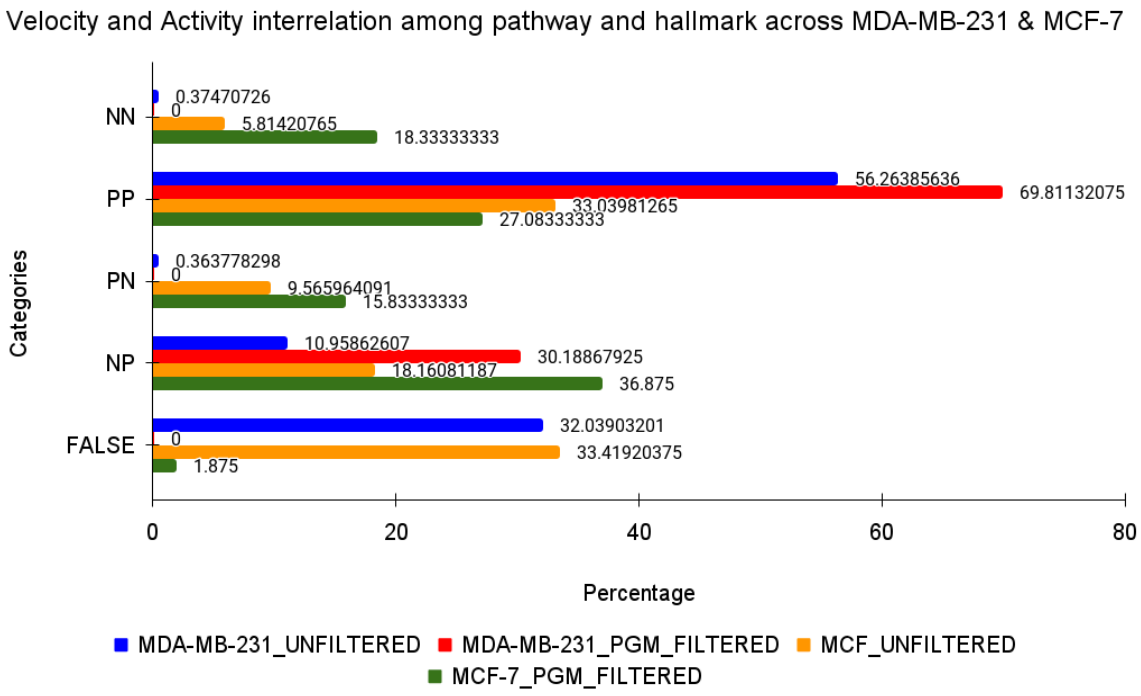

Figure 3.3: Here, the relationship between gene set activity and its regulation has been established by RNA velocity. The coherence was assessed using Spearman correlation analysis between the two sets of breast cancer cell lines MCF-7 & MDA-MB-231. The associations between different scenarios are classified into five distinct categories: NN (Negative hallmark velocity correlation: Negative pathway activity correlation), PP (Positive hallmark velocity correlation: Positive pathway activity correlation), PN (Positive hallmark velocity correlation: Negative pathway activity correlation), NP (Negative hallmark velocity correlation: Positive pathway activity correlation), and FALSE (Zero hallmark velocity correlation: Zero pathway activity correlation).

3.3.3 Inference and validation of context specific associations

Though many pertinent and established relationships between cancer hallmark properties and pathways were highlighted by significance and specificity calculations, these associations might only be linear and could also be indirect. To investigate the nonlinear and direct relationship between hallmarks and pathways, we employed probabilistic graphical modeling (PGM) using Bayesian and Markov methods. Asymmetrical and directional relationships within the network are highlighted by Bayesian network inference, whereas symmetrical and undirected relationships are better represented by the Markov model. We used Bayesian and Markov approaches to infer such dependencies by modeling the activity of specific hallmark pathway associations (obtained from co-occurrence filtration) and hallmark combined velocity (obtained from RNA velocity). Only the Hallmark pathway connections with edge weights above the threshold were chosen for validation. Figure 3.4A illustrates the ten pathways that were identified as common between the pathways filtered from PGM and GWAS-based geneset enrichment for MDA-MB-231. The highest correlation among the hallmark pathway associations for MDA-MB-231 was observed between the hallmark activity of "evading cell death" and the pathway activity of "copper homeostasis". We assessed this relationship across a number of breast cancer cell lines in order to further substantiate the association's cell line specificity (Figure 3.4B). Out of the 25 breast cell lines examined, 23 exhibited a positive correlation. The highest correlation observed was only 0.5, in contrast to the correlation of 0.7 observed in MDA-MB-231. Additionally, MDA-MB-231's rank specificity was significantly lower (0.01) than that of other breast cancer cell lines, demonstrating the effectiveness of our approach in identifying distinct hallmark-pathway associations that are specific to individual cell lines. Copper has been shown to directly influence cancer cells' ability to invade tissue and metastasize (Erler et al. 2009; Morrissey et al. 2016; MacDonald et al. 2014; Blockhuys &

Wittung-Stafshede 2017). The most recent study identified cuproptosis, a hitherto unidentified mechanism of regulating cell death. Cuproptosis is a cellular process that primarily occurs in cells engaged in active respiration and the TCA cycle. It facilitates the binding of copper with fatty acylating components, leading to the accumulation of fatty acylating protein, depletion of iron-containing sulphur cluster protein, activation of HSP70, initiation of harmful oxidative stress within the cell, and ultimately, cell death (Tsvetkov et al. 2022; Sha et al. 2022). Several studies have proposed the utilisation of copper complexes as a therapeutic approach to combat drug resistance in MDA-MB-231 (Abu-Serie & Abdelfattah 2023; Said Suliman et al. 2021; Wang et al. 2024). MDA-MB-231 had the highest expression of the Cu pump ATP7A and was the only cell line to express LOXL2 (an enzyme involved in extracellular matrix remodeling) at appreciable levels, indicating greater copper availability than other breast cancer cell lines (Vitaliti et al. 2023).

In addition, to verify additional associations of MDA-MB-231 that were identified through our analysis, we conducted a literature search for the co-occurrence of cancer hallmarks and associated pathway terms. Literature-based confirmation was carried out by looking up pathway terms that were found to co-occur for a particular hallmark in PubMed abstracts. A comparison was also conducted between various approaches (correlation, Bayesian, Markov) used for velocity and activity in a set of hallmark-related pathways. We discovered that the pathways derived from each of the methods outperformed the null distribution, indicating the significance of expected hallmark pathway connections. Figure 3.4C shows that, when compared to other methods for literature authentication, the relationships obtained using PGM of filtered pathways between hallmark activity-pathway activity for MDA-MB-231 performed the best. According to these observations, direct dependencies between these connections can be captured by PGM-based methods, whereas hallmark velocity-pathway activity appears to have indirect relationships.

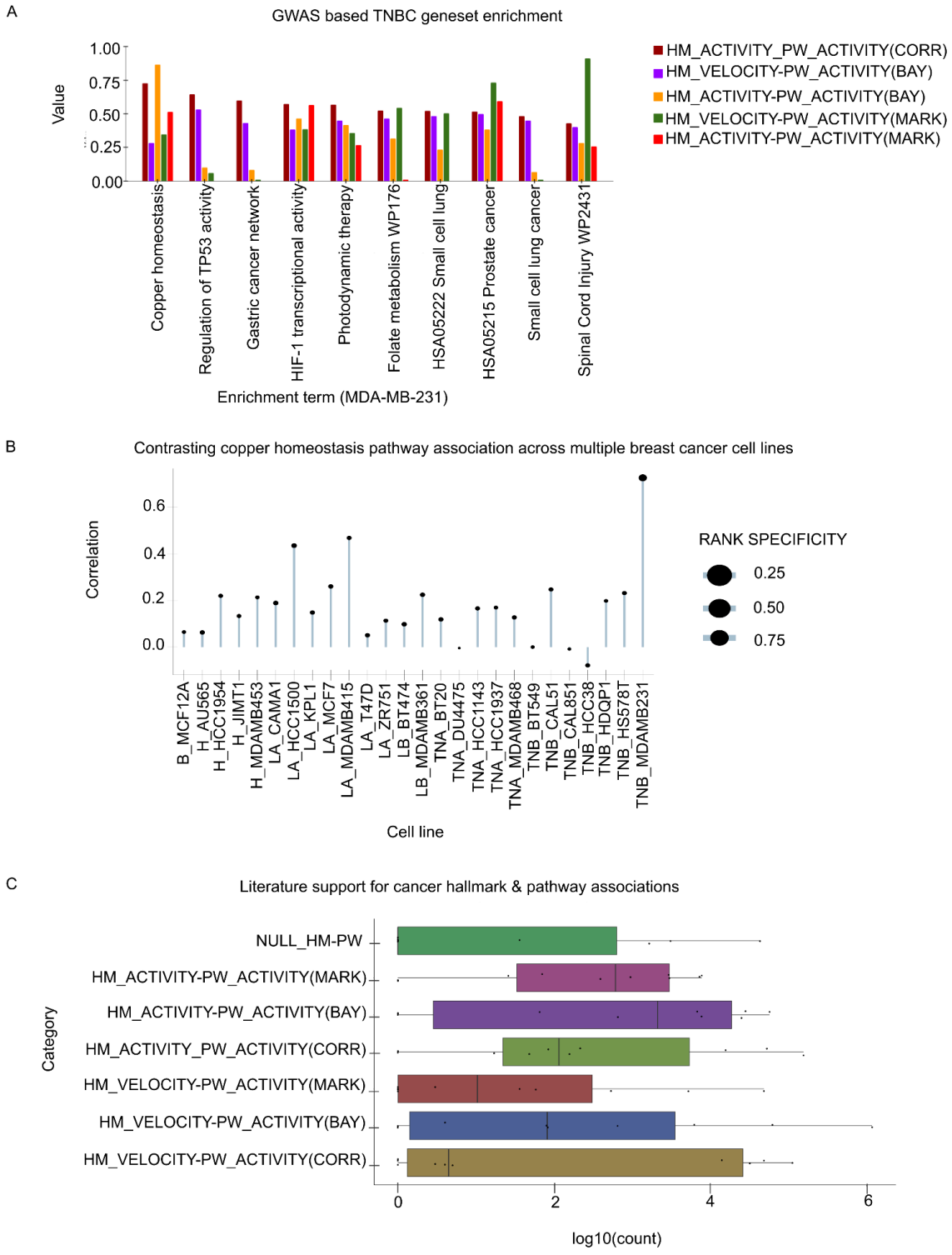

Figure 3.4: Context specific associations between cancer hallmark properties and pathways in MDA-MB-231. **A)** This figure is an illustration of the top ten pathways that were found to be shared by the pathways that were filtered from PGM and the GWAS-based geneset enrichment for MDA-MB-231. **B)** We evaluated the exact same

connection between the pathway and the hallmark property in a variety of breast cancer cell lines in order to provide further evidence that the association between the two is unique to the cell line. **C)** In order to validate the additional associations of MDA-MB-231 that have been identified by our analysis, we carried out a literature search to look for the co-occurrence of cancer hallmarks and associated pathway phrases.

3.3.4 Using context-specific associations to predict drug response

In an effort to determine the significance of the hallmark and associated pathways in relation to the context of cell line specificity, we attempted to model the activity of various cancer cell lines in relation to the AUC value of the drug. Twenty-five breast cancer cell lines were shortlisted based on the availability of their drug response in the CTRPv2 database. We used the kNN model to predict a drug's AUC value using machine learning. At first, we chose the union pathways that are linked to the cancer hallmark of "resisting cell death" as the features for our machine learning model. This model is used to determine the AUC value for drug response predictions for 25 breast cancer cell lines. Additionally, we utilized the machine learning model for estimation of the AUC by employing both the complete pathway sets and randomly generated pathway sets. Figure 3.5A illustrates the correlation between the actual and predicted AUC values for the three approaches employed. The kNN model, utilizing pathways obtained through our method, demonstrated superior predictive capability for the response to 544 drugs compared to the other two methods (all pathway set and random pathway set). To further emphasize the relevance of context-specific drug response, we made an effort to contrast the pathway-based drug response with a random drug response, specifically focusing on a particular cell line (MDA-MB-231). As shown in the Figure 3.5B, the results clearly demonstrate that pathway-specific drugs exhibited higher sensitivity (lower drug AUC values) compared to randomly chosen drugs. This contrast underscores the importance of context-specificity in predicting drug responses. Essentially, the pathway-based selection method is more aligned with the unique characteristics of the specific cell line, highlighting how context influences drug effectiveness. We believe this approach strengthens the case for context-specific drug response prediction, as seen in the improved drug sensitivity observed in pathway-targeted treatments.

Therefore, our approach can be utilized to generate hypotheses for investigating the beneficial and adverse impacts of drug-targeting pathways within specific cell types. In general, our approach, which makes use of single-cell profiles and RNA-velcoity, has the potential to assist in the reduction of the search field for the purpose of discovering alternative drugs for cancer

therapy that are cell specific. Furthermore, it is effective in predicting the effect that these drugs might have on cancer hallmark properties.

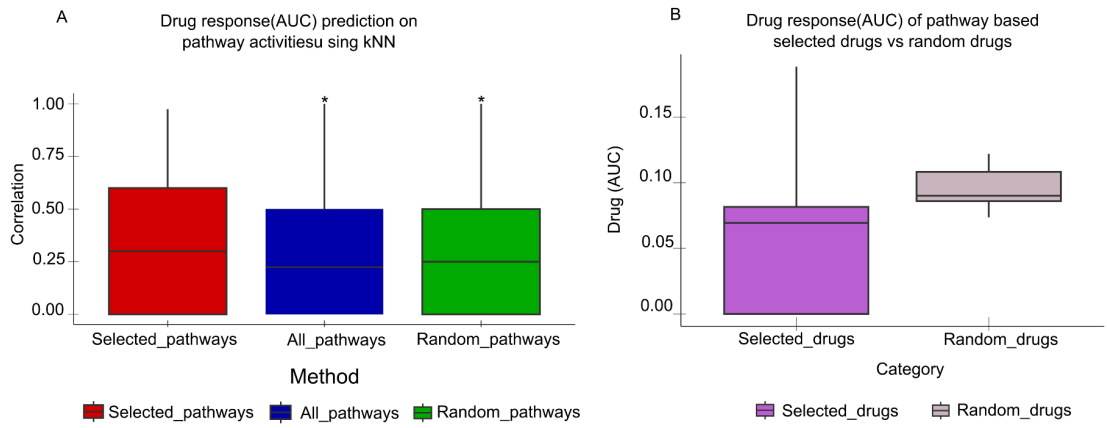

Figure 3.5: Drug response prediction using cell line specific hallmark pathway associations. **A)** Here, we used the kNN model to predict a drug's AUC from CTRPv2 database value for 25 different breast cancer cell lines. The kNN model, utilizing pathways obtained through our method ("Selected_pathways"), demonstrated superior predictive capability for the response to 544 drugs compared to the other two methods ("All_pathways" set and "Random_pathways" set). **B)** Comparison of pathway based vs. random drug response in the MDA-MB-231breast cancer cell line. This contrast demonstrates how targeting relevant biological pathways for a particular cell line can lead to more effective therapeutic outcomes, underscoring the relevance of personalized drug response strategies.

3.4 Discussion

We have introduced a computational pipeline that addresses the issue of determining direct relationships between cancer hallmark properties and pathways. This is achieved through an integrated approach that combines single-cell expression profiles and RNA velocity. We have utilised the diversity of individual cells and the patterns of RNA velocity to deduce interconnections between pathways and hallmarks. To clarify the importance and specificity of a connection across pathways, hallmarks, and cell types, we utilize extensive single-cell atlases to standardize the estimates of dependencies. In previous studies, researchers have utilised mutation profiles to establish connections between pathways and hallmark properties (Iorio et al. 2018; Chen et al. 2021). However, most previous analyses have been hampered by heterogeneity in cells, which blocks the inference of related cell-type and blurs differential expression results, and enrollment of patients issues that prevent enough samples from being collected. Therefore, our method offers an appropriate solution to the problems of cancer cell line specificity and the

interconnections between pathways and hallmark properties. Utilizing gene set enrichment instead of individual genes provides an benefit in terms of being less influenced by drop-out or sparsity in single-cell expression profiles.

We have employed multiple strategies, including filtering for cell quality, filtering for gene set relevance, and normalisation by a vast and diverse set of cell-type profiles to emphasise specificity, to further strengthen our approach. Furthermore, we have demonstrated the significance of our resulting associations through a high-throughput validation process exploiting a PubMed literature search. Additionally, we have demonstrated how our method enables the deduction of variations in the relationship between cancer hallmark characteristics and pathways among various cancer cell lines. Monitoring the changes in these connections aids in determining the distinct functions of specific pathways in relation to shared hallmarks in cell lines. Our findings indicate that mTOR plays a similar role in both MCF-7 & MDA-MB-231 breast cancer cell lines, while lipogenesis has a distinct role in these cell lines. These variations are crucial when examining disease prognosis, therapy, and drug resistance. In addition, our study emphasised the distinctiveness of copper regulation in the MDA-MB-231 breast cancer cell line compared to other cell lines associated with breast cancer. Although numerous investigations have shown various aspects of cuprotosis, we here highlighted the direct dependency between copper homeostasis and the hallmark property of "evading cell death".

One advantage of leveraging single-cell expression profiles is the ability to model cellular heterogeneity from an individual at a given time. Therefore, it allows for the identification of differences among various individuals. Therefore, employing our method to analyze cancer cells from an individual has the potential to create new opportunities for precision medicine. Furthermore, by making use of the filtered hallmark-associated pathways, our method demonstrates superior predictive capability for drug response compared to randomly selected or whole sets of pathways. These observations provide a foundation for investigating novel aspects of cancer treatment that are tailored to the specific cancer cell type and its unique characteristics.

Moreover, although our approach can effectively identify direct connections between pathways and cancer hallmarks, it does not necessarily provide causal inference. Nevertheless, pathways with a greater direct dependency are more likely to be causally related than random gene sets. Nevertheless, our approach narrows down the range of possibilities to identify the genuine causality factors that can be investigated further in laboratories.

CHAPTER 4

UTILISING GENE ENSEMBLES DEFINED BY CHROMATIN DOMAINS WITH TRANSCRIPTOMICS TO UNDERSTAND DRUG-RESPONSE AND PHENOTYPIC HETEROGENEITY IN CANCER CELLS

4.1 Introduction

4.1.1 Cancer gene expression regulation via chromatin domains

The harmonious operation of chromatin arrangement and the functioning of various genomic regions play crucial roles in the transformation of cellular states and cellular response. Multiple research projects have focused on establishing connections between the features of cancer cells and the arrangement of chromatin (L. Li et al. 2019; Polak et al. 2015). While genetic mutation and evasion of apoptosis play a vital role in the formation of cancer, regulatory alterations in the epigenome and gene expression consistently become crucial throughout this transformation (Pisco & Huang 2015). The Cancer Cell Line Encyclopaedia (CCLE) study group tried to predict the therapeutic response to drugs by analyzing mutation and expression patterns (Barretina et al. 2012). Several efforts have been undertaken to provide alternative explanations that are not related to genetics, such as reconfiguration and stemness. These factors have been associated with enhanced resistance to foreign substances and higher production of efflux pumps. Additionally, they are connected to improved DNA repair capabilities, resilience, and reaction to stress (Dean et al. 2005; Donnenberg & Donnenberg 2005). The presence of diverse tumor populations, including a small subset of cancer cells capable of transitioning between different

states, has also been identified as a contributing factor to drug resistance. Regardless of the context, chromatin organization plays an essential role in cancer cell growth and their resistance to conventional drugs (Kleppe et al. 2018; Morgan & Shilatifard 2015).

4.1.2 TADs as regulators of disease progression

Chromatin can be categorized into many domains, called TAD. Alternatively, TADs may be described as genomic areas that are naturally characterized and separated via distinct loci (Dixon et al. 2012; Liu et al. 2019). TADs have been shown to exhibit significant stability across a variety of cell types (Rao et al. 2014). Indeed, some research has shown that homologous regions in the human and mouse genomes maintain around 54% of TAD boundaries (Dixon et al. 2012). The pattern of gene co-localization is also influenced by cellular along with evolutionary mechanisms. For example, the gene-dense locus 3p21.31 on chromosome 3 is recognized as the site of the most extensive cluster of tumor suppressor genes (Jain et al. 2021). The co-localization and gene synteny engaged in similar or related tasks may be attributed to many factors, such as the prevention of recombination of the beneficial alleles at a specific location or the simultaneous expression of these genes. Salem et al. demonstrated that paralogs have a higher tendency to be located inside the same TADs compared to randomly paired genes (Ibn-Salem et al. 2017). The mutation rates inside TADs and their borders have been investigated in relation to cancer development (Akdemir et al. 2020). Various variables, including CTCF, G4 quadruplexes, and repetitions, may contribute to the development and reinforcement of TAD borders (Zhang et al. 2023; Ni et al. 2023). Barutcu *et al*., have shown that removing CTCF-rich areas in close proximity to TAD borders does not always disturb the TADs (Barutcu et al. 2018). In a similar vein, Akdemir *et al*., found that only 14 percent of TAD border deletions led to a two-fold shift in the expression of nearby genes (Akdemir et al. 2020). Therefore, it is not easy to determine the impact of a mutation on the preservation or disruption of TAD boundaries. Thus, the deviation in the functioning of genes inside a TAD might provide a more immediate indicator of the impact of the TAD border. Furthermore, there are several concerns about TADs and their impact on the control of cancer cell characteristics and their response to drugs, as well as the potential benefits of TADs for clinical research. In order to address these issues and get solutions to various concerns, we have developed a methodology for assessing TAD activity by measuring

the increased expression of a group of genes inside a TAD. This allows us to establish connections between gene expression, treatment response, and the severity of cancer.

Initially, we examined the chromatin interaction architecture in cell lines of squamous cell carcinoma of the head and neck (HNSC) to get a better understanding of the behavior of TADs in relation to drug resistance. Subsequently, we broadened our investigation by examining the transcriptome of 819 cell lines published in the CCLE database and their responses to 544 drugs listed in the Cancer Therapeutic Response Portal (CTRP) (Rees et al. 2016). Additionally, we used our methodology to analyze transcriptome data of 9014 subjects spanning 20 different forms of cancer, which were provided by the TCGA database consortium (Cancer Genome Atlas Research Network et al. 2013).

4.2 Material And Methods

4.2.1 Hi-C contact matrix generation and TAD detection

Pandey *et al*., conducted a study of ChIP-seq and single-cell expression datasets in addition to Hi-C profiles (Kumar & Others 2022). In order to generate the Hi-C contact matrix, the binary alignment mapping file was transformed into the arrowhead input format with the help of samtools (Danecek et al. 2021). The arrowhead data input was utilized to generate a normalized Hi-C contact matrix (.hic file) using the Juicer tool (Durand et al. 2016). The DomainCaller methodology developed by Dixon *et al*., was further used to detect topologically associated domains at a resolution of 25 kb across all chromosomes using in-house generated data (Dixon et al. 2015). Furthermore, TAD boundaries were collected using the TADKB database (Liu et al. 2019). The TAD sites for each cell type were determined using the Directionality Index (DI) (Rao et al. 2014; Pandey et al. 2023) .

4.2.2 Associating TAD activity and drug responsiveness

TAD gene sets were created by identifying the regions of overlap between the gene promoters and the TAD boundary positions in the hg19 human genome. The activity score of TAD is determined by measuring the degree of enrichment of its gene set. We used the Gene Set Variation Analysis (GSVA) method to determine the functional enrichment of gene ensemble inside TADs across all of the samples in our study using bulk RNA-seq expression profiles

(Hänzelmann et al. 2013). GSVA is a non-parametric, unsupervised method that eliminates the conventional approach of direct phenotype modeling within the enrichment score. The TCGA dataset, available at https://portal.gdc.cancer.gov/, consists of bulk gene expression profiles obtained from patients across different kinds of cancer (Grossman et al. 2016). These transcriptomic profiles were then converted into TAD-activity scores. To further compute TAD-activity scores, 819 samples from various cancer cell lines were also taken from CCLE (Cancer Cell Line Encyclopaedia) (Barretina et al. 2012). The drug response was obtained using the CTRP version-2 database. 819 CCLE-cancer cell lines were selected based on the availability of their treatment response data for 544 drugs in the CTRP version-2 database (Rees et al. 2016). The pIC50 value for drugs was utilized in this case. The pIC50 value is determined by taking the negative logarithm (base 10) of the drug IC50 value. Spearman correlation coefficient was evaluated between the TAD-activity scores associated with cancer cell lines and their corresponding pIC50 values for each drug.

4.2.3 Estimating patient survival P-value

To determine the importance of the relationship between gene expression and TAD-activity with TCGA patient survival, the R packages named survival and survminer were utilized (Therneau & Grambsch n.d.) (Therneau & Grambsch n.d.). The input files consisted of TAD-activity ratings and clinical data, including information like days till death, ethnicity, etc. These datasets were obtained via the Xena browser (Goldman et al. 2020). The 'Status' column in clinical data represents the binary values of 0 and 1, which correspond to the states of being alive and deceased, respectively. To determine the survival P-value for every TAD, the median value of each row in the GSVA data was computed. Additionally, a new column called "median-group" was generated, with a value of 1 indicating that the GSVA value is larger than the median and a value of 0 indicating that the GSVA value is lower than the median. The clinical data, including the number of days till death, the patient's status as 0 or 1, and their unique identifier, was combined with the GSVA scores and the column representing the median group. Therefore, for each TAD, we generated a file containing the necessary information to be used in the R package survival and survminer. The P-value for survival associated with each TAD is included in the ‘Survfit’ output. Furthermore, we computed the hazard ratio for each TAD using the ‘coxph’ function. The hazard ratio, in this case, would refer to the potential risk linked to the activity of

corresponding gene ensemble within the TAD. Various threshold combinations, including 51-49, 60-40, and 70-30 percentiles, were used to determine the relevance of TAD in survival (Pandey et al. 2023). These thresholds were used to classify patients into groups based on TAD activity in conjunction with the MaxStat package (Lausen & Schumacher 1992). Following that, which were prevalent in those pairings, were chosen for more investigation.

In order to further investigate the relationship between patient survival and drug resistance, we established two distinct collections of TADs, which we referred to as the Positive_set and the Negative_set. These sets had a correlation that was more than 0.3 and more than -0.3, respectively. We analyzed their survival likelihood ratio across various categories. In addition, we also assessed TADs distribution over various thresholds of the hazard ratio.

4.2.4 Examining the synteny

We acquired the genes that are orthologous and their respective chromosomal locations for sixty nine mammals from the Ensembl database (https://asia.ensembl.org/index.html). Furthermore, we acquired the genome fasta files of species that did not appear in Ensembl directly from the DNA Zoo database (https://www.dnazoo.org/). Subsequently, we matched the genome sequences with the human genome with the help of the last program (https://gitlab.com/mcfrith/last). Subsequently, the alignment chain data were used to establish the corresponding locations of orthologous genes. We used the 'plotSegments' and 'plotRect' functions from the 'plotgardener' R-package version to generate visual representations of the synteny maps derived from the transcription start site (TSS) of orthologous genes among mammalian species (Kramer et al. 2022).

4.2.5 Assessing the enrichment of GWAS variants linked to EGFR-TAD

We compiled an array of null model TADs consisting of randomly selected genomic areas that are about the same size as the EGFR-TAD (chr7:51500000-57100000). The GWAS-based mutations were compared with the null-model-based TAD positions and EFGR-TAD to determine the number of mutations that were linked. The enrichment of related mutations in the EGFR-TAD for the disease X was determined as follows:

$$Enrichment\ (X)\ =\ \frac{Count\ of\ GWAS\ mutations\ for\ X\ in\ EGFR\ TAD}{The\ average\ number\ of\ the\ GWAS\ mutations\ per\ X\ in\ null\ model\ TADs}$$

4.3 Results

The use of our unique methodology to examine chromatin interaction patterns and single-cell transcriptome profiles of patient derived HNSC cell lines has established the groundwork for analyzing extensive transcriptome datasets of cell lines sourced from the CCLE database and tumors obtained from TCGA databases. We analyzed the chromatin-interaction patterns of HNSC cell lines using the Hi-C technique. We utilized HNSC patients-derived primary cultures (PDCs) that were generated by Chia *et al*. (Chia et al. 2017). These cultures consist of primary HNSC cells from two patients' primary tumors (HN137P,HN148P) and THE metastatic cancer cells from patient HN120 (HN120M). The TAD activity is calculated by using the GSVA method, which utilizes the bulk expression profile. Figure 4.1 also displays a concise summary of the study (Pandey et al. 2023).

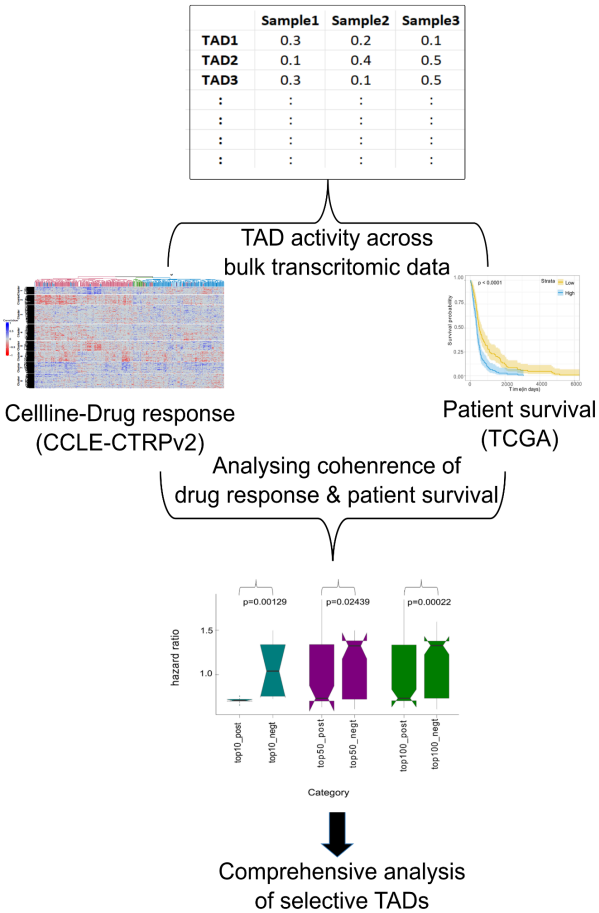

Figure 4.1: Investigation of chromatin interaction patterns and domain activity in cell lines originating from patients of head and neck cancer. These cell lines include the metastatic tumor cell line HN120M derived from patient HN120, as well as the original tumor cells HN137P along with HN148P derived by patients HN137 and HN148, respectively. The TAD boundaries produced from Hi-C profiling of HNSC cell lines were obtained through the union list and utilized for further analysis.

4.3.1 Examining TAD activity patterns and their associations with drug susceptibility

We created a comprehensive list of TADs identified in our head and neck squamous cell carcinoma (HNSC) cell lines, as well as TADs documented in the TADKB database (Liu et al. 2019). Our methodology relies on the findings of several studies about the preservation of TAD boundaries across various cell types (Dixon et al. 2012; Dixon et al. 2015). Therefore, we used a shared compilation of TADs to compute TAD-activity scores for 819 CCLE cell lines. We next examined the correlation among TAD-activity and the drug pIC50 value that were documented for the corresponding cell lines at CTRPv2. The null-model of TADs, which have random boundaries, exhibited a reduced correlation among their activity levels and pIC50 values of drugs in comparison to actual TADs.

During our analysis of the CCLE database, we observed a noteworthy pattern when examining the relationship across TAD activity and drug response across all cancer cell lines. Specifically, we found that for numerous TADs, there was no significant correlation among the expression of their own genes and the pIC50 value for the same drug. The relationship between drugs pIC50 such as Ciclopirox, Lenvatinib, LY-2183240, Foretinib and Axitinib, and TAD-activity was much stronger than their association with any of their genes (Figure 4.2) (Pandey et al. 2023). The findings suggest that utilizing the enrichment-score of a gene set of TADs may be a more effective method for evaluating response to different drugs compared to using single genes.

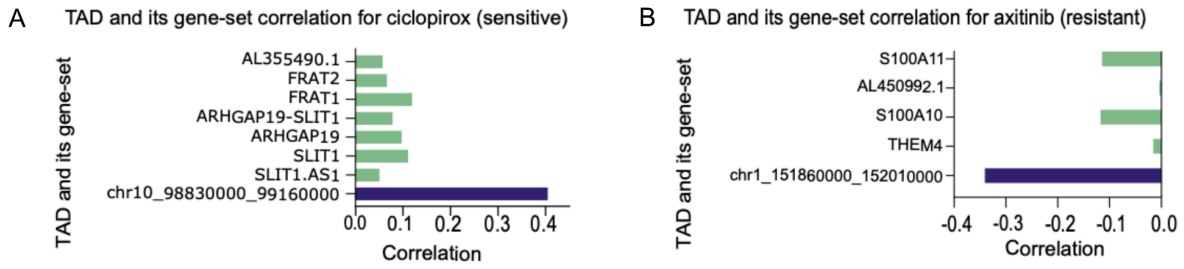

Figure 4.2: Drug response and TAD-activity association patterns. **A)** The correlation between the pIC50 of the drug Ciclopirox and the activity of chr10_98830000_99160000 TAD (blue) and its corresponding genes. As shown, the individual genes do not exhibit as strong a connection with the pIC50 value of the given drug as the activity of their respective TAD. **B)** Comparing the correlation values between the activity of chr1_151880000_152010000 TAD (blue) and the gene expression of its corresponding gene set with the pIC50 value of drug Axitinib (Pandey et al. 2023).

4.3.2 Comparing the prognostic value of TAD and gene-based patient survival

The fluctuation in gene activity within a TAD in tumors may be attributed to many factors, including alterations in gene regulation, mutations at the TAD boundaries, or genomic alterations in structure. variation in copy number (CNV) of the TAD-containing genomic regions. Notwithstanding, the TAD-activity score could be exploited to monitor their substantial impact on gene activity. Therefore, we examined changes in TAD activity by analyzing the cancerous transcriptomes furnished through the TCGA consortium. We utilized the same combined list of TAD as the study conducted on the CCLE data set. In addition, our study specifically examined TADs that are linked to survival outcomes in 20 distinct cancer types. The number of TADs that were related to survival in two or three forms of cancer was significantly higher than the number of TADs that were associated with the survival in just one type of cancer (Figure 4.3A) (Pandey et al. 2023). The introspection of hazard ratios indicated that several TADs had significant connections (P-value < 0.05) with survival in multiple types of cancer. Furthermore, the impacts of these TADs were shown to change depending on the specific cancer surroundings, as shown in Figure 4.3C.

It is possible that the link between a TAD and survival might be due to the influence of a small number of genes that have a greater impact on survival than the activity of the TAD itself. Nevertheless, our findings indicate that the activity of several TADs had a greater impact on survival compared to the individual genes included within them. In HNSC, CESC, and BRCA

cancer, respectively, the proportion of such TADs that had a greater impact on survival than any of their genes was 57%, 44%, and 56% (Figure 4.3B). A TAD, including the EGFR gene, situated at chr7:51500000-57100000, has a stronger correlation (P-value < 1E-4) across its activity and survival compared to any of its respective genes in HNSC (Figure 4.3D), BRCA, and CESC. Additional instances of TCGA-HNSC include TADs situated at chr1:156110000-156900000 and chr3:97760000-98290000. These TADs have a stronger correlation with survival than any of their corresponding genes included inside them (Figure 4.3E-F). In summary, TAD-activity-based survival analysis suggests that the co-localization of genes and their coordination play a significant role in promoting cancer malignancy by influencing important biological processes.

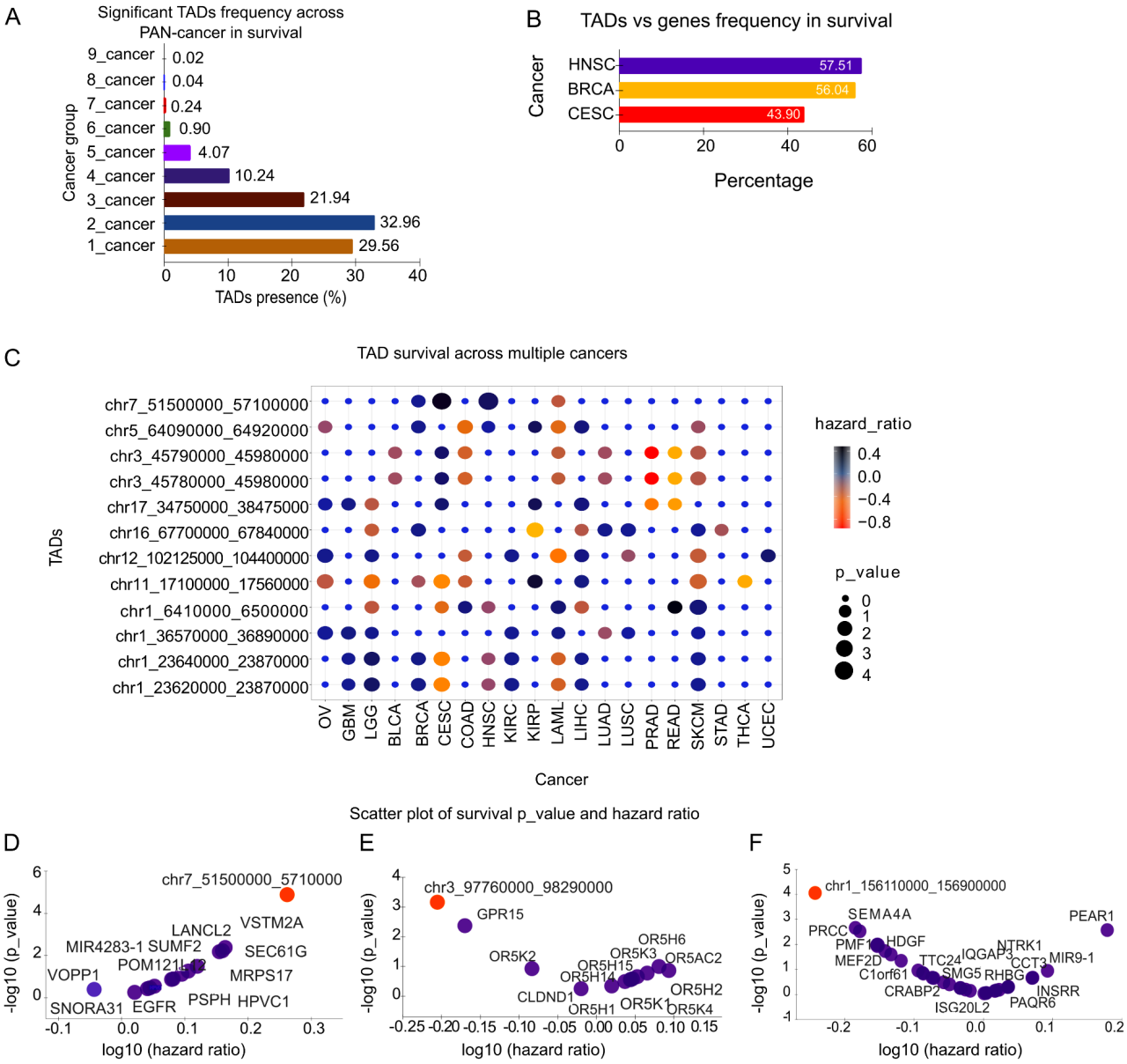

Figure 4.3: Analysis of TADs in various cancer samples from TCGA and the relationship with patient survival. **A)** The graph displays the distribution of TAD numbers linked with survival in one, two, or three cancer types. The P-value cutoff to obtain association has been set at 0.05. The fraction was obtained by dividing by the proportion of TADs that demonstrated an association with patient survival in a minimum of one cancer category. **B)** Proportion of TADs that have a greater impact on survival than any of the genes inside them. The proportion was determined based on all TADs that showed a correlation with survival in the specified cancer type. **C)** A dot-plot is shown, illustrating the hazard ratios and P-values for the relationship between TADs and survival across numerous kinds of cancer (PAN-cancer). **D)** TAD (chr7_51500000_57100000) and its genes' survival connection via scatter plots

depicting the P-values observed in TCGA-HNSC patients. **E)** Scatter plot demonstrating the comparisons of gene survival P-values for the TAD (chr3_97760000_98290000) in TCGA-HNSC. **F)** Contrasting TAD (chr1_156110000_156900000) and gene survival P-values in TCGA-HNSC (Pandey et al. 2023).

4.3.3 Correspondence between drug response and TAD-activity based patient survival

We assessed whether survival results in cancer patients with TAD-activity were coherently associated with their treatment responsiveness in cancer cell lines (refer to Supplementary Methods). We organized the TADs linked to survival i.e., with a P-value less than 0.05 in the HNSC patients from TCGA based on the count of drugs that exhibited absolute correlation greater than 0.3 across HNSC cell lines within the CCLE repository. In terms of the hazard ratio, the top 50 TADs (P-value < 0.05) that provided resistance (negatively correlated with drug pIC50 & r < -0.3) towards the maximum number of drugs had a considerably higher hazard ratio (P-value < 0.01) compared to the top fifty TADs that correlated with sensitivity against the greatest diversity of drugs (Figure 4.4A) (Pandey et al. 2023). Similar trends were seen whether the top ten or top hundred TADs were employed. Among the top ten TADs (with a survival P-value less than 0.05) that showed a correlation with resistance and the biggest number of drugs (with a correlation coefficient less than -0.3 and pIC50), 50% had hazard ratios greater than 1.25. None of the top ten TADs, which had a survival P-value less than 0.05, showed a correlation with sensitivity towards the highest number of drugs (with a correlation coefficient more than 0.3). Additionally, none of these TADs had a hazard ratio over 1.25, as shown in Figure 4.4B. Figure 4.4C shows that the same pattern was seen for the top fifty or top hundred TAD categories. These findings brought to light the coherence that exists between the effect of TAD on drug response in cell lines and malignancies in patients, as well as the valuable TAD-based features that are related to both the severity of cancer and resistance towards several drugs.

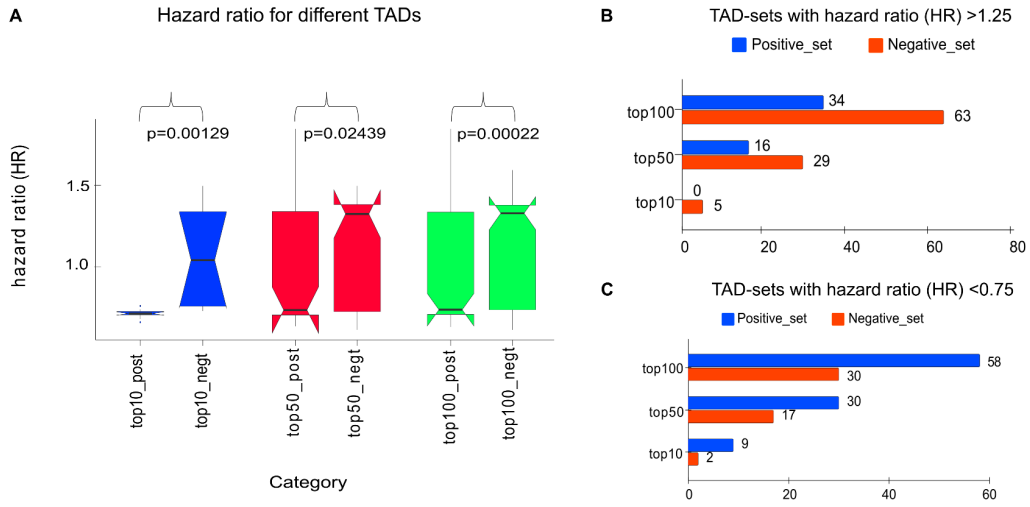

Figure 4.4: Correspondance between the connection of TAD activity with survival across TCGA patients and the association between TAD and drug response in CCLE cell lines is examined. **A)** The distribution of hazard ratios of the top ten, fifty, and hundred categories of TADs correlated with drug response. The "top10_post" variable includes the TADs that have the highest frequency of positive correlation (R > 0.3) with drug pIC50. In a similar way the "top10_neg" dataset includes the TADs that exhibit the greatest 10 frequencies of negative association with the pIC50 of drugs. The groupings “top50_post”, “top50_negt”, “top100_post”, and “top100_negt” are exhibiting trends that are similar. The hazard ratio associated with TAD activity exhibited an inverse association with the pIC50 value (or a direct correlation with IC50), indicating a value over one (indicating poor survival). Conversely, the same is true for TADs that are positively related. **B)** Frequency of topologically associated domains (TADs) in various categories (“top50_pos” or “top50_neg”) with hazard ratio cutoffs larger than 1.25 in the TCGA dataset. **C)** Frequency of topologically associated domains (TADs) in distinct groups, with hazard ratios smaller than 0.75 (Pandey et al. 2023).

4.3.4 Case report on TAD associated with the EGFR gene

The increased impact of TAD activity at chr7:51500000-57100000 on survival, compared to its gene EGFR, in several forms of cancer, has prompted inquiries into the significance of co-localization and the role of other genes within this region. To be concise, we refer to the TAD situated at chr7:51500000-57100000 as EGFR-TAD. Figure 4.5A shows that the EGFR-TAD is often amplified in the tumors of a small percentage of TCGA-HNSC (HNSC) patients (Pandey et al. 2023). Our research revealed that the genes in the EGFR-TAD exhibited synteny in apes and old world monkeys but lacked synteny in other mammals along with new world monkeys (Figure

4.5B). We looked further into the possibility that apes and old world monkeys have the same biological mechanism that contributes to brain development and cancer severity, given the increased gene synteny seen in EGFR-TAD between the two species. The impact of copy number variations (CNVs) in the genes belonging to the EGFR-TAD on survival, when compared to the EGFR gene, has been shown in many forms of cancer, such as HNSC and glioblastoma (Figure 4.5C). The analysis suggests a strong likelihood of genes being amplified and active together in the EGFR-TAD. Our analysis revealed a significant presence of genetic variants related to serine and glycine levels in the blood, the size of the brain, and glioblastoma within the EGFR-TAD dataset (Figure 4.5D). Figure 4.5E displays the enrichment scores of the top 10 enriched features and disorders. The gene PSPH in EGFR-TAD has a role in the creation of serine inside the body via aerobic glycolysis. This serine production is necessary for the functioning of cancer cells (Shunxi et al. 2023). The significant concentration of GWAS mutations linked to brain volume and the acquisition of synteny in primates with bigger brains suggests that there has been primate-specific adaptive selection in this genetic region. However, this adaptation comes at the expense of an elevated vulnerability to cancer.

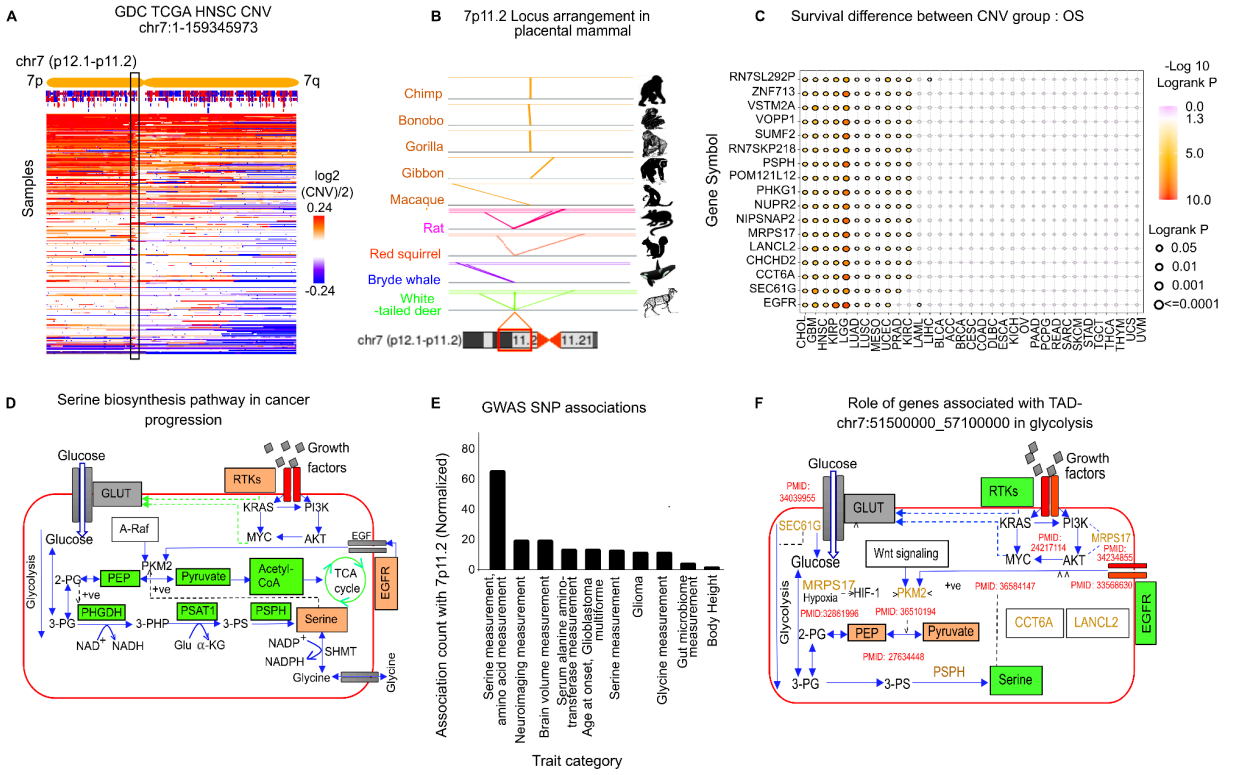

Figure 4.5: Providing a comprehensive case study on the TAD that encompasses the EGFR gene. **A)** The representation of changes in the number of copies (CNV) of DNA segments in chromosome 7 in head and neck squamous cell carcinoma (HNSC) within the TCGA cohort. The TAD segment chr7:51500000-57100000, which encompasses the EGFR gene i.e, EGFR-TAD, is visually represented by a box. **B)** An overview of the preservation of gene synteny in the EGFR-TAD. **C)** The correlation between the survival rate of TCGA-HNSC patients and copy number variations (CNV) of genes within the EGFR-TAD region. **D)** This figure provides an outline of the main steps involved in the production of serine via the intrinsic biosynthesis pathway, which occurs as a byproduct of glycolysis. **E)** The top enriched features are determined by the presence of single nucleotide variations in the EGFR-TAD region, as identified by Genome Wide Association Studies (GWAS). **F)** A summary of the genes involvement in EGFR-TAD in the process of glycolysis and other associated activities (Pandey et al. 2023).

Analysis of CCLE cell-line transcriptomes identified drugs that exhibit the most positive associations between pIC50 value and the activity of EGFR-TAD. These pharmaceuticals primarily target the AKT pathway or activities associated with aerobic glycolysis. The three drugs, piperlongumine, cerulenin, and AZD6482, were found to be positively correlated with EGFR-TAD in breast cancer cell lines (BRCA). These drugs have been shown to directly as well as indirectly alter glycolysis via the PI3K/Akt pathway. The drug Netulin-3, which is an inhibitor of MDM2 and has been shown to reduce the generation of serine and glycolysis, demonstrated sensitivity with the activity of EGFR-TAD in all three cancer types (LUSC, BRCA, and HNSC). Several genes in the EGFR-TAD region facilitate glycolysis and associated activities (Figure 4.5F), as well as activate the AKT pathway. The PI3K-AKT pathway is recognized for its ability to enhance aerobic glycolysis (Hoxhaj & Manning 2020). The data suggest that the genes in EGFR-TAD underwent synteny in apes, which facilitated cell proliferation via two distinct mechanisms. The first mode involves the activation of the PI3K-AKT pathway. The second mechanism involves facilitating aerobic glycolysis to produce biomolecules such as serine and glycine, which are essential for cellular growth. A more reliable drug selection process that is based on the EGFR-TAD activity may be achieved via the acquisition of knowledge about the systematic collaboration of genes.

4.4 Discussion

In the current study, our computational framework was used to investigate the diversity of cancer and its response to drugs. Additionally, TAD activity may reflect in part the enrichment of

functionality of other types of regulatory components inside a TAD, such as enhancers. The abundance of conserved connections within TADs suggests that TADs function as regulatory units. The limited impact of TAD boundary mutation and the potential involvement of several unidentified components in TAD boundary formation provide compelling support for our appraoch, which utilises TAD-activity as a direct means of assessing the consequences of disrupting TAD-based regulatory units (Monteza et al. 2020; Barutcu et al. 2018; Zhang et al. 2023). In addition, Mohanty *et al*., have shown that there is not always a direct relationship between the variation in the number of copies of genes and their expression levels (Mohanty et al. 2021). Therefore, the copy number variation (CNV) of a TAD area may not always indicate its level of activity. These results once again emphasize the significance of directly using gene expression to measure the activity of TADs in order to explore their underlying regulators.

In contrast to conventional biomarkers, TAD activity provides potential benefits in terms of efficacy by elucidating a more comprehensive understanding of gene regulation, particularly in diseases where chromatin architecture is a critical factor. In terms of specificity, TAD activity may be especially important for diseases that are associated with chromatin remodeling or structural variations. In terms of reliability, whereas advancements in genomic technologies might facilitate accurate assessment of TAD activity, problems persist linked to interindividual variability and the technical consistency of data.

Our data unequivocally demonstrates that TAD activity possesses the capability to act as some sort of biomarker for more accurate prediction of treatment response or assessment of survival. TAD activity often surpasses individual gene expression as a more effective indicator for drug responsiveness and analysis of survival. Redundancy may be avoided by developing a methodology for identifying potent combinations of drugs based on the understanding of relationships between TAD-activity and drug-response. However, it is important to remember that TAD-activity may not be able to accurately anticipate how a patient would react to all drugs. The finding of a connection between the severeness of cancer and an evolutionary drive linked to the growth of bigger brains in apes also offers insight into the species-specific impact and unintended consequences of oncotherapeutic medications. An example of a potential side effect is the influence on the cognitive development of primates. Furthermore, the use of EGFR-TAD to sustain aerobic glycolysis has been linked to immunological evasion of tumour cells via the regulation of T-cell activity, in addition to its role in cell proliferation and drug-resistance (Chang

et al. 2021). Therefore, the case of EGFR-TAD provides a unique opportunity to examine the transcriptome by using the TAD gene set information. This allows for the investigation of the functional causes behind the maintenance or acquisition of gene synteny, which may help in understanding the underlying mechanisms of different disorders.

CHAPTER 5

INTROSPECTING COMPUTATIONAL CHALLENGES ASSOCIATED WITH CELL-FREE CANCER DIAGNOSIS USING METHYLATION PROFILES ACROSS ENSEMBLES OF GENOMIC LOCI

5.1 Introduction

5.1.1 Comprehending the origins and characteristics of cfDNA

Conventional clinical diagnostic techniques, for example tissue biopsies or bone marrow, are intrusive and susceptible to sample bias. As a result, researchers are seeking new molecular markers. Recently, liquid biopsy-based methods for diagnosing diseases have become more significant since they provide a more secure and quicker alternative to tissue-based examinations (Feng et al. 2019). A liquid biopsy derived approach utilizes cancer traces collected from cell-free DNA (cfDNA). The fragments are referred to as circulating tumor DNA, also known as ctDNA, and have demonstrated the capacity to aid in cancer detection and prognostic.

Although there is an abundance of research on cfDNA, our understanding of the exact molecular origins of cfDNA is still limited. Recent studies have shown that the cfDNA release in the blood is influenced by several processes, including apoptosis, pyroptosis, necrosis, autophagy, NETosis, and cf-mtDNA (Aucamp et al. 2018; Grabuschnig et al. 2020). The size of these cfDNA fragments ranges from 1000 to 3000 base pairs, which is different from the shorter pieces produced during apoptosis (ranging from 90 to 166 base pairs) (L. Liu et al. 2021) . Furthermore,

cfDNA found in the bloodstream may exist either as unbound DNA or as part of complex structures such as nucleosomes, membrane fragments, vitrosomes, or enclosed inside extracellular vesicles (EVs) such as microvesicles, exosomes, and apoptotic bodies (Panagopoulou et al. 2021). Signals arising via cfDNA fragmentation patterns, nucleosome positioning, transcription factor binding, regions around transcription start sites, and peripheral cellular modifications may be used to diagnose diseases. However, the process of inferring patterns in downstream analysis is impacted by intrinsic properties of information generated from cfDNA, such as noise, sensitivity, and length of DNA fragment (Zheng et al. 2021) .

In the context of cancer, it is important to note that cfDNA is not only produced by tumour cells, but non-cancerous cells also contribute significantly to its release. The signal from malignant cells is distorted by the release of cfDNA from non-cancerous cells, making the data more erratic and noisy (Bronkhorst et al. 2019) . In addition to other contributing variables, the rate at which cfDNA is cleared from plasma is also a key factor in cancer detection (Khier & Lohan 2018).

Gene ensemble-based analysis of cfDNA looks at gene groups, usually arranged according to biological pathways or functions, to find molecular alterations linked to diseases like cancer. For example, the analysis of gene ensembles based on cfDNA fragmentation pattern, contributes to a more comprehensive understanding of disease biology by detecting altered pathways and evaluating the efficacy of treatment in a non-invasive manner (Noë et al. 2024; Peneder et al. 2021). Similarly, the concept of evaluating TADs in cfDNA is compelling, as cfDNA comprises genetic fragments that reflect segments of the genome, including those linked to TADs. (Zukowski et al. 2020). However, directly evaluating TADs in cfDNA poses many hurdles owing to fragmentation and the intricate architecture of TADs. Nevertheless, as cfDNA analysis tools advance, especially in high-resolution mapping and epigenetic profiling, the examination of TADs in cfDNA may emerge as an exciting tool for elucidating disease mechanisms and improving diagnostics (Liu 2022). In addition to genetic alterations, epigenetic modifications in cfDNA are also believed to be valuable as diagnostic biomarkers for many cancer types. One of the most reliable epigenetic markers is DNA methylation, which is achieved by adding a methyl group to the fifth carbon of cytosine using DNA methyltransferases (DNMTs) (Spector et al. 2023).

5.1.2 Computational obstacles while examining cfDNA methylation data

Computational analysis of cell-free DNA methylation data comprises some basic steps, which include pre-processing, quality check of reads, alignment, statistical analysis, and output interpretation. However, the precise method may differ based on the research objectives and the specific form of cfDNA (e.g., plasma-derived vs. tissue-specific cfDNA), with particular parts being customized to concentrate on either global methylation patterns or specific genomic regions. A variety of programs, including FastQC, NGS QC, QC–Chain, and ClinQC, have been created to do quality analysis (Patel & Jain 2012; Pandey et al. 2016). After the raw data has been analyzed, programs like Trim Galore may be used to eliminate low-quality bases and adapters. Wild card (GSNAP & BSMAP) and three-letter (BisMark, BS-Seeker2, & BRAT-BW) are some of the approaches used for aligning raw data to the reference genome (Zhou et al. 2019; Xi & Li 2009; Krueger & Andrews 2011; Guo et al. 2013). The wild card technique makes it possible to map both Cs and Ts of reads to Cs in the reference genome. On the other hand, the three-letter approach converts all Cs of reference and reads into Ts, which enables traditional alignment tools to be utilized. Data visualization tools like UCSC Genome Browser (Haeussler et al. 2019) , DNMIVD (Ding et al. 2020), Methylation plotter (Mallona et al. 2014) , and Integrative Genomics Viewer (IGV) (Thorvaldsdóttir et al. 2013) can be used to examine global methylation profiles. Several current DNA methylation calling applications (e.g., MeDUSA, RnBeads, Batman, MEDME) make use of various statistical models to measure DNA methylation coverage (Müller et al. 2019; Wilson et al. 2012; Pelizzola et al. 2008). Before making any decisions, it is important to take sequencing depth into account, which varies depending on the assay used.

For the early detection of numerous cancer types, a few large-scale controlled clinical trials are currently in progress. Preliminary findings from these extensive investigations suggest a poor level of accuracy in identifying stage I (18%) & stage II (43%) cancer, with a specificity of around 0.7% (Ofman et al. 2020) . The limited sensitivity in detecting cancer at an early stage emphasizes the need to carefully examine the different stages of cfDNA methylation studies. In our investigation, we have shown the utility of DNA methylation markers and their responsiveness in relation to the heterogeneity seen in tumors. In our analysis of cfDNA methylation patterns, we have also emphasized the advantages and limitations of the deconvolution and other machine learning techniques.

5.2 Materials and Methods

5.2.1 cfDNA methylation markers efficacy on TCGA cohort

To evaluate the suitability of cfDNA-based methylation biomarkers on 450k methylation bulk samples, the following methodology was used. Initially, we manually compiled a catalog of cfDNA methylation markers from the existing scientific literature. A brief description of the approach followed has been shown in Figure 5.1. We obtained 48 distinct cfDNA methylation markers for breast cancer using literature mining. The TCGA database provides publicly accessible molecular and medical data for multiple human cancers. This data includes the single nucleotide polymorphism (SNP), exome variant analysis, DNA methylation, microRNA (miRNA), transcriptome (mRNA), and proteome (Silva et al. 2016). From the TCGA database, an Illumina human methylation 450k array data set for bulk tissue was obtained. The Illumina human methylation 450k array is a high-resolution technology that quantifies methylation levels at about 450,000 CpG sites with single base precision (J. Wang et al. 2018). The TCGA 450k array methylation datafiles provide methylation beta values linked to various CpG sites. The beta value, which ranges from 0 (lower levels of methylation) to 1 (greater levels of methylation), indicates the proportion of the methylated array intensity and the array's total intensity. The TCGA dataset of 752 breast cancer and 96 non-cancerous samples was used to visualize the normalized methylation beta values for selected cfDNA marker promoters for differential analysis.

5.2.2 Classification of samples based on a single marker

Second, an investigation was performed on the diagnostic ability of the literature-derived breast cancer cfDNA methylation marker set on TCGA data. In order to highlight the effectiveness of using a single marker in classification, a comprehensive dataset consisting of 192 TCGA 450k methylation (bulk) samples was generated by random sampling. This dataset included 96 samples of breast cancer and 96 samples of normal tissue. Subsequently, biomarkers chosen at random from the TCGA data set were used to predict the cancer or non-cancer state of the samples. The sensitivity and False Positive Rate (FPR) values were derived from the LDA (Linear Discriminant Analysis) fitting and visualized as box plots. We take advantage of LDA, with its default configuration inside the MASS package (Venables & Ripley n.d.). The package

employs FDA (Fisher 1936) for the purpose of discrimination, and its extension, LDA (Radhakrishna Rao 2008), for the purpose of classification.

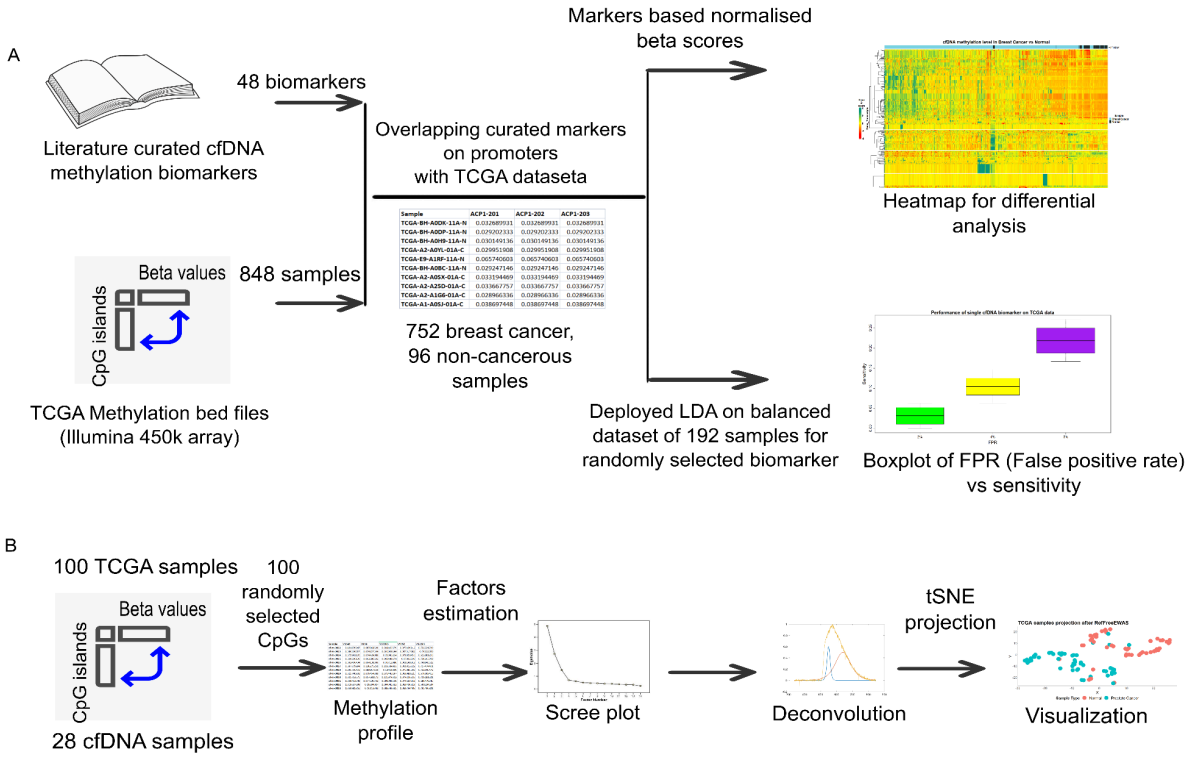

Figure 5.1: A brief description of the approach followed. **A)** Manually searching the literature yielded 48 breast cancer cfDNA methylation biomarkers. TCGA data set of both breast cancer as well as non-cancerous samples was utilized to illustrate normalized methylation beta values for chosen cfDNA marker promoters for differential analysis. To demonstrate the efficacy of a single marker in classification, biomarkers randomly selected from the balanced TCGA data set were utilized to predict sample cancer or non-cancer status. LDA (Linear Discriminant Analysis) fitting yielded box plots of sensitivity and FPR values. **B)** In order to understand how deconvolution algorithms may be used to distinguish between malignant and non-cancerous samples, we conducted a study to assess the effectiveness of three commonly used deconvolution techniques: RefFreeEWAS, ReFACTor, and SVA.

5.2.3 Implementation of deconvolution methods on cfDNA markers

To comprehend the use of deconvolution algorithms in discriminating cancerous samples from non-cancerous ones, we demonstrated the efficacy of three widely exploited deconvolution techniques: RefFreeEWAS (Houseman et al. 2014), ReFACTor (Rahmani et al. 2016), and SVA

(Leek & Storey 2007). This demonstration was conducted using bulk TCGA and cfDNA methylation data. First, using the TCGA database, 100 methylation samples (fifty prostate cancer and fifty normal) were queried for methylation beta values corresponding to 100 randomly chosen CpG islands. As the three techniques for deconvolution discussed earlier fall into the reference-free category, the estimate of the source was conducted using the scree plot. Based on the scree analysis, five components were selected to be kept. Afterward, tSNE was used on the normalized TCGA data and the corresponding deconvoluted matrices, with a perplexity of 5 and a maximum of 300 iterations. For the cfDNA 450k methylation samples (14 prostate cancer, 14 normal) that were taken from the CFEA database, a similar procedure was used (Yu et al. 2020). The estimated value for cell types was 3, and the tSNE parameters were specified with perplexity=3 and max_iter=60.

5.3.Results

5.3.1 Tumor heterogeneity and biomarkers dependency

The genetic, morphological, phenotypic and epigenetic variation among cell populations has culminated in the existence of inter as well as intra-tumor heterogeneity for decades. One of the main reasons for drug resistance and the failure of targeted therapies these days is cellular heterogeneity (Gay et al. 2016). Although studies that focus on whole cell populations may provide insights into the behavior of the majority of cells, they might obscure the significance of certain subpopulations and, therefore, the underlying basic biology. Moreover, the presence of cellular heterogeneity presents significant difficulties in diagnosing and treating diseases in studies that rely on data averaged over a population (Altschuler & Wu 2010) . Although tissue biopsies may only captivate a portion of this variability, liquid biopsies seem to be more advantageous in this situation (Castro-Giner et al. 2018).

To evaluate the consistency of several established cfDNA methylation markers from the literature, we examined their expression in a collection of 848 TCGA samples, which included 96 normal in addition to 752 breast cancer patients (Figure 5.2). It was discovered that the biomarkers' heterogeneity was significant enough to impede the design of therapeutics and diagnostics. Some confounding variables may also contribute to the observed variability, in addition to the markers utilized for disease detection.

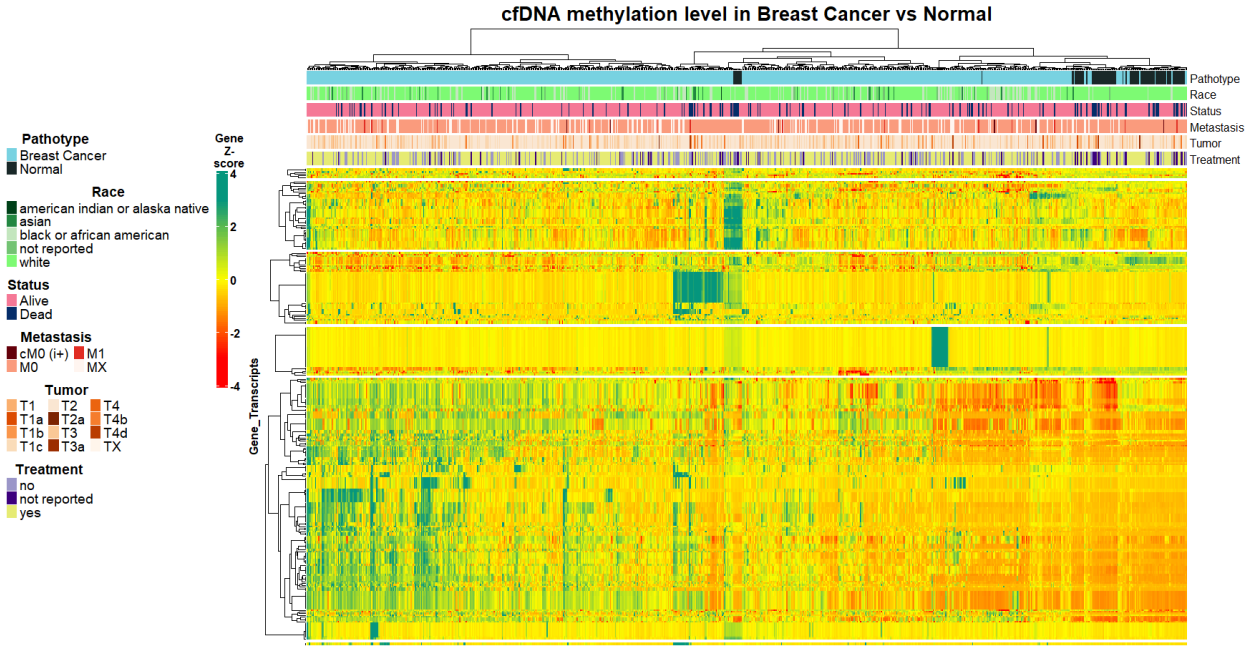

Figure 5.2: This heatmap illustrates the variability that exists across biomarkers across a certain particular type of cancer. With the purpose of conducting a differential analysis between non-cancerous and cancerous samples, a heatmap was created to visualize the normalized beta scores based on markers for all of the TCGA observations. One of the elements that may have an impact on disease diagnoses and therapies is the amount of heterogeneity that exists across biomarkers, as seen by this figure.

Tumour heterogeneity is a significant factor contributing to treatment failure, therapeutic resistance, and low survival rates in cancer patients. Nevertheless, the treatment of disease is hindered by the fact that the expected biomarkers, which are located on a non-uniform scale, are influenced by the dynamic nature of tumor cells (Russano et al. 2020) . The literature provides several examples of the heterogeneous character of druggable targets, such as stomach adenocarcinoma, lung adenocarcinoma, breast cancer, melanoma, and so on. As a result, using biomarker-based targeted therapy for heterogeneous malignancies ultimately results in recurrence (Ramón Y Cajal et al. 2020). Several diverse computational pipelines and techniques are now being developed to estimate cellular heterogeneity as a pre-processing step, aiming to get more useful insights (Huan et al. 2018; Scherer et al. 2020; Kim et al. 2016) .

5.3.2 Potential and difficulties in single marker based diagnosis

Despite being recognized for over a decade, the significance of aberrant cfDNA methylation levels in cancer as a diagnostic tool in clinical practice has not been clearly established. One major limitation of traditional biomarkers is that they are often only useful for detecting metastatic and advanced-stage cancer (Uehiro et al. 2016).

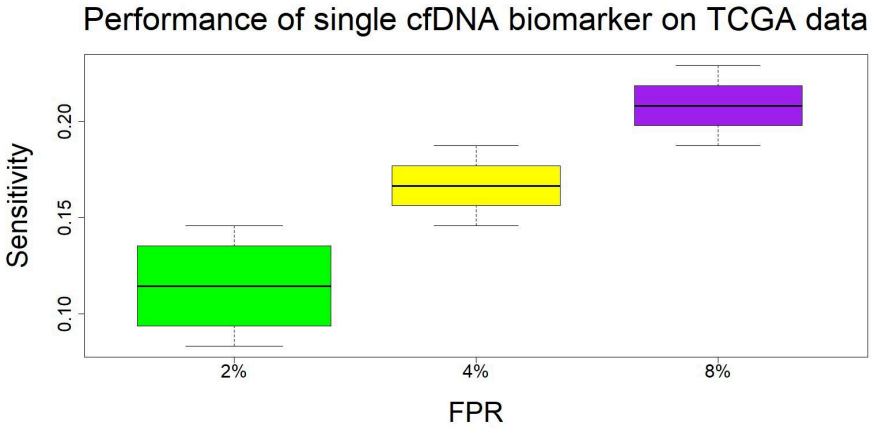

Figure 5.3: This boxplot illustrates the efficiency of a single biomarker for sample label prediction (breast cancer or non-cancerous) by comparing the false positive rate (FPR) to the sensitivity measures. These values were acquired based on the LDA (Linear Discriminant Analysis) fitting of TCGA data for one single marker. A single marker based technique for disease identification is shown to have a much lower level of sensitivity, as appears from the figure.

It is also evident from the box plot in Figure 5.3 that the concept of employing a single marker based technique for detecting diseases does not appear to ensure an appropriate degree of sensitivity when implemented in a classification model of the 192 TCGA illumina 450k methylation samples containing 96 normal and 96 breast cancer patients. It is possible that the potency of one single marker is not entirely capable of separating the malignant state from the non cancerous condition because of the small quantity of cfDNA that is generated. On the other hand, the sensitivity may be improved by using a collection involving multiple markers.

According to research by Barault *et al*., the prevalence of individual biomarkers in patients is rather modest, but it may rise when combined (Barault et al. 2018). Even if each of these indicators is relevant on its own, the multiparametric situation may help it distinguish between people with cancer and those who are healthy. When Mouliere *et al*., investigated the use of multiple markers for cfDNA in colorectal cancer (Intplex), they discovered that it was highly

specific, sensitive, and simple to operate. Furthermore, it has been shown that this approach is capable of being adjusted for repeated testing, hence simplifying further investigations in the context of personalized medicine (Mouliere et al. 2014). Nevertheless, there seem to be some drawbacks of utilizing a multi-marker panel. First and foremost, the efficacy of markers is contingent upon the population, experimental assay, and result analysis. These factors make doctors less confident in these biomarker panels. Furthermore, studies conducted to demonstrate the reliability of cfDNA markers are often retroactive and suffer from insufficient sample size and statistical expertise. To prevent such irregularities, it is necessary to conduct thorough investigations that adhere to the established rules for reporting diagnostic accuracy (Salvi et al. 2016) .

5.3.3 Deconvolution using tissue-specific biomarkers

Given the significant variations seen across different tissues, it is recommended to use biomarkers that are particular to each tissue. It has also been discovered that tissue-specific markers are more uniform in nature when evaluating plasma DNA (Gai et al. 2018). A frequently used technique for determining the source of tumor tissue using cfDNA involves the utilization of the deconvolution algorithm. This method has the ability to separate the original signal amongst an amalgamation of signals. Deconvolution strategies may be categorized into two primary types: reference-based and reference-free. Reference based deconvolution techniques rely on supervised approaches that use differentially methylated regions (DMRs) particular to each cell type. In contrast, reference free approaches do not need distinct reference points for different cell types. Instead, they estimate the fraction of cells using unsupervised deconvolution methods (Titus et al. 2017). Constrained projection [CP] uses least square minimization and is one of the first and most extensively used reference dataset methods. Regarding reference free techniques, there are methods such as reducing undesirable variation (RUV) and non-negative matrix factorization (NMF) (Teschendorff & Zheng 2017). For cfDNA deconvolution, several other reference-based [CIBERSORT,EpiDISH] and reference-free [CDSeq, CellMix, TOAST, EWASher, RefFreeEWAS, SVA] methods have recently surfaced (Teschendorff et al. 2017; Newman et al. 2015; Li & Wu 2019; Houseman et al. 2014). Research indicates that the accuracy of disease prediction improves when tissue proportion parameters are taken into account, leading to a more comprehensive understanding of biological outcomes. Moss et al. found that

deconvolution utilizing just selected groups of important CpG sites yields higher resolution and lower noise than using the whole methylome, even with low DNA levels (Moss et al. 2018).

In order to assess the distinguishability of reference free deconvolution techniques, we used three widely employed techniques: RefFreeEWAS, ReFACTor, and SVA (Figure 5.4). These methods were applied to a dataset consisting of 450k methylation profiles from both prostate cancer and normal samples obtained from TCGA having 100 samples and cfDNA with 28 samples, respectively. The present study conducted a comparison of different deconvolution techniques on a sample of hundred randomly selected CpG sites. The results indicated that the effectiveness of a particular approach is influenced, to some extent, by the characteristics of the dataset. For instance, in case of sample from TCGA, RefFreeEWAS demonstrated superior classification performance compared to other methods. On the other hand, in the cfDNA dataset, both ReFACTor and RefFreeEWAS exhibited similar performance in terms of separation capabilities (Figure 5.4). Additional limitations include unaccountable covariates associated with CpG island methylation, limited datasets, and batch effects.

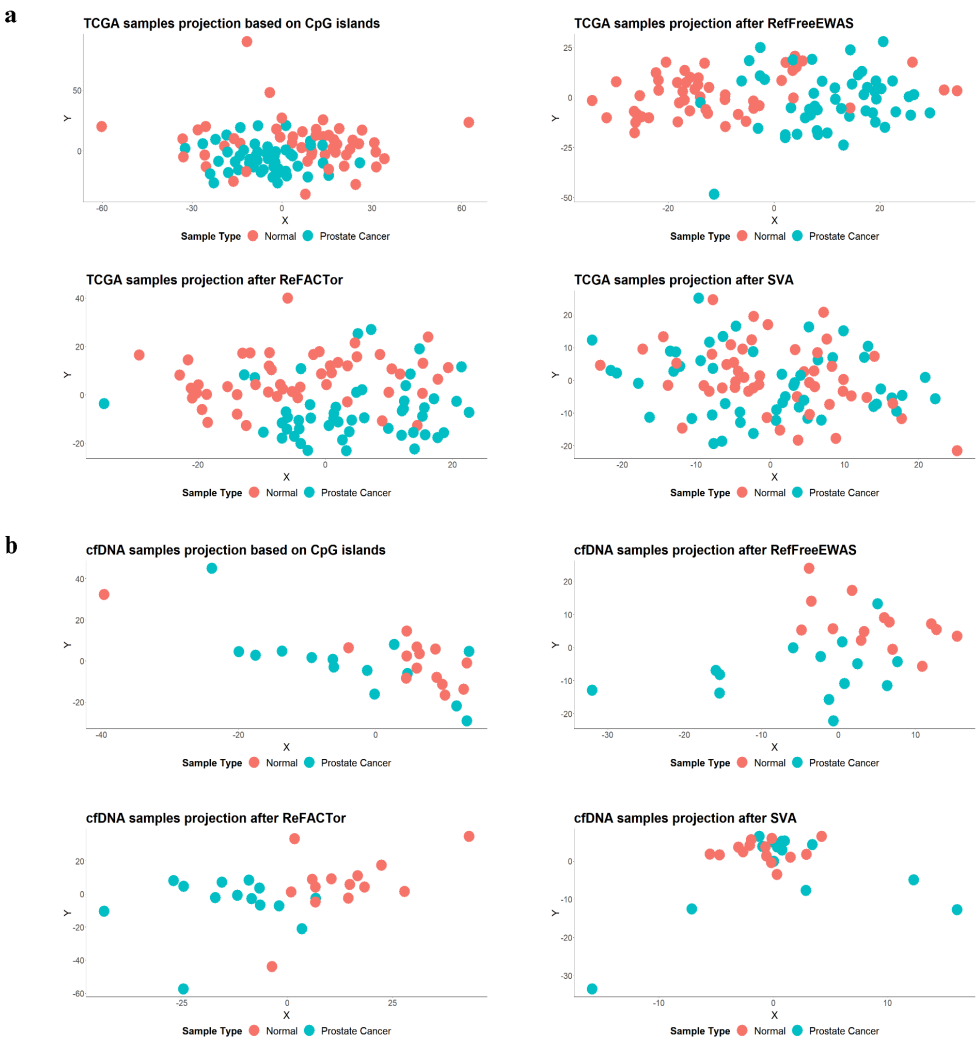

Figure 5.4: Implementing various deconvolution methods to analyze DNA methylation patterns of cancerous and non-cancerous samples. Deconvolution algorithms that do not need a reference, for example RefFreeEWAS, SVA, and ReFACTor, were used to evaluate DNA methylation profiles. The results were then projected on tSNE coordinates to assess the separability of the samples. **A)** The DNA methylation patterns from solid tumors of prostate cancer accessible in the TCGA consortium were used. **B)** Deconvolution techniques were used to analyze DNA methylation patterns of cfDNA obtained from the plasma of people with both prostate cancer and non-cancerous samples from CFEA database. The comparison study was conducted using a collection of 100 CpG sites that were randomly chosen.

5.4 Discussion

In this chapter, we outline the advantages and disadvantages of several methods used to identify cancer via the analysis of cfDNA methylation (Sharma, Verma, et al. 2022). Through the

examination of preexisting DNA methylation patterns obtained from tumour samples and cfDNA, we have identified constraints in the use of particular markers as a result of the presence of cancer heterogeneity. Moreover, there exists another kind of bias that further complicates the computing task. The presence of bias in various methods of DNA methylation detection diminishes the value of identifying certain markers. For example, the HM450k methylation array may identify several markers that may not be detectable at all, utilizing RRBS-based cfDNA methylation profiling. Although there are certain collections of cfDNA methylation profiles available for cancer patients, it is challenging to identify biomarkers that can be universally applied for any form of cancer using various cfDNA methylation profiling approaches. In other areas of genomics, such as the analysis of single-cell expression profiles, there have been a limited number of efforts to do integrative analysis regardless of the biases introduced by the platform and procedure utilised. Nevertheless, there have been few endeavors to address the computational challenge of integrating cfDNA methylation data for analysis. The discrepancy may arise from the fact that single-cell expression profiles consist of recognized cell types, while cfDNA methylation profiles include a combination of signals from a variety of cell types.

The method used by several clinical trials to train machine learning models on one dataset and verify them on a separate dataset is often referred to as transfer learning. Significant progress has been made in enhancing the adaptability of transfer learning (Cao et al. 2010) to fresh data sets in order to mitigate the batch effect. Nevertheless, adaptive transfer learning often requires a limited amount of data from the target domain in order to fine-tune its performance. There may be daily fluctuations in the analysis of cfDNA methylation, within the same patient. Therefore, it is yet unknown to what extent adaptive transfer learning may be used to determine the origin of tissue by analyzing cfDNA methylation, regardless of batch effect and variations caused by blood cells.

While certain clinical studies have shown high precision in identifying advanced-stage cancer, the diagnosis of cancer in its early stages remains a formidable task (J. Li et al. 2019; Uehiro et al. 2016). The limited precision of early cancer identification diminishes the effectiveness of liquid biopsy since advanced stage tumors are often untreatable. Therefore, there is a continuing need for innovative computational methods to enhance the identification of early stage cancer by utilizing cfDNA methylation patterns.

CHAPTER 6

CONCLUSION

Since single-gene based approaches overgeneralize the complex and interdependent structure of biological systems, they have usually failed to address disease complexity. Multifactorial interactions between genes, regulatory components, and environmental factors cause diseases, especially complicated ones. Focusing on individual genes ignores the complex network of relationships that underpins cellular activity and disease. This reductionist approach cannot completely understand biological systems' redundancy and robustness, where genes may compensate for each other and regulatory structures can adapt to disturbances. Thus, understanding and appropriately concentrating on diseases requires a comprehensive systems biology approach that considers genetic, epigenetic, and environmental components within the biological framework. Despite technological advances, several factors make it challenging to treat many diseases using a systems biology approach. Biological systems are complex and dynamic, with extensive regulatory networks that make modeling and understanding difficult. Modern bioinformatics applications and changing computational models are needed to integrate massive information from high-throughput technologies across several biological levels. Genetic variety, epigenetic, and environmental factors complicate diseases and make it difficult to identify universally effective treatments. The slow process of examining and authorization novel drugs and providing treatment hinders advancement in medicine. Economic constraints, specialized infrastructure, and transdisciplinary expertise limit systems biology's clinical use. These difficulties need advances in computational and integrative methods, specific medical techniques, infrastructure, and interdisciplinary collaboration. To overcome these limits, more advanced, integrated, and tailored diagnostic methods are needed to fully comprehend disease biology and enhance diagnosis and therapy.

6.1 Overview of the contribution

The purpose of this part is to offer a condensed summary of the chapters that convey the thesis altogether.

6.2 Using single-cell expression profiles to associate pathways to diseases and make inferences about potential drugs

The identification of direct dependencies across diseases and biological pathways has been the focus of many different kinds of academic investigations due to the fact that it offers a wide range of potential applications. However, in the past, such investigations have encountered obstacles owing to the heterogeneity of the cells and the constraints on the quantity of samples that are accessible for bulk expression profiles. In this thesis, we present a technique for doing single-cell expression-based inference of the relationship between pathway, disease, and cell type (sci-PDC), which may assist in understanding the source and effect of these factors and guiding precision medicine. Within a small number of diseases and pathways, our methodology brought to light reliable connections between them. Employing diabetes as a case study, we have proven the way sci-PDC may assist in monitoring variations in the association between pathways and diseases as a consequence of variations in age and species. Significant knowledge on the applicability of the mouse system to certain pathways in the context of diabetes was uncovered by comparing the pathways-disease relationships between humans and mice. Its dependability for multifaceted applications is shown by the consistency of our approach's outcomes with information from earlier research papers, including details regarding the drug targeted pathways.

6.3 Using tumor specific single-cell profiles to explore relationships between biological pathways and cancer hallmark properties for precision therapy

It is well recognized that cancer exhibits inter along with, intra-tumor heterogeneity as well as stemness, both of which generate therapy hurdles. In order to have a complete understanding of its complexity, it is necessary to recognize the distinguishing traits that distinguish malignant cells, also known as cancer hallmark properties. These traits have a close connection with certain

biochemical pathways, each of which is necessary for the development, expansion, and response to therapies for cancer. However, depending on the particular context for various types of cells, these hallmarks may manifest and behave differently. There are several cancer cell lines, each of which has its own distinct origins as well as distinct epigenetic and genetic variations that influence their behavior. Consequently, the relevance of cellular specificity emphasizes the necessity for customized therapies that concentrate on the different genetic vulnerabilities of tumor cells in order to deliver precision medicine. Despite the fact that typical scRNA-seq-based studies could only give a snapshot of such relationships, it is necessary to have an attribute that incorporates temporal dimension in order to comprehend their direction and poising. The purpose of this investigation is to provide a strategy that utilizes an amalgamation of scRNA-seq and RNA-velocity characterization in order to establish direct connections between cancer hallmark features and pathways. In addition to exposing hallmarks and pathways that are pertinent to cell-line-specific reliance, this method has the potential to enhance our understanding of the cell's drug response in a variety of cancer microenvironments.

6.4 Utilising gene ensembles defined by chromatin domains with transcriptomics to understand drug-response and phenotypic heterogeneity in cancer cells

A comprehensive investigation is required to determine the impact of co-localizing genes within TADs along with their function as an independent regulatory component in cancer tissues,along with drug responsiveness. In order to comprehend the relationship between the TAD activity and drug response, we examined their activity utilizing cancer-cell transcriptomes in conjunction with chromatin interaction as well as epigenomic profiles. After doing an examination of the transcriptomes of 819 cancerous cell lines, it was shown that the response of these cells to different drugs was more closely associated with activity of distinct TADs rather than with genes. When we applied our methodology to the data of 9014 cancer patients, of whom there were twenty distinct forms of cancer, we discovered that there was a stronger association across survival and the activity of numerous individual TADs than there was among their genes. The increased usefulness of TAD-activity-based analysis is shown by its identification of

primate-specific synteny gain inside a TAD, including the EGFR gene, and its contribution towards cancer malignancy.

6.5 Introspecting computational challenges associated with cell-free cancer diagnosis using ensemble of DNA methylation markers

cfDNA methylation profiling is regarded as a novel and perhaps sustainable technique for liquid biopsy, which may be used to investigate disease progression and establish precise and uniform diagnosing and prognosis markers. Multiple processes are accountable for the shedding of cfDNA in the plasma of the blood, which subsequently enables it to provide insights into the dynamic alterations occurring in the human body. The study of cfDNA for cancer detection has various hurdles, including its fragmented nature, small quantity, and significant noise in the background. The examination of the methylation pattern in cfDNA is made more difficult by factors like heterogeneity, marker sensitivity, technological biases, as well as batch effects. The present study examines the source of cfDNA methylation, its characterization, and the computational difficulties involved in analyzing it for diagnostic purposes. Here, we also consider the use of a multi-marker strategy to address the issue of cancer heterogeneity. In addition, we provide a comprehensive evaluation of deconvolution techniques and machine learning methods for the investigation of cfDNA methylation. Our method demonstrates the possibility of enhancing analytical techniques for early cancer detection by using cfDNA methylation.

6.6 Future directions

The future prospects for using gene ensembles in disease diagnostics and therapies have the potential to completely transform personalized medicine. It should be noted that the application of these ensembles may be contingent upon the specific context. Even though they can offer valuable insights into specific biological processes, their generalizability and robustness may be restricted across various cell types or conditions. Sometimes these ensembles don't convey the intricacy of gene interactions or the dynamic nature of gene expression in varied biological conditions. Moreover, the techniques employed to construct gene ensembles may include biases that influence the accuracy as well as interpretation of these gene ensembles. Thus, while gene

ensembles are compelling their relevance and correctness must be carefully assessed, especially when applied to biological systems beyond their original environment.

Recent developments in genomic technology, such as next-generation sequencing and CRISPR-based gene editing, allow for the discovery of distinct collections of genes that are linked to certain conditions. Within the field of therapeutics, comprehending gene connections and pathways may result in precise therapies that are customized to an individual's genetic makeup, hence minimizing adverse effects and enhancing effectiveness. Integrating gene ensemble data alongside additional omics data, including metabolomics and proteomics, may provide a comprehensive understanding of the underlying processes contributing to the disease. This can facilitate the creation of innovative treatment approaches and the repurposing of drugs. With the ongoing advancement of bioinformatics and computational biology, the evaluation and interpretation of intricate collections of genes shall grow more precise. This will greatly improve the capacity to forecast the course of diseases and anticipate how they will respond to therapy. This comprehensive approach has the potential to revolutionize healthcare by providing more accurate, anticipatory, and personalized therapies.

REFERENCES

Abbas, M. & El-Manzalawy, Y., 2020. Machine learning based refined differential gene expression analysis of pediatric sepsis. *BMC medical genomics*, 13(1), p.122.

Abderrahmani, A. et al., 2000. Genetic variation in the hepatocyte nuclear factor-3beta gene (HNF3B) does not contribute to maturity-onset diabetes of the young in French Caucasians. *Diabetes*, 49(2), pp.306–308.

Abu-Serie, M.M. & Abdelfattah, E.Z.A., 2023. A comparative study of smart nanoformulations of diethyldithiocarbamate with Cu4O3 nanoparticles or zinc oxide nanoparticles for efficient eradication of metastatic breast cancer. *Scientific reports*, 13(1), p.3529.

Adewale, Q. et al., 2024. Single-nucleus RNA velocity reveals critical synaptic and cell-cycle dysregulations in neuropathologically confirmed Alzheimer's disease. *Scientific reports*, 14(1), p.7269.

Adriaens, M.E. et al., 2008. The public road to high-quality curated biological pathways. *Drug discovery today*, 13(19-20), pp.856–862.

Agrawal, A. et al., 2024. WikiPathways 2024: next generation pathway database. *Nucleic acids research*, 52(D1), pp.D679–D689.

Aibar, S. et al., 2017. SCENIC: single-cell regulatory network inference and clustering. *Nature methods*, 14(11), pp.1083–1086.

Aissa, A.F. et al., 2021. Single-cell transcriptional changes associated with drug tolerance and response to combination therapies in cancer. *Nature communications*, 12(1), p.1628.

Akdemir, K.C. et al., 2020. Disruption of chromatin folding domains by somatic genomic rearrangements in human cancer. *Nature genetics*, 52(3), pp.294–305.

Alamri, B.N. et al., 2016. The role of ghrelin in the regulation of glucose homeostasis. *Hormone molecular biology and clinical investigation*, 26(1), pp.3–11.

Al-Awar, A. et al., 2016. Experimental Diabetes Mellitus in Different Animal Models. *Journal of diabetes research*, 2016, p.9051426.

Al-Hakeim, H.K. & Alhillawi, Z.H., 2018. Effect of serum fibroblast growth factor receptor 2 and CAPS proteins on calcium status in β-thalassaemia major patients who are free from overt inflammation. *Growth factors* , 36(3-4), pp.178–185.

Altschuler, S.J. & Wu, L.F., 2010. Cellular heterogeneity: do differences make a difference? *Cell*, 141(4), pp.559–563.

Alves, F. et al., 2001. Inhibitory effect of a matrix metalloproteinase inhibitor on growth and spread of human pancreatic ductal adenocarcinoma evaluated in an orthotopic severe combined immunodeficient (SCID) mouse model. *Cancer letters*, 165(2), pp.161–170.

Anandhan, A. et al., 2017. Metabolic Dysfunction in Parkinson's Disease: Bioenergetics, Redox Homeostasis and Central Carbon Metabolism. *Brain research bulletin*, 133, pp.12–30.

Andrews, S.J. et al., 2023. The complex genetic architecture of Alzheimer's disease: novel insights and future directions. *EBioMedicine*, 90, p.104511.

Angione, C., 2019. Human Systems Biology and Metabolic Modelling: A Review-From Disease Metabolism to Precision Medicine. *BioMed research international*, 2019, p.8304260.

Anon, A case of Tardive Dyskinesia and Parkinsonism Following Use of phentermine for Weight Loss. Available at: https://neurology.ufl.edu/2018/04/18/a-case-of-tardive-dyskinesia-and-parkinsonism-following-use-of-phentermine-for-weight-loss/ [Accessed June 15, 2024a].

Anon, Ingenuity Pathways Analysis (IPA). Available at: https://www.nihlibrary.nih.gov/resources/tools/ingenuity-pathways-analysis-ipa [Accessed July 3, 2024b].

Araki, H. et al., 2012. GeneSetDB: A comprehensive meta-database, statistical and visualisation framework for gene set analysis. *FEBS open bio*, 2, pp.76–82.

Ariaans, G. et al., 2017. Anti-tumor effects of everolimus and metformin are complementary and glucose-dependent in breast cancer cells. *BMC cancer*, 17(1), p.232.

Armingol, E. et al., 2021. Deciphering cell-cell interactions and communication from gene expression. *Nature reviews. Genetics*, 22(2), pp.71–88.

Asati, V., Mahapatra, D.K. & Bharti, S.K., 2016. PI3K/Akt/mTOR and Ras/Raf/MEK/ERK signaling pathways inhibitors as anticancer agents: Structural and pharmacological perspectives. *European journal of medicinal chemistry*, 109, pp.314–341.

Ashburner, M. et al., 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, 25(1), pp.25–29.

Atkins, A. et al., 2021. Off-target analysis in gene editing and applications for clinical translation of CRISPR/Cas9 in HIV-1 therapy. *Frontiers in genome editing*, 3, p.673022.

Aucamp, J. et al., 2018. The diverse origins of circulating cell-free DNA in the human body: a critical re-evaluation of the literature. *Biological reviews of the Cambridge Philosophical Society*, 93(3), pp.1649–1683.

Bailey, S.M., 2023. Editorial: Hallmark of cancer: replicative immortality. *Frontiers in oncology*, 13, p.1204094.

Balaban, S. et al., 2018. Heterogeneity of fatty acid metabolism in breast cancer cells underlies differential sensitivity to palmitate-induced apoptosis. *Molecular oncology*, 12(9), pp.1623–1638.

Bannister, A.J. & Kouzarides, T., 2011. Regulation of chromatin by histone modifications. *Cell research*, 21(3), pp.381–395.

Barault, L. et al., 2018. Discovery of methylated circulating DNA biomarkers for comprehensive non-invasive monitoring of treatment response in metastatic colorectal cancer. *Gut*, 67(11), pp.1995–2005.

Barretina, J. et al., 2012. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391), pp.603–607.

Barutcu, A.R. et al., 2018. A TAD boundary is preserved upon deletion of the CTCF-rich Firre locus. *Nature communications*, 9(1), p.1444.

Basu, A. et al., 2013. An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. *Cell*, 154(5), pp.1151–1161.

Belton, J.-M. et al., 2012. Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* , 58(3), pp.268–276.

Benjamini, Y. & Hochberg, Y., 2018. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 57(1), pp.289–300.

Bergen, V. et al., 2020. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nature biotechnology*, 38(12), pp.1408–1414.

Bergen, V. et al., 2021. RNA velocity-current challenges and future perspectives. *Molecular systems biology*, 17(8), p.e10282.

Blanchet, E. et al., 2015. Feedback inhibition of CREB signaling promotes beta cell dysfunction in insulin resistance. *Cell reports*, 10(7), pp.1149–1157.

Blockhuys, S. & Wittung-Stafshede, P., 2017. Copper chaperone Atox1 plays role in breast cancer cell migration. *Biochemical and biophysical research communications*, 483(1), pp.301–304.

Blüthgen, N. et al., 2005. Biological profiling of gene groups utilizing Gene Ontology. *Genome informatics. International Conference on Genome Informatics*, 16(1), pp.106–115.

Bobadilla, J.L. et al., 2002. Cystic fibrosis: a worldwide analysis of CFTR mutations--correlation with incidence data and application to screening. *Human mutation*, 19(6), pp.575–606.

Bronkhorst, A.J., Ungerer, V. & Holdenrieder, S., 2019. The emerging role of cell-free DNA as a molecular marker for cancer management. *Biomolecular detection and quantification*, 17,

p.100087.

Buphamalai, P. et al., 2021. Network analysis reveals rare disease signatures across multiple levels of biological organization. *Nature communications*, 12(1), p.6306.

Cancer Genome Atlas Research Network et al., 2013. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature genetics*, 45(10), pp.1113–1120.

Cantley, J. & Ashcroft, F.M., 2015. Q&A: insulin secretion and type 2 diabetes: why do β-cells fail? *BMC biology*, 13(1), p.33.

Cao, B. et al., 2010. Adaptive Transfer Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 24(1), pp.407–412.

Carlice-Dos-Reis, T. et al., 2017. Investigation of mutations in the HBB gene using the 1,000 genomes database. *PloS one*, 12(4), p.e0174637.

Carthew, R.W., 2021. Gene Regulation and Cellular Metabolism: An Essential Partnership. *Trends in genetics: TIG*, 37(4), pp.389–400.

Castro-Giner, F. et al., 2018. Cancer Diagnosis Using a Liquid Biopsy: Challenges and Expectations. *Diagnostics (Basel, Switzerland)*, 8(2). Available at: http://dx.doi.org/10.3390/diagnostics8020031.

Cefalu, W.T., 2006. Animal models of type 2 diabetes: clinical presentation and pathophysiological relevance to the human condition. *ILAR journal / National Research Council, Institute of Laboratory Animal Resources*, 47(3), pp.186–198.

Chandra, O. et al., 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*, p.2024.05.06.592622. Available at: https://www.biorxiv.org/content/biorxiv/early/2024/05/08/2024.05.06.592622 [Accessed August 31, 2024].

Chandra, O. et al., 2023. 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, pp.3590–3603.

Chang, H. et al., 2021. Lactate secreted by PKM2 upregulation promotes Galectin-9-mediated immunosuppression via inhibiting NF-κB pathway in HNSCC. *Cell death & disease*, 12(8), p.725.

Chang, J., Chapman, B., Friedberg, I., Hamelryck, T., de Hoon, M., Cock, P., Antao, T., Talevich, E. and Wilczynski, B., 2010. Biopython tutorial and cookbook. *Update*, pp.15-19.

Chawla, S. et al., 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), p.e13.

Chen, C., Ge, Y. & Lu, L., 2023. Opportunities and challenges in the application of single-cell and spatial transcriptomics in plants. *Frontiers in plant science*, 14, p.1185377.

Chen, E.Y. et al., 2013. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC bioinformatics*, 14, p.128.

Chen, S. et al., 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* , 34(17), pp.i884–i890.

Chentli, F., Azzoug, S. & Mahgoun, S., 2015. Diabetes mellitus in elderly. *Indian journal of endocrinology and metabolism*, 19(6), pp.744–752.

Chen, Y., Verbeek, F.J. & Wolstencroft, K., 2021. Establishing a consensus for the hallmarks of cancer based on gene ontology and pathway annotations. *BMC bioinformatics*, 22(1), p.178.

Chia, C.W., Egan, J.M. & Ferrucci, L., 2018. Age-Related Changes in Glucose Metabolism, Hyperglycemia, and Cardiovascular Risk. *Circulation research*, 123(7), pp.886–904.

Chiang, C.-C. et al., 2023. Transcriptome analysis creates a new era of precision medicine for managing recurrent hepatocellular carcinoma. *World journal of gastroenterology: WJG*, 29(5), pp.780–799.

Chia, S. et al., 2017. Phenotype-driven precision oncology as a guide for clinical decisions one patient at a time. *Nature communications*, 8(1), p.435.

Chiou, K.L. & Bergey, C.M., 2018. Methylation-based enrichment facilitates low-cost, noninvasive genomic scale sequencing of populations from feces. *Scientific reports*, 8(1), p.1975.

Cho, I.-S. et al., 2012. Deregulation of CREB signaling pathway induced by chronic hyperglycemia downregulates NeuroD transcription. *PloS one*, 7(4), p.e34860.

Choi, Y.H. & Kim, J.K., 2019. Dissecting Cellular Heterogeneity Using Single-Cell RNA Sequencing. *Molecules and cells*, 42(3), pp.189–199.

Ciccarone, F. & Ciriolo, M.R., 2023. Editorial: Hallmark of cancer: sustained proliferative signalling. *Frontiers in oncology*, 13, p.1328827.

Cohen, K. et al., 2015. Pharmacological treatment of diabetic peripheral neuropathy. *P & T: a peer-reviewed journal for formulary management*, 40(6), pp.372–388.

Conner, S.J. et al., 2024. Cell morphology best predicts tumorigenicity and metastasis in vivo across multiple TNBC cell lines of different metastatic potential. *Breast cancer research: BCR*, 26(1), p.43.

Cordell, H.J., 2009. Detecting gene-gene interactions that underlie human diseases. *Nature reviews. Genetics*, 10(6), pp.392–404.

Cosentino, G., Conrad, A.O. & Uwaifo, G.I., 2013. Phentermine and topiramate for the management of obesity: a review. *Drug design, development and therapy*, 7, pp.267–278.

Crick, F., 1970. Central dogma of molecular biology. *Nature*, 227(5258), pp.561–563.

Croft, D. et al., 2011. Reactome: a database of reactions, pathways and biological processes. *Nucleic acids research*, 39(Database issue), pp.D691–7.

Cubeñas-Potts, C. & Corces, V.G., 2015. Architectural proteins, transcription, and the three-dimensional organization of the genome. *FEBS letters*, 589(20 Pt A), pp.2923–2930.

D'Agostino, N., Li, W. & Wang, D., 2022. High-throughput transcriptomics. *Scientific reports*, 12(1), p.20313.

Damian, D. and Gorfine, M., 2004. Statistical concerns about the GSEA procedure. *Nature genetics*, 36(7), pp.663-663.

van Dam, S. et al., 2018. Gene co-expression analysis for functional classification and gene-disease predictions. *Briefings in bioinformatics*, 19(4), pp.575–592.

Danecek, P. et al., 2021. Twelve years of SAMtools and BCFtools. *GigaScience*, 10(2). Available at: http://dx.doi.org/10.1093/gigascience/giab008.

Das, S., McClain, C.J. & Rai, S.N., 2020. Fifteen Years of Gene Set Analysis for High-Throughput Genomic Data: A Review of Statistical Approaches and Future Challenges. *Entropy* , 22(4). Available at: http://dx.doi.org/10.3390/e22040427.

Dean, M., Fojo, T. & Bates, S., 2005. Tumour stem cells and drug resistance. *Nature reviews. Cancer*, 5(4), pp.275–284.

Dewanjee, S. et al., 2021. The Emerging Role of HDACs: Pathology and Therapeutic Targets in Diabetes Mellitus. *Cells* , 10(6). Available at: http://dx.doi.org/10.3390/cells10061340.

Dhar, G.A. et al., 2021. DNA methylation and regulation of gene expression: Guardian of our health. *The Nucleus/ an international journal of cytology and allied topics*, 64(3), pp.259–270.

Dhawan, B.N. et al., 1996. International Union of Pharmacology. XII. Classification of opioid receptors. *Pharmacological reviews*, 48(4), pp.567–592.

Ding, W. et al., 2020. DNMIVD: DNA methylation interactive visualization database. *Nucleic acids research*, 48(D1), pp.D856–D862.

Dixon, J.R. et al., 2015. Chromatin architecture reorganization during stem cell differentiation. *Nature*, 518(7539), pp.331–336.

Dixon, J.R. et al., 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398), pp.376–380.

Dobin, A. et al., 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* , 29(1), pp.15–21.

Donnenberg, V.S. & Donnenberg, A.D., 2005. Multiple drug resistance in cancer revisited: the cancer stem cell hypothesis. *Journal of clinical pharmacology*, 45(8), pp.872–877.

Dorrity, M.W. et al., 2020. Dimensionality reduction by UMAP to visualize physical and genetic interactions. *Nature communications*, 11(1), p.1537.

Dostie, J. et al., 2006. Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome research*, 16(10), pp.1299–1309.

Durand, N.C. et al., 2016. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell systems*, 3(1), pp.95–98.

Emmert-Streib, F. & Glazko, G.V., 2011. Pathway analysis of expression data: deciphering functional building blocks of complex diseases. *PLoS computational biology*, 7(5), p.e1002053.

Enge, M. et al., 2017. Single-Cell Analysis of Human Pancreas Reveals Transcriptional Signatures of Aging and Somatic Mutation Patterns. *Cell*, 171(2), pp.321–330.e14.

Erler, J.T. et al., 2009. Hypoxia-induced lysyl oxidase is a critical mediator of bone marrow cell recruitment to form the premetastatic niche. *Cancer cell*, 15(1), pp.35–44.

Erwin, A. et al., 2011. Intrathecal baclofen in multiple sclerosis: too little, too late? *Multiple sclerosis* , 17(5), pp.623–629.

Fan, J. et al., 2016. Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nature methods*, 13(3), pp.241–244.

Fares, J. et al., 2020. Molecular principles of metastasis: a hallmark of cancer revisited. *Signal transduction and targeted therapy*, 5(1), p.28.

Faubert, B. et al., 2013. AMPK is a negative regulator of the Warburg effect and suppresses tumor growth in vivo. *Cell metabolism*, 17(1), pp.113–124.

Feizi, S. et al., 2013. Network deconvolution as a general method to distinguish direct dependencies in networks. *Nature biotechnology*, 31(8), pp.726–733.

Feng, H., Jin, P. & Wu, H., 2019. Disease prediction by cell-free DNA methylation. *Briefings in bioinformatics*, 20(2), pp.585–597.

Ferrara, N., 2002. VEGF and the quest for tumour angiogenesis factors. *Nature reviews. Cancer*, 2(10), pp.795–803.

Fisher, R.A., 1936. The use of multiple measurements in taxonomic problems. *Annals of*

eugenics, 7(2), pp.179–188.

Friedman, N., 2004. Inferring cellular networks using probabilistic graphical models. *Science*, 303(5659), pp.799–805.

Fu, Z., Gilbert, E.R. & Liu, D., 2013. Regulation of insulin synthesis and secretion and pancreatic Beta-cell dysfunction in diabetes. *Current diabetes reviews*, 9(1), pp.25–53.

Gai, W. et al., 2018. Liver- and Colon-Specific DNA Methylation Markers in Plasma for Investigation of Colorectal Cancers with or without Liver Metastases. *Clinical chemistry*, 64(8), pp.1239–1249.

Galupa, R. & Heard, E., 2017. Topologically Associating Domains in Chromosome Architecture and Gene Regulatory Landscapes during Development, Disease, and Evolution. *Cold Spring Harbor symposia on quantitative biology*, 82, pp.267–278.

Gambardella, G. et al., 2022. A single-cell analysis of breast cancer cell lines to study tumour heterogeneity and drug response. *Nature communications*, 13(1), p.1714.

Gao, M., Qiao, C. & Huang, Y., 2022. UniTVelo: temporally unified RNA velocity reinforces single-cell trajectory inference. *Nature communications*, 13(1), p.6586.

García-Cortés, D. et al., 2020. Gene Co-expression Is Distance-Dependent in Breast Cancer. *Frontiers in oncology*, 10, p.1232.

Garcia-Moreno, A. et al., 2022. Functional Enrichment Analysis of Regulatory Elements. *Biomedicines*, 10(3). Available at: http://dx.doi.org/10.3390/biomedicines10030590.

Gay, L., Baker, A.-M. & Graham, T.A., 2016. Tumour Cell Heterogeneity. *F1000Research*, 5. Available at: http://dx.doi.org/10.12688/f1000research.7210.1.

Gest, C. et al., 2013. Rac3 induces a molecular pathway triggering breast cancer cell aggressiveness: differences in MDA-MB-231 and MCF-7 breast cancer cell lines. *BMC cancer*, 13, p.63.

Gilbert-Diamond, D. & Moore, J.H., 2011. Analysis of gene-gene interactions. *Current protocols in human genetics / editorial board, Jonathan L. Haines ... [et al.]*, Chapter 1, p.Unit1.14.

Gilbert, S.F., 2000. The developmental mechanics of cell specification. *Developmental Biology*, *340*, p.341.

Glaab, E. et al., 2012. EnrichNet: network-based gene set enrichment analysis. *Bioinformatics* , 28(18), pp.i451–i457.

Glasspool, R.M., Teodoridis, J.M. & Brown, R., 2006. Epigenetics as a mechanism driving polygenic clinical drug resistance. *British journal of cancer*, 94(8), pp.1087–1092.

Goeman, J.J. & Bühlmann, P., 2007. Analyzing gene expression data in terms of gene sets:

methodological issues. *Bioinformatics* , 23(8), pp.980–987.

Goldman, M.J. et al., 2020. Visualizing and interpreting cancer genomics data via the Xena platform. *Nature biotechnology*, 38(6), pp.675–678.

Goldman, S.L. et al., 2019. The Impact of Heterogeneity on Single-Cell Sequencing. *Frontiers in genetics*, 10, p.8.

Goswami, C.P. et al., 2015. A New Drug Combinatory Effect Prediction Algorithm on the Cancer Cell Based on Gene Expression and Dose-Response Curve. *CPT: pharmacometrics & systems pharmacology*, 4(2), p.e9.

Govindarajan, R. et al., 2012. Microarray and its applications. *Journal of pharmacy & bioallied sciences*, 4(Suppl 2), pp.S310–2.

Grabuschnig, S. et al., 2020. Putative Origins of Cell-Free DNA in Humans: A Review of Active and Passive Nucleic Acid Release Mechanisms. *International journal of molecular sciences*, 21(21). Available at: http://dx.doi.org/10.3390/ijms21218062.

Groenewoud, D., Shye, A. & Elkon, R., 2022. Incorporating regulatory interactions into gene-set analyses for GWAS data: A controlled analysis with the MAGMA tool. *PLoS computational biology*, 18(3), p.e1009908.

Grossman, R.L. et al., 2016. Toward a Shared Vision for Cancer Genomic Data. *The New England journal of medicine*, 375(12), pp.1109–1112.

Guan, J., Lin, Y. & Ji, G., 2020. Cell Type-Specific Gene Network-Based Analysis Depicts the Heterogeneity of Autism Spectrum Disorder. *Frontiers in cellular neuroscience*, 14, p.59.

Gui, H. et al., 2011. Comparisons of seven algorithms for pathway analysis using the WTCCC Crohn's Disease dataset. *BMC research notes*, 4, p.386.

Guo, W. et al., 2013. BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data. *BMC genomics*, 14, p.774.

Gupta, M.K. et al., 2023. The role of DNA methylation in personalized medicine for immune-related diseases. *Pharmacology & therapeutics*, 250, p.108508.

Gurule, N.J. et al., 2021. A tyrosine kinase inhibitor-induced interferon response positively associates with clinical response in EGFR-mutant lung cancer. *NPJ precision oncology*, 5(1), p.41.

Haeussler, M. et al., 2019. The UCSC Genome Browser database: 2019 update. *Nucleic acids research*, 47(D1), pp.D853–D858.

Hanahan, D., 2022. Hallmarks of Cancer: New Dimensions. *Cancer discovery*, 12(1), pp.31–46.

Hanahan, D. & Weinberg, R.A., 2011. Hallmarks of cancer: the next generation. *Cell*, 144(5),

pp.646–674.

Hanahan, D. & Weinberg, R.A., 2000. The hallmarks of cancer. *Cell*, 100(1), pp.57–70.

Han, J., Zhang, Z. & Wang, K., 2018. 3C and 3C-based techniques: the powerful tools for spatial genome organization deciphering. *Molecular cytogenetics*, 11, p.21.

Han, X. et al., 2020. Construction of a human cell landscape at single-cell level. *Nature*, 581(7808), pp.303–309.

Han, X. et al., 2018. Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell*, 172(5), pp.1091–1107.e17.

Hänzelmann, S., Castelo, R. & Guinney, J., 2013. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC bioinformatics*, 14, p.7.

Haselkorn, J.K. et al., 2005. Overview of spasticity management in multiple sclerosis. Evidence-based management strategies for spasticity treatment in multiple sclerosis. *The journal of spinal cord medicine*, 28(2), pp.167–199.

Haythorne, E. et al., 2019. Diabetes causes marked inhibition of mitochondrial metabolism in pancreatic β-cells. *Nature communications*, 10(1), p.2474.

Heneka, M.T. et al., 2011. Impact and Therapeutic Potential of PPARs in Alzheimer's Disease. *Current neuropharmacology*, 9(4), pp.643–650.

Hodgson, D.R. et al., 2018. Candidate biomarkers of PARP inhibitor sensitivity in ovarian cancer beyond the BRCA genes. *British journal of cancer*, 119(11), pp.1401–1409.

Hofker, M.H., Fu, J. & Wijmenga, C., 2014. The genome revolution and its role in understanding complex diseases. *Biochimica et biophysica acta*, 1842(10), pp.1889–1895.

Houseman, E.A., Molitor, J. & Marsit, C.J., 2014. Reference-free cell mixture adjustments in analysis of DNA methylation data. *Bioinformatics* , 30(10), pp.1431–1439.

Hoxhaj, G. & Manning, B.D., 2020. The PI3K-AKT network at the interface of oncogenic signalling and cancer metabolism. *Nature reviews. Cancer*, 20(2), pp.74–88.

Huang, B., Bates, M. & Zhuang, X., 2009. Super-resolution fluorescence microscopy. *Annual review of biochemistry*, 78, pp.993–1016.

Huang, D.W., Sherman, B.T. & Lempicki, R.A., 2009a. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research*, 37(1), pp.1–13.

Huang, D.W., Sherman, B.T. & Lempicki, R.A., 2009b. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols*, 4(1), pp.44–57.

Huang, W. et al., 2019. Sodium valproate induced acute pancreatitis in a bipolar disorder patient:

a case report. *BMC pharmacology & toxicology*, 20(1), p.71.

Huan, Q. et al., 2018. HeteroMeth: A Database of Cell-to-cell Heterogeneity in DNA Methylation. *Genomics, proteomics & bioinformatics*, 16(4), pp.234–243.

Hua, Y. et al., 2023. Metformin and cancer hallmarks: shedding new lights on therapeutic repurposing. *Journal of translational medicine*, 21(1), p.403.

Ibn-Salem, J., Muro, E.M. & Andrade-Navarro, M.A., 2017. Co-regulation of paralog genes in the three-dimensional chromatin architecture. *Nucleic acids research*, 45(1), pp.81–91.

Imkeller, K. et al., 2022. Metabolic balance in colorectal cancer is maintained by optimal Wnt signaling levels. *Molecular systems biology*, 18(8), p.e10874.

Iorio, F. et al., 2018. Pathway-based dissection of the genomic heterogeneity of cancer hallmarks' acquisition with SLAPenrich. *Scientific reports*, 8(1), p.6713.

Jain, Y. et al., 2021. Convergent evolution of a genomic rearrangement may explain cancer resistance in hystrico- and sciuromorpha rodents. *NPJ aging and mechanisms of disease*, 7(1), p.20.

James, C. et al., 2024. Evolutionary analysis of gene ages across TADs associates chromatin topology with whole-genome duplications. *Cell reports*, 43(4), p.113895.

Jeon, J., Han, E.Y. & Jung, I., 2023. MOPA: An integrative multi-omics pathway analysis method for measuring omics activity. *PloS one*, 18(3), p.e0278272.

Jha, I.P. et al., 2021. Learning the Mental Health Impact of COVID-19 in the United States With Explainable Artificial Intelligence: Observational Study. *JMIR mental health*, 8(4), p.e25097.

Jiang, M., Chen, H. & Guo, G., 2021. Studying Kidney Diseases at the Single-Cell Level. *Kidney diseases (Basel, Switzerland)*, 7(5), pp.335–342.

Joly, J.H., Lowry, W.E. & Graham, N.A., 2021. Differential Gene Set Enrichment Analysis: a statistical approach to quantify the relative enrichment of two gene sets. *Bioinformatics (Oxford, England)*, 36(21), pp.5247–5254.

Jovic, D. et al., 2022. Single-cell RNA sequencing technologies and applications: A brief overview. *Clinical and translational medicine*, 12(3), p.e694.

Ju, W. et al., 2013. Defining cell-type specificity at the transcriptional level in human disease. *Genome research*, 23(11), pp.1862–1873.

Kahn, B.B., 1998. Type 2 diabetes: when insulin secretion fails to compensate for insulin resistance. *Cell*, 92(5), pp.593–596.

Kalofoutis, A. et al., 1980. Erythrocyte phospholipid fatty acid fluctuations in patients with

beta-thalassemia minor. *Clinical biochemistry*, 13(6), pp.273–276.

Kanehisa, M. & Goto, S., 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1), pp.27–30.

Kaur, B. et al., 2023. Editorial: Hallmark of cancer: Reprogramming of cellular metabolism. *Frontiers in oncology*, 12. Available at: https://www.frontiersin.org/journals/oncology/articles/10.3389/fonc.2022.1126913.

Kempfer, R. & Pombo, A., 2020. Methods for mapping 3D chromosome architecture. *Nature reviews. Genetics*, 21(4), pp.207–226.

Khier, S. & Lohan, L., 2018. Kinetics of circulating cell-free DNA for biomedical applications: critical appraisal of the literature. *Future science OA*, 4(4), p.FSO295.

Kim, B.N. et al., 2020. TGF-β induced EMT and stemness characteristics are associated with epigenetic regulation in lung cancer. *Scientific reports*, 10(1), p.10597.

Kim, M.-C. et al., 2016. An Integrated Analysis of the Genome-Wide Profiles of DNA Methylation and mRNA Expression Defining the Side Population of a Human Malignant Mesothelioma Cell Line. *Journal of Cancer*, 7(12), pp.1668–1679.

Kim, S.K. & Cho, S.W., 2022. The Evasion Mechanisms of Cancer Immunity and Drug Intervention in the Tumor Microenvironment. *Frontiers in pharmacology*, 13, p.868695.

Kiselev, I.S. et al., 2021. DNA Methylation As an Epigenetic Mechanism in the Development of Multiple Sclerosis. *Acta naturae*, 13(2), pp.45–57.

Klemm, S.L., Shipony, Z. & Greenleaf, W.J., 2019. Chromatin accessibility and the regulatory epigenome. *Nature reviews. Genetics*, 20(4), pp.207–220.

Kleppe, A. et al., 2018. Chromatin organisation and cancer prognosis: a pan-cancer study. *The lancet oncology*, 19(3), pp.356–369.

Kochar, D.K. et al., 2004. Sodium valproate for painful diabetic neuropathy: a randomized double-blind placebo-controlled study. *QJM: monthly journal of the Association of Physicians*, 97(1), pp.33–38.

Koch, C.M. et al., 2018. A Beginner's Guide to Analysis of RNA Sequencing Data. *American journal of respiratory cell and molecular biology*, 59(2), pp.145–157.

Kottaisamy, C.P.D. et al., 2021. Experimental animal models for diabetes and its related complications-a review. *Laboratory animal research*, 37(1), p.23.

Kramer, N.E. et al., 2022. Plotgardener: cultivating precise multi-panel figures in R. *Bioinformatics* , 38(7), pp.2042–2045.

Krueger, F. & Andrews, S.R., 2011. Bismark: a flexible aligner and methylation caller for

Bisulfite-Seq applications. *Bioinformatics* , 27(11), pp.1571–1572.

Kumar, V. & Others, 2022. *Unraveling cellular heterogeneity and phenotypic drug responses using chromatin profiles*. IIIT-Delhi. Available at: https://repository.iiitd.edu.in/xmlui/handle/123456789/1258.

Kwiatkowska, E. et al., 2016. Effect of 3-bromopyruvate acid on the redox equilibrium in non-invasive MCF-7 and invasive MDA-MB-231 breast cancer cells. *Journal of bioenergetics and biomembranes*, 48(1), pp.23–32.

Kwon, H.-J. et al., 2023. Advances in methylation analysis of liquid biopsy in early cancer detection of colorectal and lung cancer. *Scientific reports*, 13(1), p.13502.

La Manno, G. et al., 2018. RNA velocity of single cells. *Nature*, 560(7719), pp.494–498.

Lamb, J. et al., 2006. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, 313(5795), pp.1929–1935.

Langfelder, P. & Horvath, S., 2008. WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics*, 9(1), p.559.

Lantz, K.A. et al., 2004. Foxa2 regulates multiple pathways of insulin secretion. *The Journal of clinical investigation*, 114(4), pp.512–520.

Lappalainen, T. & MacArthur, D.G., 2021. From variant to function in human disease genetics. *Science*, 373(6562), pp.1464–1468.

Lausen, B. & Schumacher, M., 1992. Maximally Selected Rank Statistics. *Biometrics*, 48(1), pp.73–85.

Leek, J.T. & Storey, J.D., 2007. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS genetics*, 3(9), pp.1724–1735.

de Leeuw, C.A. et al., 2018. Conditional and interaction gene-set analysis reveals novel functional pathways for blood pressure. *Nature communications*, 9(1), p.3768.

Liang, P.-I. et al., 2020. Curation of cancer hallmark-based genes and pathways for in silico characterization of chemical carcinogenesis. *Database: the journal of biological databases and curation*, 2020. Available at: http://dx.doi.org/10.1093/database/baaa045.

Liberzon, A. et al., 2011. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* , 27(12), pp.1739–1740.

Li, G. et al., 2014. Chromatin Interaction Analysis with Paired-End Tag (ChIA-PET) sequencing technology and application. *BMC genomics*, 15 Suppl 12(Suppl 12), p.S11.

Li, J. et al., 2019. Detection of Colorectal Cancer in Circulating Cell-Free DNA by Methylated CpG Tandem Amplification and Sequencing. *Clinical chemistry*, 65(7), pp.916–926.

Li, L. et al., 2019. Cancer Is Associated with Alterations in the Three-Dimensional Organization of the Genome. *Cancers*, 11(12). Available at: http://dx.doi.org/10.3390/cancers11121886.

Lin, L. et al., 2015. SLC transporters as therapeutic targets: emerging opportunities. *Nature reviews. Drug discovery*, 14(8), pp.543–560.

Li, Q., Hermanson, P.J. & Springer, N.M., 2018. Detection of DNA Methylation by Whole-Genome Bisulfite Sequencing. *Methods in molecular biology* , 1676, pp.185–196.

Liu, C.-J. et al., 2023. GSCA: an integrated platform for gene set cancer analysis at genomic, pharmacogenomic and immunogenomic levels. *Briefings in bioinformatics*, 24(1). Available at: http://dx.doi.org/10.1093/bib/bbac558.

Liu, L. et al., 2021. Characterization of cell free plasma methyl-DNA from xenografted tumors to guide the selection of diagnostic markers for early-stage cancers. *Frontiers in oncology*, 11. Available at: https://www.frontiersin.org/articles/10.3389/fonc.2021.615821/full.

Liu, S. et al., 2020. Finding new cancer epigenetic and genetic biomarkers from cell-free DNA by combining SALP-seq and machine learning: esophageal cancer as an example. *bioRxiv*. Available at: http://dx.doi.org/10.1101/2020.01.18.911172.

Liu, T. et al., 2019. TADKB: Family classification and a knowledge base of topologically associating domains. *BMC genomics*, 20(1), p.217.

Liu, X. et al., 2018. Digital gene expression profiling analysis and its application in the identification of genes associated with improved response to neoadjuvant chemotherapy in breast cancer. *World journal of surgical oncology*, 16(1), p.82.

Liu, Y., 2022. At the dawn: cell-free DNA fragmentomics and gene regulation. *British journal of cancer*, 126(3), pp.379–390.

Liu, Y. et al., 2021. DNA methylation-calling tools for Oxford Nanopore sequencing: a survey and human epigenome-wide evaluation. *Genome biology*, 22(1), p.295.

Liu, Y. & Chance, M.R., 2013. Pathway analyses and understanding disease associations. *Current genetic medicine reports*, 1(4). Available at: http://dx.doi.org/10.1007/s40142-013-0025-3.

Li, Y., Hu, M. & Shen, Y., 2018. Gene regulation in the 3D genome. *Human molecular genetics*, 27(R2), pp.R228–R233.

Li, Y. & Tollefsbol, T.O., 2010. Impact on DNA methylation in cancer prevention and therapy by bioactive dietary components. *Current medicinal chemistry*, 17(20), pp.2141–2151.

Li, Z. & Wu, H., 2019. TOAST: improving reference-free cell composition estimation by cross-cell type differential analysis. *Genome biology*, 20(1), p.190.

MacDonald, G. et al., 2014. Memo is a copper-dependent redox protein with an essential role in

migration and metastasis. *Science signaling*, 7(329), p.ra56.

Mackerer, C.R., Clay, G.A. & Dajani, E.Z., 1976. Loperamide binding to opiate receptor sites of brain and myenteric plexus. *The Journal of pharmacology and experimental therapeutics*, 199(1), pp.131–140.

Madsen, R.R. et al., 2021. Positive correlation between transcriptomic stemness and PI3K/AKT/mTOR signaling scores in breast cancer, and a counterintuitive relationship with PIK3CA genotype. *PLoS genetics*, 17(11), p.e1009876.

Maleki, F. et al., 2020. Gene Set Analysis: Challenges, Opportunities, and Future Research. *Frontiers in genetics*, 11, p.654.

Malhi, H. & Gores, G.J., 2008. Cellular and molecular mechanisms of liver injury. *Gastroenterology*, 134(6), pp.1641–1654.

Mallona, I., Díez-Villanueva, A. & Peinado, M.A., 2014. Methylation plotter: a web tool for dynamic visualization of DNA methylation data. *Source code for biology and medicine*, 9, p.11.

Mar, J.C. et al., 2011. attract: A method for identifying core pathways that define cellular phenotypes. *PloS one*, 6(10), p.e25445.

Mashili, F. et al., 2013. Constitutive STAT3 phosphorylation contributes to skeletal muscle insulin resistance in type 2 diabetes. *Diabetes*, 62(2), pp.457–465.

Mathur, R. et al., 2018. Gene set analysis methods: a systematic comparison. *BioData mining*, 11, p.8.

Mercer, S.L. & Coop, A., 2011. Opioid analgesics and P-glycoprotein efflux transporters: a potential systems-level contribution to analgesic tolerance. *Current topics in medicinal chemistry*, 11(9), pp.1157–1164.

Michael Conn, P., 2013. *Animal Models for the Study of Human Disease*, Academic Press.

Mi, H. & Thomas, P., 2009. PANTHER pathway: an ontology-based pathway database coupled with data analysis tools. *Methods in molecular biology* , 563, pp.123–140.

Mishra, S., Pandey, N., Rawat, A., et al., 2023. An Explainable Model Using Graph-Wavelet for Predicting Biophysical Properties of Proteins and Measuring Mutational Effects. *IEEE Access*, 11, pp.135222–135234.

Mishra, S., Pandey, N., Chawla, S., et al., 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), pp.218–231.

Mishra, S., Srivastava, D. & Kumar, V., 2021. Improving gene network inference with graph wavelets and making insights about ageing-associated regulatory changes in lungs.

Briefings in bioinformatics, 22(4). Available at: http://dx.doi.org/10.1093/bib/bbaa360.

Mladkova, J. et al., 2010. Phenotyping breast cancer cell lines EM-G3, HCC1937, MCF7 and MDA-MB-231 using 2-D electrophoresis and affinity chromatography for glutathione-binding proteins. *BMC cancer*, 10, p.449.

Mohanta, T.K., Mishra, A.K. & Al-Harrasi, A., 2021. The 3D Genome: From Structure to Function. *International journal of molecular sciences*, 22(21). Available at: http://dx.doi.org/10.3390/ijms222111585.

Mohanty, V. et al., 2021. Uncoupling of gene expression from copy number presents therapeutic opportunities in aneuploid cancers. *Cell reports. Medicine*, 2(7), p.100349.

Monteza, H.G.V. et al., 2020. Identificación de áreas con aptitud para el desarrollo de una maricultura sostenible en la región Lambayeque – Perú. , pp.32–54.

Morgan, H.D. et al., 1999. Epigenetic inheritance at the agouti locus in the mouse. *Nature genetics*, 23(3), pp.314–318.

Morgan, M.A. & Shilatifard, A., 2015. Chromatin signatures of cancer. *Genes & development*, 29(3), pp.238–249.

Morrissey, M.A. et al., 2016. SPARC Promotes Cell Invasion In Vivo by Decreasing Type IV Collagen Levels in the Basement Membrane. *PLoS genetics*, 12(2), p.e1005905.

Moss, J. et al., 2018. Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. *Nature communications*, 9(1), p.5068.

Motsinger, A.A. & Ritchie, M.D., 2006. Multifactor dimensionality reduction: an analysis strategy for modelling and detecting gene-gene interactions in human genetics and pharmacogenomics studies. *Human genomics*, 2(5), pp.318–328.

Mouliere, F. et al., 2014. Multi-marker analysis of circulating cell-free DNA toward personalized medicine for colorectal cancer. *Molecular oncology*, 8(5), pp.927–941.

Mourad, R., Sinoquet, C. & Leray, P., 2012. Probabilistic graphical models for genetic association studies. *Briefings in bioinformatics*, 13(1), pp.20–33.

Müller, F. et al., 2019. RnBeads 2.0: comprehensive analysis of DNA methylation data. *Genome biology*, 20(1), p.55.

Mumbach, M.R. et al., 2016. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nature methods*, 13(11), pp.919–922.

Nagarajan, R., Scutari, M. & Lèbre, S., 2014. *Bayesian Networks in R: with Applications in Systems Biology*, Springer Science & Business Media.

Negrini, S., Gorgoulis, V.G. & Halazonetis, T.D., 2010. Genomic instability--an evolving

hallmark of cancer. *Nature reviews. Molecular cell biology*, 11(3), pp.220–228.

Newman, A.M. et al., 2015. Robust enumeration of cell subsets from tissue expression profiles. *Nature methods*, 12(5), pp.453–457.

Ng, C.K.Y. et al., 2014. Predictive performance of microarray gene signatures: impact of tumor heterogeneity and multiple mechanisms of drug resistance. *Cancer research*, 74(11), pp.2946–2961.

Nguyen, H. et al., 2021. CPA: a web-based platform for consensus pathway analysis and interactive visualization. *Nucleic acids research*, 49(W1), pp.W114–W124.

Ni, L. et al., 2023. Pan-3D genome analysis reveals structural and functional differentiation of soybean genomes. *Genome biology*, 24(1), p.12.

Noë, M. et al., 2024. DNA methylation and gene expression as determinants of genome-wide cell-free DNA fragmentation. *Nature communications*, 15(1), p.6690.

Nomura, S., 2021. Single-cell genomics to understand disease pathogenesis. *Journal of human genetics*, 66(1), pp.75–84.

Novelle, M.G. et al., 2016. Metformin: A Hopeful Promise in Aging Research. *Cold Spring Harbor perspectives in medicine*, 6(3), p.a025932.

Ofman, J., Hall, M. & Aravanis, A., 2020. GRAIL and the quest for earlier multi-cancer detection. *Nature*.

Okhovat, M. et al., 2023. TAD evolutionary and functional characterization reveals diversity in mammalian TAD boundary properties and function. *Nature communications*, 14(1), p.8111.

Olivieri, O. et al., 1994. Oxidative damage and erythrocyte membrane transport abnormalities in thalassemias. *Blood*, 84(1), pp.315–320.

Olsson, M. & Zhivotovsky, B., 2011. Caspases and cancer. *Cell death and differentiation*, 18(9), pp.1441–1449.

Ospina, O.E. et al., 2024. Differential gene expression analysis of spatial transcriptomic experiments using spatial mixed models. *Scientific reports*, 14(1), p.10967.

Pacifico, A. & Leone, G., 2007. Role of p53 and CDKN2A inactivation in human squamous cell carcinomas. *Journal of biomedicine & biotechnology*, 2007(3), p.43418.

Panagopoulou, M., Esteller, M. & Chatzaki, E., 2021. Circulating Cell-Free DNA in Breast Cancer: Searching for Hidden Information towards Precision Medicine. *Cancers*, 13(4). Available at: http://dx.doi.org/10.3390/cancers13040728.

Pandey, N., Sharma, M., Mathur, A., Anene-Nzel, C.G., Hakimullah, M., Patel, P., et al., 2023. Deciphering drug response and phenotypic heterogeneity of cancer cells using gene

ensembles of regulatory units defined by chromatin domains. *bioRxiv*. Available at: https://doi.org/10.1101/2023.01.15.524115.

Pandey, N., Sharma, M., Mathur, A., Anene-Nzel, C.G., Hakimullah, M., Jha, I.P., et al., 2023. Deciphering the phenotypic heterogeneity and drug response in cancer cells using genome-wide activity and interaction of chromatin domains. *bioRxiv*, p.2023.01.15.524115. Available at: https://www.biorxiv.org/content/10.1101/2023.01.15.524115v1 [Accessed August 22, 2024].

Pandey, N. et al., 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, p.738194.

Pandey, R.V. et al., 2016. ClinQC: a tool for quality control and cleaning of Sanger and NGS data in clinical research. *BMC bioinformatics*, 17, p.56.

Pardoll, D.M., 2012. The blockade of immune checkpoints in cancer immunotherapy. *Nature reviews. Cancer*, 12(4), pp.252–264.

Patel, R.K. & Jain, M., 2012. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PloS one*, 7(2), p.e30619.

Pelizzola, M. et al., 2008. MEDME: an experimental and analytical methodology for the estimation of DNA methylation levels based on microarray derived MeDIP-enrichment. *Genome research*, 18(10), pp.1652–1659.

Peneder, P. et al., 2021. Multimodal analysis of cell-free DNA whole-genome sequencing for pediatric cancers with low mutational burden. *Nature communications*, 12(1), p.3230.

Pisco, A.O. & Huang, S., 2015. Non-genetic cancer cell plasticity and therapy-induced stemness in tumour relapse: "What does not kill me strengthens me." *British journal of cancer*, 112(11), pp.1725–1732.

Polak, P. et al., 2015. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature*, 518(7539), pp.360–364.

Pulkkinen, L. et al., 2010. Ghrelin in diabetes and metabolic syndrome. *International journal of peptides*, 2010. Available at: http://dx.doi.org/10.1155/2010/248948.

Pulumati, A. et al., 2023. Technological advancements in cancer diagnostics: Improvements and limitations. *Cancer reports*, 6(2), p.e1764.

Qiu, X. et al., 2020. Inferring Causal Gene Regulatory Networks from Coupled Single-Cell Expression Dynamics Using Scribe. *Cell systems*, 10(3), pp.265–274.e11.

Qi, Z. & Tan, H., 2020. Association between MGMT status and response to alkylating agents in patients with neuroendocrine neoplasms: a systematic review and meta-analysis. *Bioscience reports*, 40(3). Available at: http://dx.doi.org/10.1042/BSR20194127.

Quinodoz, S.A. et al., 2022. SPRITE: a genome-wide method for mapping higher-order 3D interactions in the nucleus using combinatorial split-and-pool barcoding. *Nature protocols*, 17(1), pp.36–75.

Rachmilewitz, E.A., Shohet, S.B. & Lubin, B.H., 1976. Lipid membrane peroxidation in beta-thalassemia major. *Blood*, 47(3), pp.495–505.

Radhakrishna Rao, C., 2008. *Linear Statistical Inference and Its Applications*, Wiley & Sons, Incorporated, John.

Rahmani, E. et al., 2016. Sparse PCA corrects for cell type heterogeneity in epigenome-wide association studies. *Nature methods*, 13(5), pp.443–445.

Rahmatallah, Y., Emmert-Streib, F. & Glazko, G., 2014. Gene Sets Net Correlations Analysis (GSNCA): a multivariate differential coexpression test for gene sets. *Bioinformatics* , 30(3), pp.360–368.

Ramón Y Cajal, S. et al., 2020. Clinical implications of intratumor heterogeneity: challenges and opportunities. *Journal of molecular medicine* , 98(2), pp.161–177.

Rao, S.S.P. et al., 2014. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7), pp.1665–1680.

Raudvere, U. et al., 2019. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic acids research*, 47(W1), pp.W191–W198.

Rees, M.G. et al., 2016. Correlating chemical sensitivity and basal gene expression reveals mechanism of action. *Nature chemical biology*, 12(2), pp.109–116.

Rekand, T. & Grønning, M., 2011. Treatment of spasticity related to multiple sclerosis with intrathecal baclofen: a long-term follow-up. *Journal of rehabilitation medicine: official journal of the UEMS European Board of Physical and Rehabilitation Medicine*, 43(6), pp.511–514.

Riba, A. et al., 2022. Cell cycle gene regulation dynamics revealed by RNA velocity and deep-learning. *Nature communications*, 13(1), p.2865.

de Rivero Vaccari, J.P. et al., 2014. RIG-1 receptor expression in the pathology of Alzheimer's disease. *Journal of neuroinflammation*, 11, p.67.

Roncati, L. & Figueiredo, C.R., 2023. Editorial: Hallmark of cancer: tumor promoting inflammation. *Frontiers in oncology*, 13, p.1242407.

Russano, M. et al., 2020. Liquid biopsy and tumor heterogeneity in metastatic solid tumors: the potentiality of blood samples. *Journal of experimental & clinical cancer research: CR*, 39(1), p.95.

Said Suliman, A. et al., 2021. Cyclodextrin Diethyldithiocarbamate Copper II Inclusion

Complexes: A Promising Chemotherapeutic Delivery System against Chemoresistant Triple Negative Breast Cancer Cell Lines. *Pharmaceutics*, 13(1). Available at: http://dx.doi.org/10.3390/pharmaceutics13010084.

Salvi, S. et al., 2016. Cell-free DNA as a diagnostic marker for cancer: current insights. *OncoTargets and therapy*, 9, pp.6549–6559.

Saman, H. et al., 2020. Inducing angiogenesis, a key step in cancer vascularization, and treatment approaches. *Cancers*, 12(5). Available at: https://www.mdpi.com/2072-6694/12/5/1172.

Samuels, D.S. et al., 2021. Gene Regulation and Transcriptomics. *Current issues in molecular biology*, 42, pp.223–266.

Scheel, C.H. & Schäfer, R., 2023. Editorial: Hallmark of cancer: Evasion of growth suppressors. *Frontiers in oncology*, 13, p.1170115.

Scherer, M. et al., 2020. Quantitative comparison of within-sample heterogeneity scores for DNA methylation data. *Nucleic acids research*, 48(8), p.e46.

Schumacher, A. et al., 2006. Microarray-based DNA methylation profiling: technology and applications. *Nucleic acids research*, 34(2), pp.528–542.

Scutari, M., 2009. Learning Bayesian Networks with the bnlearn R Package. *arXiv [stat.ML]*. Available at: http://arxiv.org/abs/0908.3817.

Segerstolpe, Å. et al., 2016. Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes. *Cell metabolism*, 24(4), pp.593–607.

Shakoori, A.R., 2017. Fluorescence In Situ Hybridization (FISH) and Its Applications. In T. A. Bhat & A. A. Wani, eds. *Chromosome Structure and Aberrations*. New Delhi: Springer India, pp. 343–367.

Sharma, M., Jha, I.P., et al., 2022. Associating pathways with diseases using single-cell expression profiles and making inferences about potential drugs. *Briefings in bioinformatics*, 23(4). Available at: http://dx.doi.org/10.1093/bib/bbac241.

Sharma, M., Verma, R.K., et al., 2022. Computational challenges in detection of cancer using cell-free DNA methylation. *Computational and structural biotechnology journal*, 20, pp.26–39.

Sharma, R. et al., 2020. FITs: forest of imputation trees for recovering true signals in single-cell open chromatin profiles. *NAR genomics and bioinformatics*, 2(4), p.lqaa091.

Sha, S. et al., 2022. Prognostic analysis of cuproptosis-related gene in triple-negative breast cancer. *Frontiers in immunology*, 13, p.922780.

Shchetynsky, K. et al., 2017. Discovery of new candidate genes for rheumatoid arthritis through

integration of genetic association data with expression pathway analysis. *Arthritis research & therapy*, 19(1), p.19.

Shibuya, M., 2011. Vascular Endothelial Growth Factor (VEGF) and Its Receptor (VEGFR) Signaling in Angiogenesis: A Crucial Target for Anti- and Pro-Angiogenic Therapies. *Genes & cancer*, 2(12), pp.1097–1105.

Shi, Y., Ye, P. & Long, X., 2017. Differential Expression Profiles of the Transcriptome in Breast Cancer Cell Lines Revealed by Next Generation Sequencing. *Cellular physiology and biochemistry: international journal of experimental cellular physiology, biochemistry, and pharmacology*, 44(2), pp.804–816.

Shmakov, S.A. et al., 2018. Systematic prediction of genes functionally linked to CRISPR-Cas systems by gene neighborhood analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 115(23), pp.E5307–E5316.

Shroff, H. et al., 2024. Live-cell imaging powered by computation. *Nature reviews. Molecular cell biology*, 25(6), pp.443–463.

Shunxi, W. et al., 2023. Serine Metabolic Reprogramming in Tumorigenesis, Tumor Immunity, and Clinical Treatment. *Advances in nutrition* , 14(5), pp.1050–1066.

SIGMA Type 2 Diabetes Consortium et al., 2014. Sequence variants in SLC16A11 are a common risk factor for type 2 diabetes in Mexico. *Nature*, 506(7486), pp.97–101.

Silva, P.P. et al., 2022. A machine learning-based SNP-set analysis approach for identifying disease-associated susceptibility loci. *Scientific reports*, 12(1), p.15817.

Silva, T.C. et al., 2016. TCGA Workflow: Analyze cancer genomics and epigenomics data using Bioconductor packages. *F1000Research*, 5, p.1542.

Silvestri, A. et al., 2015. Metformin Induces Apoptosis and Downregulates Pyruvate Kinase M2 in Breast Cancer Cells Only When Grown in Nutrient-Poor Conditions. *PloS one*, 10(8), p.e0136250.

Sollis, E. et al., 2023. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic acids research*, 51(D1), pp.D977–D985.

Song, S. & Black, M.A., 2008. Microarray-based gene set analysis: a comparison of current methods. *BMC bioinformatics*, 9, p.502.

Sorrentino, V.G. et al., 2021. Hypomethylating Chemotherapeutic Agents as Therapy for Myelodysplastic Syndromes and Prevention of Acute Myeloid Leukemia. *Pharmaceuticals* , 14(7). Available at: http://dx.doi.org/10.3390/ph14070641.

Spector, B.L. et al., 2023. The methylome and cell-free DNA: current applications in medicine and pediatric disease. *Pediatric research*, 94(1), pp.89–95.

Sportelli, C. et al., 2020. Metformin as a Potential Neuroprotective Agent in Prodromal Parkinson's Disease-Viewpoint. *Frontiers in neurology*, 11, p.556.

Srivastava, D. et al., 2018. CellAtlasSearch: a scalable search engine for single cells. *Nucleic acids research*, 46(W1), pp.W141–W147.

Stetson, L.C. et al., 2021. Single cell RNA sequencing of AML initiating cells reveals RNA-based evolution during disease progression. *Leukemia*, 35(10), pp.2799–2812.

Stoney, R. et al., 2018. Mapping biological process relationships and disease perturbations within a pathway network. *NPJ systems biology and applications*, 4, p.22.

Subramanian, A. et al., 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43), pp.15545–15550.

Sun, R. et al., 2019. Powerful gene set analysis in GWAS with the Generalized Berk-Jones statistic. *PLoS genetics*, 15(3), p.e1007530.

Szabo, Q., Bantignies, F. & Cavalli, G., 2019. Principles of genome folding into topologically associating domains. *Science advances*, 5(4), p.eaaw1668.

Sze, C.C. & Shilatifard, A., 2016. MLL3/MLL4/COMPASS Family on Epigenetic Regulation of Enhancer Function and Cancer. *Cold Spring Harbor perspectives in medicine*, 6(11). Available at: http://dx.doi.org/10.1101/cshperspect.a026427.

Tanay, A. & Regev, A., 2017. Scaling single-cell genomics from phenomenology to mechanism. *Nature*, 541(7637), pp.331–338.

Tena, J.J. & Santos-Pereira, J.M., 2021. Topologically Associating Domains and Regulatory Landscapes in Development, Evolution and Disease. *Frontiers in cell and developmental biology*, 9, p.702787.

Teschendorff, A.E. et al., 2017. A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies. *BMC bioinformatics*, 18(1), p.105.

Teschendorff, A.E. & Zheng, S.C., 2017. Cell-type deconvolution in epigenome-wide association studies: a review and recommendations. *Epigenomics*, 9(5), pp.757–768.

Therneau, T.M. & Grambsch, P.M., *Modeling Survival Data: Extending the Cox Model*, Springer New York.

Thomas, C.K., Häger-Ross, C.K. & Klein, C.S., 2010. Effects of baclofen on motor units paralysed by chronic cervical spinal cord injury. *Brain: a journal of neurology*, 133(Pt 1), pp.117–125.

Thorvaldsdóttir, H., Robinson, J.T. & Mesirov, J.P., 2013. Integrative Genomics Viewer (IGV):

high-performance genomics data visualization and exploration. *Briefings in bioinformatics*, 14(2), pp.178–192.

Titus, A.J. et al., 2017. Cell-type deconvolution from DNA methylation: a review of recent applications. *Human molecular genetics*, 26(R2), pp.R216–R224.

Tsvetkov, P. et al., 2022. Copper induces cell death by targeting lipoylated TCA cycle proteins. *Science*, 375(6586), pp.1254–1261.

Uehiro, N. et al., 2016. Circulating cell-free DNA-based epigenetic assay can detect early breast cancer. *Breast cancer research: BCR*, 18(1), p.129.

University of Illinois, 2017. Phentermine: Side effects, dosage, uses, and more. *Healthline*. Available at: https://www.healthline.com/health/drugs/phentermine-oral-capsule [Accessed June 15, 2024].

Velayos, T. et al., 2017. An Activating Mutation in STAT3 Results in Neonatal Diabetes Through Reduced Insulin Synthesis. *Diabetes*, 66(4), pp.1022–1029.

Venables, W.N. & Ripley, B.D., *Modern Applied Statistics with S*, Springer New York.

Vieira, M.N.N. et al., 2017. Protein Tyrosine Phosphatase 1B (PTP1B): A Potential Target for Alzheimer's Therapy? *Frontiers in aging neuroscience*, 9, p.7.

Vitaliti, A. et al., 2023. AKT-driven epithelial-mesenchymal transition is affected by copper bioavailability in HER2 negative breast cancer cells via a LOXL2-independent mechanism. *Cellular oncology* , 46(1), pp.93–115.

Vogelstein, B. & Kinzler, K.W., 2004. Cancer genes and the pathways they control. *Nature medicine*, 10(8), pp.789–799.

Wagner, A., Regev, A. & Yosef, N., 2016. Revealing the vectors of cellular identity with single-cell genomics. *Nature biotechnology*, 34(11), pp.1145–1160.

Wang, H. et al., 2015. Individual-level analysis of differential expression of genes and pathways for personalized medicine. *Bioinformatics* , 31(1), pp.62–68.

Wang, J. et al., 2018. Integrated analysis of DNA methylation profiling and gene expression profiling identifies novel markers in lung cancer in Xuanwei, China. *PloS one*, 13(10), p.e0203155.

Wang, K.C. & Chang, H.Y., 2018. Epigenomics: Technologies and Applications. *Circulation research*, 122(9), pp.1191–1199.

Wang, L. et al., 2020. Single-cell RNA-seq reveals the immune escape and drug resistance mechanisms of mantle cell lymphoma. *Cancer biology & medicine*, 17(3), pp.726–739.

Wang, Q. et al., 2022. Gene body methylation in cancer: molecular mechanisms and clinical

applications. *Clinical epigenetics*, 14(1), p.154.

Wang, V. et al., 2013. GeneTopics--interpretation of gene sets via literature-driven topic models. *BMC systems biology*, 7 Suppl 5(Suppl 5), p.S10.

Wang, Y. et al., 2024. Cuproptosis: A novel therapeutic target for overcoming cancer drug resistance. *Drug resistance updates: reviews and commentaries in antimicrobial and anticancer chemotherapy*, 72, p.101018.

Wang, Y. et al., 2022. The shared biomarkers and pathways of systemic lupus erythematosus and metabolic syndrome analyzed by bioinformatics combining machine learning algorithm and single-cell sequencing analysis. *Frontiers in immunology*, 13, p.1015882.

Wang, Y., Chen, P.-M. & Liu, R.-B., 2018. Advance in plasma SEPT9 gene methylation assay for colorectal cancer early detection. *World journal of gastrointestinal oncology*, 10(1), pp.15–22.

Wang, Y. & Xian, H., 2022. Identifying Genes Related to Acute Myocardial Infarction Based on Network Control Capability. *Genes*, 13(7). Available at: http://dx.doi.org/10.3390/genes13071238.

Wan, Y.-W. et al., 2016. XMRF: an R package to fit Markov Networks to high-throughput genetics data. *BMC systems biology*, 10 Suppl 3, p.69.

Wilson, G.A. et al., 2012. Resources for methylome analysis suitable for gene knockout studies of potential epigenome modifiers. *GigaScience*, 1(1), p.3.

Wilson, N.K. et al., 2015. Combined Single-Cell Functional and Gene Expression Analysis Resolves Heterogeneity within Stem Cell Populations. *Cell stem cell*, 16(6), pp.712–724.

Wolfrum, C. et al., 2003. Insulin regulates the activity of forkhead transcription factor Hnf-3beta/Foxa-2 by Akt-mediated phosphorylation and nuclear/cytosolic localization. *Proceedings of the National Academy of Sciences of the United States of America*, 100(20), pp.11624–11629.

Wu, H. & Sun, Y.E., 2006. Epigenetic regulation of stem cell differentiation. *Pediatric research*, 59(4 Pt 2), p.21R–5R.

Xiao, N. et al., 2022. Disentangling direct from indirect relationships in association networks. *Proceedings of the National Academy of Sciences of the United States of America*, 119(2). Available at: http://dx.doi.org/10.1073/pnas.2109995119.

Xi, Y. & Li, W., 2009. BSMAP: whole genome bisulfite sequence MAPping program. *BMC bioinformatics*, 10, p.232.

Yaari, G. et al., 2013. Quantitative set analysis for gene expression: a method to quantify gene set differential expression including gene-gene correlations. *Nucleic acids research*, 41(18), p.e170.

Yang, M. et al., 2017. Role of the JAK2/STAT3 signaling pathway in the pathogenesis of type 2 diabetes mellitus with macrovascular complications. *Oncotarget*, 8(57), pp.96958–96969.

Yao, K., Tong, C.-Y. & Cheng, C., 2022. A framework to predict the applicability of Oncotype DX, MammaPrint, and E2F4 gene signatures for improving breast cancer prognostic prediction. *Scientific reports*, 12(1), p.2211.

Yao, S. et al., 2021. Case Report: Combination Therapy With PD-1 Blockade for Acute Myeloid Leukemia After Allogeneic Hematopoietic Stem Cell Transplantation Resulted in Fatal GVHD. *Frontiers in immunology*, 12, p.639217.

Yaswen, P. et al., 2015. Therapeutic targeting of replicative immortality. *Seminars in cancer biology*, 35 Suppl(Suppl), pp.S104–S128.

Yin, K. et al., 2019. Using weighted gene co-expression network analysis to identify key modules and hub genes in tongue squamous cell carcinoma. *Medicine*, 98(37), p.e17100.

Yuan, G.-C. et al., 2017. Challenges and emerging directions in single-cell analysis. *Genome biology*, 18(1), p.84.

Yuan, H. et al., 2019. CancerSEA: a cancer single-cell state atlas. *Nucleic acids research*, 47(D1), pp.D900–D908.

Yu, F. et al., 2020. CFEA: a cell-free epigenome atlas in human diseases. *Nucleic acids research*, 48(D1), pp.D40–D44.

Zhang, B., Kirov, S. & Snoddy, J., 2005. WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic acids research*, 33(Web Server issue), pp.W741–8.

Zhang, D. et al., 2020. CHG: A Systematically Integrated Database of Cancer Hallmark Genes. *Frontiers in genetics*, 11, p.29.

Zhang, R. et al., 2023. G-Quadruplex Structures Are Key Modulators of Somatic Structural Variants in Cancers. *Cancer research*, 83(8), pp.1234–1248.

Zhang, Y. et al., 2019. The SLC transporter in nutrient and metabolic sensing, regulation, and drug development. *Journal of molecular cell biology*, 11(1), pp.1–13.

Zhang, Z. et al., 2022. Recent progress in DNA methyltransferase inhibitors as anticancer agents. *Frontiers in pharmacology*, 13, p.1072651.

Zhao, T. et al., 2021. SC2disease: a manually curated database of single-cell transcriptome for human diseases. *Nucleic acids research*, 49(D1), pp.D1413–D1419.

Zhao, Z. et al., 2006. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nature genetics*, 38(11), pp.1341–1347.

Zheng, H., Zhu, M.S. & Liu, Y., 2021. FinaleDB: a browser and database of cell-free DNA fragmentation patterns. *Bioinformatics* , 37(16), pp.2502–2503.

Zhou, Q. et al., 2019. An integrated package for bisulfite DNA methylation data analysis with Indel-sensitive mapping. *BMC bioinformatics*, 20(1), p.47.

Zhou, W. et al., 2020. DNA methylation enables transposable element-driven genome expansion. *Proceedings of the National Academy of Sciences of the United States of America*, 117(32), pp.19359–19366.

Zhou, Y. et al., 2022. Therapeutic target database update 2022: facilitating drug discovery with enriched comparative data of targeted agents. *Nucleic acids research*, 50(D1), pp.D1398–D1407.

Zhu, S. et al., 2019. GIGSEA: genotype imputed gene set enrichment analysis using GWAS summary level data. *Bioinformatics (Oxford, England)*, 35(1), pp.160–163.

Zhu, Y. et al., 2022. Editorial: Hallmark of cancer: Resisting cell death. *Frontiers in oncology*, 12. Available at: https://www.frontiersin.org/journals/oncology/articles/10.3389/fonc.2022.1069947.

Zucker, S., Cao, J. & Chen, W.T., 2000. Critical appraisal of the use of matrix metalloproteinase inhibitors in cancer treatment. *Oncogene*, 19(56), pp.6642–6650.

Zukowski, A., Rao, S. & Ramachandran, S., 2020. Phenotypes from cell-free DNA. *Open biology*, 10(9), p.200119.